These notes are made publicly available in the hope that they will be useful. All reports of errata will be gratefully received. I will also be glad to hear from anyone who reads them, whether or not you find errors: pcn@upenn.edu.
Brief Contents

PART I  Preliminaries

Prologue  2

Chapter 1  |  Warmup: Newtonian Gravitation  17
Unbound comets follow hyperbolic trajectories.

PART II  Static and Almost-Static

Chapter 2  |  Electrostatics Introduced  27
As a proton beam penetrates tissue, it suddenly loses most of its energy in a narrow range of depth.

Chapter 3  |  Electrostatic Multipole Expansion  37
Molecular symmetry gives some quick, qualitative predictions about molecular interactions.

Chapter 4  |  Vista: Fluorescence Resonance Energy Transfer  58
Resonance energy transfer creates a “private communication channel” between two fluorophores with a characteristic dependence on distance, and on the relative orientations of the two molecules and the vector separating them.

Chapter 5  |  Curvilinear Coordinates and Separation of Variables  67
The tip of a nearfield scanning optical microscope generates huge electrostatic fields localized to nanometer regions.

Chapter 6  |  Capacitors  77
A charged capacitor will pull dielectric material into its gap.

Chapter 7  |  Vista: Electrohydrostatics  99
The interface between a conducting and an insulating fluid can form a sharp conical point, despite surface tension.

Chapter 8  |  Charge Flux, Continuity Equation, and Ohmic Conductors  115
A traveling nerve impulse or muscle contraction leads to a multipolar electrical disturbance that can be measured from far away.

Chapter 9  |  Vista: Cell Membrane Capacitance  132
The capacitance of a cell can be measured without placing electrodes on either side of it.

Chapter 10  |  Statistical Electrostatics of Solutions  139
DNA falls apart into separate strands in pure water.

Chapter 11 | The Cable Equation 168
Small electrical disturbances on a nerve axon spread diffusively.

Chapter 12 | Vista: The Action Potential in Neurons 181
A nerve axon carries signals that preserve their form and amplitude as they travel long distances.

Chapter 13 | Examples of 3-Tensors in Physics 195
Molecular polarizability is in general anisotropic.

Chapter 14 | Tensors from Heaven 209
Although nematic liquid crystals are made from complicated molecules, only a few physical constants are needed to describe their overall behavior.

Chapter 15 | Magnetostatics 221
Tiny magnetic field disturbances can reveal brain activity without requiring invasive probes.

Chapter 16 | Units and Dimensional Analysis 241
Eddy currents slow the fall of a magnet through a conducting tube.

Chapter 17 | Magnetostatic Multipole Expansion 252
Just three constants suffice to characterize the far fields of even a complicated stationary current distribution.

PART III Dynamic

Chapter 18 | Beyond Statics 267
Light can be created in helicity states, that is, states with electric field vector that rotates instead of oscillating as the wave advances.

Chapter 19 | [Vista: Waveguides] 296

Chapter 20 | First Look at Energy and Momentum Transport by Waves 299
The expansion of the early Universe was faster than predicted from gas pressure alone.

Chapter 21 | Geometrical Optics and the Eikonal 307
Light generically focuses into networks of bright lines against a diffuse background when it encounters curved interfaces.

Chapter 22 | Diffraction and Semiclassical Approximation 340
Light can go to places forbidden by geometrical optics.
Chapter 23 | Diffraction at Caustics 362
Even when spherical and chromatic aberrations are small, diffraction still prevents perfect focusing by a lens.

Chapter 24 | Partial Polarization 385
For optical purposes, polarization of light may be described by a point inside an abstract sphere.

Chapter 25 | Generation of Radiation: First Look 391
An antenna emits energy with a specific directional pattern.

PART IV  Relativity: Low Tech

Chapter 26 | Galilean Relativity 405
Light from distant objects arrives at Earth with a delay related to distance, but not velocity, of the source.

Chapter 27 | Springs, Strings, and Local Conservation Laws 420
Energy and momentum are locally conserved on a vibrating string; they cannot disappear at one point and reappear at a distant point without passing through the intervening region.

Chapter 28 | Einstein’s Version of Relativity: Overview 427
Vacuum is a unique state; it has no measurable descriptors analogous to the density or velocity of a medium that carries sound waves.

Chapter 29 | Provisional Lorentz Transformations and the Fizeau Experiment 439
The speed of light in flowing water is different from that in still water (the light is “dragged along”), but in a quantitatively different way from the newtonian expectation.

Chapter 30 | Aberration of Starlight and Doppler Effects 451
Each star’s apparent position is shifted relative to others, depending on Earth’s momentary velocity.

Chapter 31 | Relativistic Momentum and Energy of Particles 475
The mass of a deuterium nucleus is less than the sum of the proton and neutron masses.

PART V  Relativity: High Tech

Chapter 32 | Four-Vectors 487
Electrons, protons, and other “material” particles also show wavelike behavior.

Chapter 33 | The Faraday Tensor 510
The orbital period of a charged particle in uniform magnetic field starts to depend on its energy, when that energy is increased beyond $mc^2$.

Chapter 34 | Manifestly Invariant Form of Maxwell 525
A suddenly accelerated charge emits a pulse of electromagnetic radiation.
Chapter 35 | Energy and Momentum of Fields  554
Superconducting magnets can fail catastrophically.

Chapter 36 | Interlude: Faraday’s Field Lines  568
Like charges, and like magnetic pole tips, repel; opposite charges and pole tips attract.

Chapter 37 | Plane Waves in 4D Language  574
The two polarizations of light do not give rise to visible interference fringes when they are combined; they seem to act as independent channels.

Chapter 38 | A Simple Spherical Wave  584
Each far-field wavefront of a dipole spherical wave is spherical, but the resulting energy flux is not isotropic.

Chapter 39 | Radiation Green Function Revisited  590
In vacuum, a charged particle in uniform motion carries fields along with it but does not radiate.

Chapter 40 | Vista: Variational Formulation  605

PART VI  Beams, Radiation and Scattering

Chapter 41 | Beams: Gaussian, Vortex, and Bessel  619
A structured beam of light can transfer far more angular momentum than a circularly polarized plane wave.

Chapter 42 | Vista: J. J. Thomson’s Pictorial Explanation of Radiation  633
The pulse of radiation from a suddenly accelerated charge consists of fields that are transversely polarized, have maximal strength in the plane perpendicular to the acceleration, and fall with distance as \( \frac{1}{r} \).

Chapter 43 | Electric Dipole Radiation  641
Diatomic, homonuclear molecules have little infrared activity, but more complex ones can be strongly IR-active gases.

Chapter 44 | Higher-Multipole Radiation  654
Some radiative nuclear transitions are much faster than others.

[[Vista: Transition radiation]]  662

Chapter 45 | Synchrotron Radiation  663

[[Vista: Radiation Reaction]]  673

Chapter 46 | The Microwave Polarizer  674
An array of aligned, linear conductors can act as a polarizing filter, and can even regenerate a missing polarization in an incoming beam.

Contents Index Notation Glossary November 14, 2023
Chapter 47 | Scattering by Free and Bound Charges 678
The angular modulation of the cosmic microwave background radiation’s polarization tells us about inhomogeneity of the early Universe.

Chapter 48 | [[Vista: Scattering by Many Objects]] 685
[[Vista: Scattering by a Dielectric Sphere]] 686

PART VII  Light in Linear Media

Chapter 49 | Light in Simple Media 688
How can light “tunnel” into a zone forbidden by geometrical optics?

[[Vista: Microscopy]] 705
[[Vista: Optical Solitons in Nonlinear Media]] 705

Chapter 50 | Čerenkov Radiation 706
When a charged particle moves through a medium faster than the local speed of light, it emits radiation even without accelerating.

Chapter 51 | Chiral Media 711
A solution of randomly oriented molecules is completely isotropic, yet nevertheless it can rotate polarized light.

Chapter 52 | Anisotropic Media 724
A transparent, crystalline material can change the polarization of light, without chirality.

Chapter 53 | Energy in Media 736
Printed text appears doubled when viewed through a crystal of calcite.

[[Photonic Bandgap]] 747
x.

Chapter 54 | Waves in a Cold Plasma and Faraday’s Rotation Effect 753
When viewed with a polarizer, the synthesized image of a black hole shows fine structure in its polarization.

[[Metamaterials]] 763
x.

PART VIII  Plus Ultra

Chapter 55 | Vista: Field Quantization 768
Insects and crustaceans can “see” the polarization state of light.
xii

[[Einstein’s Gravitation]] 785

[[Yang-Mills Theory]] 787

Epilogue 789

Appendix A | Units and Dimensional Analysis 796
Appendix B | Global List of Symbols 801
Appendix C | Numerical Values 813
Appendix D | Matrix Refresher 815
Appendix E | Animated graphics 816
Appendix F | Formulas 817
Appendix G | Glossary 818

Contents Index Notation Glossary November 14, 2023
Detailed Contents

Problems 3
Web resources xxxvii
To the student xxxviii
To the instructor xliiv

PART I Preliminaries

Prologue 2
0.1 In Their Glory 2
  0.1.1 The Maxwell equations 2
  0.1.2 The Lorentz force law 3
  0.1.3 In words and a picture 3
0.2 Explanation of Symbols 4
  0.2.1 3-vectors 4
  0.2.2 Right-hand rules and the Levi-Civita symbol 5
  0.2.3 The Kronecker symbol 7
0.3 Mathematical Miscellany 7
  0.3.1 Streamlines 7
  0.3.2 Index conventions 8
  0.3.3 Divergence theorem 9
  0.3.4 Stokes theorem 9
  0.3.5 Some useful vector identities 10
  0.3.6 Euler theorem 10
  0.3.7 Angle and solid angle 10
  0.3.8 Delta “function” 11
0.4 What Lies Ahead 12
  0.4.1 Einstein’s critique 12
  0.4.2 Some more hanging questions 13
Track 2 15
  0.1.2’ The notion of point charge 15
Problems 16

Chapter 1 Warmup: Newtonian Gravitation 17
1.1 Framing: Interplay 17
1.2 Space carries a physical field called the newtonian potential 17
1.3 An immobile point mass yields the familiar 1/r potential 18
1.4 Newton’s law unifies celestial, terrestrial, and even laboratory measurements 20
1.5 Extended objects can be handled by combining fundamental solutions 20
1.6 Two remarks 21
PART II  Static and Almost-Static

Chapter 2  Electrostatics Introduced 27
2.1 Framing: Coequal Partners 27
2.2 Rephrase in Terms of a Potential 28
2.2.1 A static electric field can be re-expressed via an integrability lemma 28
2.2.2 Force law 30
2.2.3 An integrability lemma underlies the success of the potential method 30
2.3 Basic Solutions 30
2.3.1 Point charge 30
2.3.2 Continuous charge distribution 31
2.4 Conductors 31
2.5 Upcoming 32
2.5.1 Reality of electric field 32
2.5.2 Quasi-static 33
2.5.3 Beyond static 33
2.6 Problems 34

Chapter 3  Electrostatic Multipole Expansion 37
3.1 Framing: Distillation 37
3.2 The Electrostatic Multipole Expansion 37
3.3 Some Taylor Expansions 39
3.4 Derivation of the Formula 40
3.5 Multipole Moments Organize the Features of a Distribution According to Importance 40
3.6 More Remarks 41
3.6.1 Summary so far 41
3.6.2 From potentials to fields 41
3.6.3 Apparent singularity 41
3.6.4 All moments after the first nonzero one depend on choice of base point 42
3.6.5 Spherical distributions 42
3.6.6 Symmetry may dictate that some moments equal zero 43
3.6.7 Pure dipole is an idealization arising as a limiting case 44
3.6.8 The quadrupole moment may be uniaxial or biaxial 44
3.7 Force and Torque on a Fixed Charge Distribution 44
3.7.1 Potential energy depends both on position and on orientation 44
3.7.2 Force and torque arise as derivatives of potential energy 46
3.7.3 Several intermolecular forces are dipolar in character 47
3.7.4 Dipole moment can be induced by an external field 48
3.8 Plus Ultra 48
Chapter 7 | Vista: Electrohydrostatics 99

7.1 Framing: An Impossible Shape 99
7.2 Some Geometry of Curves and Surfaces 100
    7.2.1 Curves in a plane can be characterized by a single curvature function 100
    7.2.2 Mechanical equilibrium of an interface in a plane 102
    7.2.3 Surfaces in space have two distinct curvature functions 103
    7.2.4 The Young–Laplace formula describes a trade-off between surface tension and pressure 105
7.3 Effect of Electric Field 106
    7.3.1 An electric field jump across an interface modifies the energy balance 106
    7.3.2 The modified mechanical equilibrium admits a conical point solution 107
7.4 Technological applications 108
7.5 Plus Ultra 108
    7.5.1 A look ahead 109
    7.5.2 Other physical surfaces 109
    7.5.3 A glimpse of general relativity 109
Track 2 111
    7.2.3 Metric and second fundamental form 111
    7.2.4 Derivations of variational formulas 111
Problems 113

Chapter 8 | Charge Flux, Continuity Equation, and Ohmic Conductors 115

8.1 Framing: Conservation 115
8.2 A Graphical Argument for the 1D Case 115
8.3 Two or More Dimensions 116
    8.3.1 Any local conservation rule leads to a continuity equation 116
    8.3.2 The continuity equation bridges local and global conservation 118
8.4 Remarks 118
8.5 Nonstatic Situations 118
    8.5.1 Conductivity, resistivity, conductance, resistance 119
    8.5.2 Salt water conducts electricity via the motions of ions 119
    8.5.3 Conductivity depends on temperature 120
8.6 Quasi-static Approximation 121
8.7 Electroencephalogram/Electrocardiogram 121
    8.7.1 A steady current source in solution again leads to a Poisson equation 121
    8.7.2 An isolated neuron creates an exterior potential 122
    8.7.3 Electroencephalogram 125
    8.7.4 Electrocardiogram 126

Contents Index Notation Glossary November 14, 2023
14.3 Components of the Levi-Civita Tensor 211
14.3.1 The components of $\varepsilon$ are the same in any right-handed cartesian system 211
14.3.2 Components only specify a unique $\varepsilon$ after a right-hand convention is chosen 212
14.3.3 Ne Plus Ultra 213
14.4 Connect to Familiar Things 213
14.4.1 Dot product 213
14.4.2 Cross product 213
14.5 Useful Identities 214
14.5.1 Swap dot and cross 214
14.5.2 Triple cross product 214
14.6 Plus Ultra 215
Track 2 217
14.3' Spatial inversion invariance 217
14.4' Twisted tensors 217
Problems 220

Chapter 15 Magnetostatics 221
15.1 Framing: Integrability 221
15.2 A New Force Awakens 222
15.3 Vector Potential 223
15.3.1 No scalar potential this time 223
15.3.2 Lemma to a lemma 224
15.3.3 Revisit electrostatics 224
15.3.4 The magnetic Gauss law expresses an integrability condition 225
15.3.5 The Poincaré lemma applies in any number of dimensions, and to tensors of any rank 226
15.4 Gauge Invariance and Coulomb Gauge 227
15.5 Back to Physics 227
15.5.1 Steady currents 227
15.5.2 Axial symmetry suggests a solution to the Ørsted problem 228
15.5.3 The same Green function from electrostatics also solves the magnetostatic equations 229
15.5.4 Self-consistency 229
15.5.5 Some of the equations are automatically satisfied, resolving a counting puzzle 230
15.6 Biot–Savart Formula 230
15.6.1 Second solution to Ørsted, via vector potential 230
15.6.2 $\vec{B}$ for a general current distribution 231
15.6.3 More about thin wire approximation 231
15.7 Boundary Conditions 232
15.8 Magnetoencephalography 233
15.9 Plus Ultra 233
Track 2 235
15.2' Puzzle about angular momentum conservation 235
15.9'a About magnetic monopoles 235
15.9'b Elimination of $\vec{B}$ 235
15.9'c Differential forms 236
Problems 237

Chapter 16 Units and Dimensional Analysis 241
16.1 Framing: Communication 241
16.2  Base Units in Mechanics and Thermodynamics  241
16.3  Units in Electrodynamics  242
  16.3.1  A new dimension  243
  16.3.2  The SI base unit of charge is the coulomb  243
  16.3.3  Derived SI units  244
16.4  The gaussian system  245
  16.4.1  First point: The gaussian base unit of charge is the statcoulomb  245
  16.4.2  Second point: The gaussian system uses a modified $\vec{B}$ field  245
  16.4.3  Third point: The gaussian system eliminates charge dimensions  246
  16.4.4  What is an “esu”?  247
16.5  Remarks  248

Track 2  249
  16.3.2'  Why base the SI on the proton?  249
  16.4.1'a  Planck units  249
  16.4.1'b  Other schemes  249

Problems  250

Chapter 17  |  Magnetostatic Multipole Expansion  252
17.1  Framing: Distillation Again  252
17.2  Tensor Preliminaries  252
17.3  Far Fields of a Steady, Localized Current Distribution  253
  17.3.1  The magnetic dipole vector potential is the leading term in a series expansion  253
  17.3.2  A familiar example  254
17.4  Higher Moments  254
  17.4.1  The magnetic quadrupole fields fall faster than the dipole  254
  17.4.2  All moments after the first nonzero one are basepoint-dependent  256
17.5  Magnetic Dipole in an External Field  256
  17.5.1  Force and torque on a dipole  256
  17.5.2  Diamagnetism, paramagnetism, ferromagnetism  259
  17.5.3  Purification of oxygen via magnetic forces  259
  17.5.4  Magnetic levitation of macroscopic objects at room temperature  259

Track 2  261
  17.5.2'  Higher multipole polarization  261

Problems  262

PART III  Dynamic

Chapter 18  |  Beyond Statics  267
18.1  Framing: Self-consistency  267
18.2  Review  267
  18.2.1  Field equations  267
  18.2.2  A coil carrying constant current  267
18.3  Slowly time-varying Currents  269
  18.3.1  Faraday observed an $\vec{E}$ field associated to a time-varying magnetic field  269
  18.3.2  Work must be done to increase current through a solenoid  270
  18.3.3  Self-inductance also affects signal propagation along a cable  272
  18.3.4  Magnetic field energy is proportional to volume  272
18.4  Maxwell’s Modification to Ampère’s Law  273

Contents Index Notation Glossary November 14, 2023
18.4.1 Mathematical consistency hinges on the continuity equation for charge 273
18.4.2 Boundary conditions 274
18.5 Wave Solutions 275
18.5.1 About traveling plane waves 275
18.5.2 The final form of the vacuum Maxwell equations have plane wave solutions 276
18.6 EM Waves Resemble Light 277
18.7 Complex Exponential Notation for Waves 278
18.7.1 Electric Gauss law 279
18.7.2 Faraday law 279
18.7.3 Magnetic Gauss law 280
18.7.4 Ampère law 280
18.7.5 Traveling wave with attenuation 280
18.7.6 Summary 281
18.8 Potentials Beyond Statics 281
18.8.1 and can still be represented by using potentials 281
18.8.2 Gauge invariance and Coulomb gauge also extend beyond statics 281
18.8.3 Coulomb gauge can be augmented if the charge density is zero 282
18.9 Waves via Potentials 283
18.10 Complex Polarizations 284
18.10.1 Linear, circular, elliptical 284
18.10.2 Helicity basis for circular polarization 285
18.11 Plus Ultra 285
18.11.1 Spherical waves foreshadowed 285
18.11.2 Dielectric media 285
18.11.3 Necessity of the field concept 286
Track 2 287
18.4.1’a Connection to ohmic materials 287
18.4.1’b Stumbling yet pulled forward 287
18.6’a On the speed of light 288
18.11’ On the guidance of mathematical consistency 289
Problems 290

Chapter 19 | [[Vista: Waveguides]] 296
Problems 296

[[Vista: Johnson noise]] 298

Chapter 20 | First Look at Energy and Momentum Transport by Waves 299
20.1 Framing: Pressure 299
20.2 Linear Polarization 299
20.2.1 Electromagnetic waves transport energy 299
20.2.2 Although momentum is a vector, its transport in a wave does not time-average to zero 300
20.2.3 Radiation pressure underpins many electromagnetic phenomena 301
20.3 Light Cannot Be Interpreted As a Stream of Newtonian Particles 302
20.4 Circular and Elliptical Polarizations 302
20.5 Electromagnetic Waves can Also Transport Angular Momentum 303
Track 2 304
20.2.2’ [[Ponderomotive force and the Paul trap]] 304
Chapter 21 | Geometrical Optics and the Eikonal 307

21.1 Framing: Locally-Plane Waves 307
21.2 Light has Plane Wave Solutions in a Uniform Medium 308
21.3 Piecewise-uniform Medium 308
  21.3.1 The law of reflection 308
  21.3.2 The law of refraction arises from matching fields across a planar boundary 309
  21.3.3 Geometrical optics extends these exact results to approximate rules for non-planar boundaries 311
  21.3.4 Optical tweezers exploit the momentum transfer implied by refraction 312
  21.3.5 Spherical aberration limits the focusing power of simple lenses 312
  21.3.6 Dispersion from a triangular prism 313
  21.3.7 Chromatic aberration 314
  21.3.8 Total internal reflection arises when the law of refraction cannot be satisfied 314
  21.3.9 Interfering-ray optics: Michelson interferometer 315
21.4 Rainbows and Other Caustics 316
  21.4.1 Millennia of confusion 316
  21.4.2 Natural versus contrived focusing 316
  21.4.3 Rainbows as caustics 317
  21.4.4 Primary and secondary bows 320
  21.4.5 Thomas Young's critique of geometrical optics results 321
21.5 Gradient-index Medium 322
  21.5.1 Rays of light can be regarded as streamlines of energy flux 322
  21.5.2 Locally-plane waves are a useful idealization when there is a separation of length scales 323
  21.5.3 The eikonal equation controls propagation of a locally-plane wave 323
  21.5.4 Rays in vacuum 325
  21.5.5 Rays bend continuously in a gradient-index medium 326
  21.5.6 Shortwave radio skip (skywave transmission) 327
21.6 More Phenomena 328
  21.6.1 Mirages rely on our brains' assumptions about light propagation 328
  21.6.2 A gradient-index lens can minimize spherical aberration 329
  21.6.3 Gravitational fields can bend light rays even in vacuum 329
21.7 Plus Ultra: From Geometrical Optics to Wave Mechanics 331
  21.5.4' Fermat's principle 333
  21.7' Eikonal approach to classical mechanics, and to wave mechanics 334

Problems 335

Chapter 22 | Diffraction and Semiclassical Approximation 340

22.1 Framing: Stationary-phase Points and Their Bifurcations 340
22.2 Knife-Edge Diffraction 340
  22.2.1 Sommerfeld's exact solution reproduces classic diffractive features 341
  22.2.2 The solution is expressed in terms of Fresnel integrals 342
22.3 Stationary-phase Approximation of Oscillatory Integrals 342
  22.3.1 The Fresnel integral illustrates the stationary-phase principle 342
  22.3.2 Extensions 345
  22.3.3 Breakdown of stationary-phase approximation 346
22.4 Semiclassical Approximation to Scalar Diffraction 346
  22.4.1 Overview 346

Contents Index Notation Glossary November 14, 2023
22.4.2 Vector waves resemble scalar waves at large distances 347
22.4.3 The Huygens principle 347
22.4.4 Knife-edge diffraction and its disappearing stationary-phase path 351
22.4.5 Qualitative origin of diffusive phenomena 352

Track 2 354
22.2.1 Lamb's exact solution of knife-edge diffraction 354
22.3.1 Another approach to the Fresnel integral 358
22.3.2 Corrections to stationary-phase approximation 359
22.4.1 About the term “semiclassical” 359
22.4.4 Vista: Physical optics 360

Problems 361

Chapter 23 | Diffraction at Caustics 362
23.1 Framing: Superstationarity 362
23.2 The Semiclassical Rainbow 363
23.2.1 Somewhere under the rainbow 363
23.2.2 Earlier results rederived as stationary-phase applied to the Huygens integral 363
23.2.3 Stationary-phase paths can merge and annihilate as parameters change 367
23.2.4 A semiclassical calculation addresses the shortcomings of geometrical optics 368
23.2.5 The second-order variation of phase degenerates at the caustic 368
23.2.6 Extension of stationary phase to a function of several variables 370
23.2.7 Reduction to a single integral 371
23.2.8 Other simplifications 372
23.2.9 Expectations and results 373
23.2.10 Comparison with experiment 374
23.2.11 Compare and contrast with knife-edge diffraction 375
23.3 Lens Focusing, Point Caustics, and the Rayleigh Criterion 375
23.3.1 Geometrical optics of 2D lens focusing 375
23.3.2 Diffraction blurs the point focus 377
23.3.3 Extension to a circular lens 378
23.4 Plus Ultra 379
Track 2 381
23.2.2 Catastrophe theory 381
23.2.6 Reduction of the rainbow integral to a cubic phase function 381
23.2.10 High-tech rainbow calculations 381

Problems 382

Chapter 24 | Partial Polarization 385
24.1 Framing: Stokes Parameters 385
24.2 Light as an Ensemble 385
24.2.1 Most sources give chaotic light 385
24.2.2 Optical instruments ultimately measure energy deposition 385
24.2.3 Steady sources: Replace time average by ensemble average 387
24.3 Some Convenient Models of Light 387
24.3.1 Fully polarized light corresponds to the periphery of the Poincaré sphere 387
24.3.2 Simplified model of unpolarized light 388
24.3.3 Partial polarization 388
24.4 How to Measure the Stokes Parameters 389

Problems 389
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>27.6</td>
<td>Continuity Equations for Energy and Momentum</td>
<td>424</td>
</tr>
<tr>
<td>27.6.1</td>
<td>Energy and momentum each have local expressions for their density and flux</td>
<td>424</td>
</tr>
<tr>
<td>27.6.2</td>
<td>Energy and momentum densities of a traveling wave</td>
<td>426</td>
</tr>
<tr>
<td>27.7</td>
<td>Plus Ultra</td>
<td>426</td>
</tr>
<tr>
<td></td>
<td>Problems</td>
<td>426</td>
</tr>
<tr>
<td>Chapter 28</td>
<td>Einstein’s Version of Relativity: Overview</td>
<td>427</td>
</tr>
<tr>
<td>28.1</td>
<td>Framing: Conservative Revolution</td>
<td>427</td>
</tr>
<tr>
<td>28.2</td>
<td>The æther Hypothesis</td>
<td>427</td>
</tr>
<tr>
<td>28.3</td>
<td>The No-æther Hypothesis</td>
<td>429</td>
</tr>
<tr>
<td>28.3.1</td>
<td>The vacuum is a unique state</td>
<td>429</td>
</tr>
<tr>
<td>28.3.2</td>
<td>Follow the symmetry</td>
<td>430</td>
</tr>
<tr>
<td>28.4</td>
<td>Where We are Heading</td>
<td>430</td>
</tr>
<tr>
<td>Track 2</td>
<td>432</td>
<td></td>
</tr>
<tr>
<td>28.2’a</td>
<td>Michelson–Morley 1887</td>
<td>432</td>
</tr>
<tr>
<td>28.2’b</td>
<td>More about uniqueness of the vacuum state</td>
<td>436</td>
</tr>
<tr>
<td>28.2’c</td>
<td>In praise of æther</td>
<td>436</td>
</tr>
<tr>
<td>28.3’</td>
<td>Poincaré’s work</td>
<td>437</td>
</tr>
<tr>
<td></td>
<td>Problems</td>
<td>438</td>
</tr>
<tr>
<td>Chapter 29</td>
<td>Provisional Lorentz Transformations and the Fizeau Experiment</td>
<td>439</td>
</tr>
<tr>
<td>29.1</td>
<td>Framing: Dragging Light</td>
<td>439</td>
</tr>
<tr>
<td>29.2</td>
<td>Review</td>
<td>439</td>
</tr>
<tr>
<td>29.2.1</td>
<td>Galilean invariance predicts simple addition of velocities</td>
<td>439</td>
</tr>
<tr>
<td>29.2.2</td>
<td>æther skeptics have some explaining to do</td>
<td>440</td>
</tr>
<tr>
<td>29.3</td>
<td>Graphical Explorations Suggest a Form for Boost Transformations</td>
<td>441</td>
</tr>
<tr>
<td>29.4</td>
<td>The Wave Equation is Invariant Under Provisional Lorentz Transformations</td>
<td>442</td>
</tr>
<tr>
<td>29.4.1</td>
<td>Coordinate transformation</td>
<td>442</td>
</tr>
<tr>
<td>29.4.2</td>
<td>Active viewpoint</td>
<td>443</td>
</tr>
<tr>
<td>29.4.3</td>
<td>Passive viewpoint</td>
<td>443</td>
</tr>
<tr>
<td>29.5</td>
<td>Einstein’s Velocity Addition</td>
<td>444</td>
</tr>
<tr>
<td>29.6</td>
<td>A Nonnull, Falsifiable Prediction</td>
<td>445</td>
</tr>
<tr>
<td>29.6.1</td>
<td>Fizeau’s experiment</td>
<td>445</td>
</tr>
<tr>
<td>29.6.2</td>
<td>Galilean versus relativistic predictions</td>
<td>446</td>
</tr>
<tr>
<td>29.7</td>
<td>Plus ultra</td>
<td>448</td>
</tr>
<tr>
<td></td>
<td>Problems</td>
<td>450</td>
</tr>
<tr>
<td>Chapter 30</td>
<td>Aberration of Starlight and Doppler Effects</td>
<td>451</td>
</tr>
<tr>
<td>30.1</td>
<td>Framing: A Greedy Principle</td>
<td>451</td>
</tr>
<tr>
<td>30.2</td>
<td>Again No Dilation Invariance</td>
<td>452</td>
</tr>
<tr>
<td>30.3</td>
<td>Lorentz Transformations in One Space Dimension</td>
<td>453</td>
</tr>
<tr>
<td>30.3.1</td>
<td>A subgroup that excludes dilations</td>
<td>454</td>
</tr>
<tr>
<td>30.3.2</td>
<td>Rapidity parameter</td>
<td>455</td>
</tr>
<tr>
<td>30.4</td>
<td>A Typical Paradox and its Resolution</td>
<td>456</td>
</tr>
<tr>
<td>30.5</td>
<td>Lorentz Transformations in Three Space Dimensions</td>
<td>457</td>
</tr>
<tr>
<td>30.6</td>
<td>More Key Experiments: Aberration of Starlight and Doppler Shift</td>
<td>458</td>
</tr>
<tr>
<td>30.6.1</td>
<td>Light-speed trajectories change apparent direction but not their speed</td>
<td>458</td>
</tr>
</tbody>
</table>
Chapter 41  |  Beams: Gaussian, Vortex, and Bessel  619
41.1 Framing: Diffractive Spreading  619
41.2 Gaussian Beams  619
41.3 Optical-Vortex Beams Transport Angular Momentum Even When Linearly Polarized  622
41.4 Bessel Beams  623
41.4.1 An idealized solution with no diffractive spreading at all  623
41.4.2 A physically realizable approximation to the ideal  623
41.4.3 Application to microscopy  626
Problems  630

Chapter 42  |  Vista: J. J. Thomson’s Pictorial Explanation of Radiation  633
42.1 Framing: Kinks  633
42.2 Electric Fields From a Suddenly Accelerated Charge  633
42.3 Magnetic Fields Also Have a Radiation Contribution  636
Problems  638

Chapter 43  |  Electric Dipole Radiation  641
43.1 Framing: Double Expansion  641
43.2 Three Length Scales  641
43.3 Electric Dipole Radiation  644
43.4 The Electric and Magnetic Fields Fall Slowly With Distance  645
43.5 Concrete Examples  646

Contents Index Notation Glossary November 14, 2023
47.3 Scattering Cross Section and the Thomson Formula 679
47.4 Light Propagates Diffusively in a Stellar Interior 680
47.5 Polarized Incoming Light Retains its Polarization Upon Scattering 681
47.6 Unpolarized Incoming Light Acquires Partial Polarization 681
47.6.1 Selective scattering can create polarization 681
47.6.2 Polarization of the cosmic microwave background as a reporter for early-Universe conditions 681
47.7 The Case of Bound Charges 681
47.7.1 Rayleigh scattering cross section 681
47.7.2 The blue, polarized sky 682
47.7.3 Blue, polarized scattering from colloidal suspensions 682

Track 2 683
47.3’ The transition to Compton scattering 683

Problems 684

Chapter 48 | [[Vista: Scattering by Many Objects]] 685

[[Vista: Scattering by a Dielectric Sphere]] 686

PART VII  Light in Linear Media

Chapter 49 | Light in Simple Media 688
49.1 Framing: Evanescent Wave 688
49.2 Polarizable Media 688
49.2.1 The electric displacement combines \( \vec{\varepsilon} \) with a contribution from induced electric dipole moment 688
49.2.2 The \( \vec{H} \) field combines \( \vec{B} \) with a contribution from induced magnetic dipole moment 689
49.2.3 The new fields summarize the effects of bound charge 690
49.2.4 Boundary conditions 691
49.3 Linear Regime 692
49.3.1 In linear media, electric polarization effectively modifies \( \varepsilon_0 \) 692
49.3.2 In linear media, magnetization effectively modifies \( \mu_0 \) 693
49.3.3 Maxwell plus constitutive relations form a closed system 693
49.3.4 Macroscopic physical realizations 693
49.3.5 Remarks and further examples 694
49.4 “Total” Internal Reflection and its Evanescent Fields 695
49.4.1 At high angle of incidence, the fields emerging from a dielectric have complex wavenumber 695
49.4.2 Evanescent fields carry no energy to infinity 697
49.4.3 Evanescent fields can nevertheless excite fluorescence 698
49.4.4 Total internal reflection fluorescence microscopy 698
49.4.5 Frustrated total internal reflection 700

Track 2 701
49.2’ More realistic treatments of polarizable material 701
49.2.1’ Dissipation and frequency dependence 701
49.3’ More general response functions 702

Problems 703

[[Vista: Microscopy]] 705

Contents Index Notation Glossary November 14, 2023
Chapter 50 | Čerenkov Radiation 706
50.1 Framing: Bow Shock 706
50.2 In Vacuum, a Charged Particle has one Source Point in the Observer’s Past Light Cone 706
50.3 In a Dielectric Medium, There Can Be Two Source Points, or None, in the Observer’s Past Light Cone 706
50.4 Interpretation 708
50.5 Application: Particle Identification in Underground Detectors 709
Problems 710

Chapter 51 | Chiral Media 711
51.1 Framing: Cross-susceptibility 711
51.2 Circular Birefringence Seems to Present a Paradox 711
51.3 Cross-susceptibility 714
51.3.1 Macroscopic physical realizations give intuition about a possible new effect 714
51.3.2 The general linear response function has new cross-terms 715
51.4 The Origin of Circular Birefringence 716
51.4.1 Historical remarks 717
51.5 How to Observe Circular Birefringence 718
51.6 Extensions 718
51.7 Circular Dichroism 719
51.8 Plus Ultra 720
Track 2 721
51.5’a Just two enantiomers 721
51.5’b Relativistic formulation 721
Problems 723

Chapter 52 | Anisotropic Media 724
52.1 Framing: Molecular order 724
52.2 The Susceptibility Tensor Defines Principal Directions 724
52.3 Applications of Anisotropy 725
52.3.1 Half-wave plate 725
52.3.2 Linear dichroism 726
52.3.3 Imaging with polarized light 726
52.3.4 How to make a circular polarizing filter 727
52.3.5 E. Land’s optical ring sight 727
52.3.6 A medical application 729
52.4 Induced Birefringence 729
52.4.1 Electro-optic (Pockels or Kerr) effects 729
52.4.2 Liquid crystals 729
52.5 Optical Torque Wrench 732
Track 2 733
52.2’ Magnetic anisotropy 733
52.4.2’ Contrast to polar liquids 733
Problems 734

Chapter 53 | Energy in Media 736
55.4.3 Coherent states mimic classical states in the limit of large occupation numbers 775
55.5 Interaction with Electrons 776
55.5.1 Classical interactions involve adding source terms to the field equations 776
55.5.2 Electromagnetic interactions can be treated perturbatively 777
55.5.3 The dipole emission pattern 778
55.6 Vistas 779
55.6.1 Many invertebrates can detect the polarization of light 779
55.6.2 Invertebrate photoreceptors have a different morphology from vertebrates’ 780
55.6.3 Some transitions are far more probable than others 782
55.6.4 Lasers exploit a preference for emission into an already occupied state 783

[[Einstein’s Gravitation]] 785

[[Yang–Mills Theory]] 787
55.7 Atiyah-Hitchin-Drinfeld-Manin instanton 788

Epilogue 789

Acknowledgments 795

Appendix A | Units and Dimensional Analysis 796
A.1 Base Units 796
A.2 Dimensions Versus Units 797
A.2.1 Densities 798
A.2.2 Angles 798
A.2.3 Functions applied to dimensional quantities 799
A.2.4 Additional units 799
A.3 About Graphs 800
A.4 Payoff 800

Appendix B | Global List of Symbols 801
B.1 Mathematical Notation 801
B.2 Dimensions and Units 805
B.3 Named Quantities 806

Appendix C | Numerical Values 813
C.1 Fundamental Constants 813
C.2 Material Properties at low frequency 813
C.3 Properties at optical frequencies 813
C.3.1 Refractive index for visible light 813
C.4 Miscellaneous 814

Appendix D | Matrix Refresher 815

Appendix E | Animated graphics 816

Appendix F | Formulas 817
Web Resources

The book's Web site (https://www.physics.upenn.edu/~pcn/EMP/) contains links to the following resources:

- Student Resources/Media...[[Not ready]]
To the Student

One goal of this book is to help you teach yourself the foundations, working knowledge, and fluency in some core theory ideas that even the most hard-nosed experimentalist must know. Another goal is to help you teach yourself the foundations, working knowledge, and fluency in some key real-world phenomena that even the most abstruse theorist must know. My choices of what, precisely, constitute that dual core are what distinguish this treatment from the many others available.

This book assumes that you have already studied this subject at the upper undergraduate level. If you have not already seen some basics of special relativity, for example, then you may wish to use a more introductory text instead of, or as a supplement to, this one. The mathematical prerequisite is some experience with the first ideas in linear algebra and differential equations, such as what you have obtained on the fly in undergraduate physics courses.

Goals of this book

1. Organize, systematize, consolidate, integrate. In particular, systematize the notion of symmetry, and its connection to tensors and tensor calculus (what is the cross product really?). We’ll start in three dimensions, because most of us grew up in a (seemingly) 3D world. But then we’ll see great advantages when we bump vectors and tensors up to 4D. Also, you know there’s a relation between symmetry and conservation laws—Chapter 40 will make it precise.

2. Forge links to other areas of physics. Do problems that make contact with those fields instead of working in a hermetically sealed silo.

3. Reflect on “Where do good theories come from?” Electrodynamics is the gateway to all of current fundamental physics, for example, Yang-Mills theory and general relativity.

4. Survey some remarkable phenomena; develop applications. If you’re a PhD student, your #1 question may not be truth/beauty, but rather, “What will I do my dissertation on?” So I wish to offer vistas to current topics.

5. Strengthen problem-solving and computer skills. A PhD is about research, and in research you keep getting stuck. This book gives opportunities to develop the generically useful faculty of getting unstuck, but with more real-world problems than you may be accustomed to.
Features of this book

- There will be math, certainly. But a key feature of the book is the emphasis on the wide world of phenomena (hence the book’s title). Even if some of them are already familiar to you, it is now time to see how they interlock. Your understanding of physics deepens when you can draw analogies to a rich tapestry of phenomena. To give a sense of where we’re going, the brief Contents singles out one key phenomenon emblematic of each chapter; if you find any of these intriguing, now you know where they are discussed. Each chapter head also hints at a physical idea that will be used to understand that phenomenon.¹
- Readers may come from many backgrounds, so some terminology may be unfamiliar. If a term is not defined, is defined but too briefly, or was defined many pages ago, you may find that the Glossary (Appendix G) says something helpful. Similarly, if a named variable or math notation is unfamiliar, the global list of symbols (Appendix B) may list it. Also, Appendix D reviews some notation specifically about matrices.
- The notations “Equation x.y” and “Idea x.y” both refer to a single numbered series.
- Many chapters end with an optional appendix labeled “Track 2.” These are sidelights, not required for understanding the main text. Some of them assume more background than the main text. There are also Track 2 footnotes and problems, marked with the symbol $T$. They are for readers who already know the basics, so they may make forward references to the main text.
- The square root of minus one is indicated in roman type (i) to distinguish it from, say, an index. (Some software packages instead refer to this quantity as I or as j.) The base of natural logarithms is indicated in roman type (e) to distinguish it from the charge on a proton (e). The differential operation is indicated in roman type (d) to distinguish it from any variable called d, which could denote a distance. Again, Appendix B summarizes other mathematical notation and lists key symbols that are used consistently throughout the book.
- Units appear in sans-serif font, dimensions in blackboard-bold. This way, you can visually distinguish between m (meters), M (dimension of mass), and m (a variable that could denote a particular object’s mass, or an integer index, and so on). In handwriting, I personally can’t do a distinct sans font, so I sometimes find it helpful not to use standard abbreviations for units to avoid confusion (that is, I write “meters,” “sec,” and “coul” instead of m, s and C). In fact, even in this book coulombs are written as coul.
- Errata will appear on the book’s web site.

Some uncomfortable questions

I might as well mention some unmentionable topics, because you are surely thinking of them.

¹What is a “physical” idea? The edges are fuzzy, but basically it’s an idea that has proven useful for understanding many seemingly different physical phenomena.
• “Why take this course a fourth time?”2 One reason is that now that you’ve taken many other physics classes, we can integrate electrodynamics with other areas. For example, now that you’ve studied statistical physics, we get to apply its insights, extending the practical reach of electrodynamics. (Recall the long title of this book.) Finally, some basic topics must be developed because they are prototypes for interesting extensions.3
• “I’m not planning to work on applications X and Y.” When I defended my own dissertation, I had no inkling that my research directions would change completely a few years later; neither do you know your future in detail. I was glad that some people had required me to pick up a lot of general physics background. The applications developed here were chosen because they seemed interesting; they are approachable, even though not always seeming so at first; and they sometimes require building up core skills like data visualization that are portable across fields.4
• “Some of this material isn’t really classical electrodynamics.” Indeed, some material here was chosen to illustrate how, at higher altitudes, we can see the various watersheds of physics merging into a connected network.
• “No physicist believes that classical physics is true.” Certainly visualizing light as a stream of tiny packets of energy is helpful for understanding how single molecules emit and absorb light. But some other simple phenomena, like what is in the space outside a permanent magnet, are not easy to describe in this way. To go beyond black-body radiation, we must invent a more detailed version of electrodynamics. This book will mainly discuss its classical limit, which for many advanced applications is a fantastically accurate approximation to the full quantum world and much simpler to handle. The coherent response of many electrons in an antenna to a coherent state of electromagnetic radiation is another example. We like simple theories not (just) because we’re lazy, but because with them, we can see farther without getting lost in formalism.

Ultimately, the complete picture does require that we quantize the theory. But Chapter 55 will show that the full structure of the classical theory is still needed as the first step. For example, polarization effects at the single photon level are mysterious if we naïvely think of photons as little marbles, yet they are important for understanding some modern microscopy techniques; they will emerge naturally when we quantize the full Maxwell theory. Prior to then, we’ll have our hands full with the many important Electromagnetic Phenomena that are adequately described by the classical approximation.

---

2Typically high school, first-year undergrad, third-year undergrad, and again now.
3For example, the familiar construction of a scalar potential from a curl-free vector field in Chapter 2 is the prototype for the Poincaré lemma developed in Chapter 15 and then is used again in Chapter 34 and Section 34.8.1’b (page 544).
4See Appendix E.
“The Facts”

The first principle [of science] is that you must not fool yourself, and you are the easiest person to fool. … After you’ve not fooled yourself, it’s easy not to fool other scientists. You just have to be honest in a conventional way after that.

— Richard Feynman

At some points you may wonder, “Why try so hard to convince me that the theory is true? Just tell me The Facts, so I can get on with becoming a scientist!” Actually,

*Some of today’s accepted theories are wrong, but we don’t know which.*

We all need finely honed critical skills. Studying past revolutions is useful to be ready for future revolutions.

In fact, science is a system of tools to prove that your wonderful new theory, which you love so much, is not true. Discovering that unfortunate fact is the first step to letting go and finding your next wonderful new theory, which may be true. When you find it, its truth may still not be clear to the world. It is instructive to see how classic theories gained the assent of a world that initially was opposed to them, by surviving tests that could have falsified them.

**Agile, fluent, compelling**

Some people like to say, “It’s not what you know; it’s who you know.” I don’t hear a lot of scientists saying that. I have heard physicists say, “It’s not what nor whom you know; it’s what you can do.”

I’ll add that a lot of success comes down to what you can do that’s never been done before. For that, you need the agility and fluency to get all the way to a goal without getting tangled up in the middle. If your instructor asks you to do a humble thing, and it’s easy—great. If it’s not easy, it’s an opportunity to build that agility and fluency.

It’s tempting to say, “I know that stuff; I’m just a little rusty.” Believe me, I too get rusty on anything after a surprisingly short time, and then for practical purposes I no longer really know it. For the rest of your scientific life, you may need to be removing that sort of rust. So get good at it. Eventually it does get easier.

Furthermore, a lot of success actually comes down to whom you can convince that your idea is correct, interesting, and important. Getting an idea all the way from your brain into another brain is hard. It starts with engaging the listener so that they actually pay attention. Every single assignment you turn in for a science class is an opportunity to improve this skill—so make sure your work tells a logical story. Communication can also be enhanced with computer-generated graphics, another research skill you will strengthen while working some of this book’s problems.

**On being right**

If you are wearied by this procedure, take pity on me, who carried it out at least 70 times.

— Kepler, on a difficult calculation
If you wish to study Nature, but you do inaccurate work, then you have accomplished nothing. Nature does not give partial credit. Nature does not care about your special circumstances. If you find this indifference appalling, try turning it around: Nature also does not care about your race, gender, or other identification; Nature does not privilege one category of people. If you do accurate work, Nature may offer you some secret not yet known to anyone else.

Sadly, everybody makes errors. But some people seem to make fewer errors, because they catch them. How does that work?

**Step 0:** Carry units everywhere (see Appendix A and Chapter 16). That’s really important, but just the start. What if your units are correct, but you dropped a term?

**Step 1:** Impose general reasonableness tests—features that the correct solution must have.

**Step 2:** You can lock your work in a drawer and do it over from scratch, then reconcile. That’s a good approach, but it won’t help if you’ve got a conceptual problem. And/or you can get symbolic software to carry out steps for you (same limitation). And/or you can collaborate, hoping that your collaborator will make a disjoint set of errors, then reconcile. And/or you can come to office hours and ask your instructor or teaching assistant. (But that stops working when the course is over.)

These steps will take you a long way, but you also need the secret weapon, the most powerful of the Rings of Power. You need to

**Step 3:** Identify limiting cases in which the answer is obvious, or at least doesn’t require computer math, or is available from some independent authoritative source. Specialize your answer to such a case and reconcile if necessary.

For example, suppose that you are asked to compute the near- and far-fields of an oscillating dipole. Work hard, but then specialize your answer to the limiting case of a static dipole and compare to the answer you know. Next, work up to considering just the far fields of an oscillating dipole and compare to your physical expectations, and so on.

**On being wrong**

As a student, you get told many times that you said or wrote something wrong. It can get discouraging. But let me offer a viewpoint.

Only in a field where you can be, and often are, objectively wrong can you ever be objectively right! Only in such a discipline can your factual rightness alone convince a skeptic, overturning their initial opinion. Only in such a discipline can you do that regardless of the mighty institutional authority of your skeptic, the politics of their tribal affiliations, and other factors that enforce groupthink in certain other disciplines. It is true that social behavior plays a role in the most seismic scientific revolutions, so be aware of that. But at the daily level, correct calculations and experiments generally win out rapidly as they get replicated by others. If you were wrong this time, you can learn how to be right next time. Study examples of this process, both on the grand and the daily levels, so you’ll be ready when it’s Your Turn.
Sparks from the anvil

11 May. Hard at work on Maxwellian electromagnetics.
13 May. Nothing but electromagnetics.
16 May. Worked on electromagnetics all day.
17 July. Depressed; could not get on with anything.
24 July. Did not feel like working.
7 August. Saw from Ries’s book that most of what I have found so far is already known.

— From the Diary of Heinrich Hertz, 1884

If it becomes hard, take heart from Hertz’s struggles. Everything worth doing is hard at first. Every physicist has a story of bottoming-out at some point.\(^5\) It never gets easier, but if you keep the fire on, eventually the kettle will boil, even if nothing seems to be happening at first. Later, you get to remember the previous difficulties and how you overcame them. Ask for help, and don’t wait until just before an exam or due date.

On computers

Many problems request that you evaluate and display results with a computer. Various platforms are available. Your instructor may require you to use one in particular, but if not, I have found that Python is freely available and can do everything requested, and that developing your skills with Python pays dividends in physics and beyond. On the few occasions when the text mentions Python specifically, I follow the widespread convention that \texttt{plt} is short for the module \texttt{matplotlib.pyplot} and \texttt{np} is short for the module \texttt{numpy}. Concerning animation, see Appendix E.

Other books

Here is a tiny subset of the available books on this subject. Many others are cited at the ends of chapters. No two sources use exactly the same units and notation, so beware. [[[Not ready]]]

\textit{Plus Ultra}—

—means “more beyond.” Let’s get started.

\footnote{One of mine involved spinor algebra.}
To the Instructor

The ultimate importance of the Maxwell theory is far greater than its immediate achievement in explaining and unifying the phenomena of electricity and magnetism. Its ultimate importance is to be the prototype for all the great triumphs of twentieth-century physics. It is the prototype for Einstein’s theories of relativity, for quantum mechanics, for the Yang–Mills theory of generalised gauge invariance, and for the Standard Model of particle physics.

— Freeman Dyson

It may well seem that yet another treatment of classical electrodynamics is needed about as much as a revised table of logarithms, so I should begin with a brief explanation.

Certainly the course can feel like a diversion from the truly urgent business of a first-year graduate student, which is often to find a research group. So I decided to present it as a series of applications relevant to every area of current research in my department. I also want to open some peepholes to more advanced theories, both for students who will study them later, and even more for those who will not.

At my institution, this course is also the last occasion where experimental and theoretical PhD students (and adventurous undergraduates) sit in the same room thinking the same thoughts. So I knew that the course was our last chance to cover topics that I wanted each group to know about the others’ intellectual worlds, but which many had not yet seen. I have tried to make this activity seem worthwhile, even enjoyable; if I succeed, then the habit may persist beyond the end of the course.

The class is also an opportunity, for some students the first one, for the systematic development of a mode of thinking that is unique to Physics, that is, placing symmetry principles front and center and letting them lead us as we work through the interplay between experimental observations and their theoretical distillations. By developing the world’s first relativistic field theory, we set the stage for the generalization to the other theories currently accepted as fundamental (see the epigraph to this section). Thus, another background agenda for this course is for students to emerge fluent in tensors and their uses. My approach is to introduce tensors informally via familiar examples, then gradually build up to the more general setting. Once students have mastered tensors, they gain access to a wealth of literature in areas like liquid crystals, cosmology, analytical mechanics, and even supersymmetry and superstrings—not just electrodynamics.

1 Sometimes years later they come back together on faculty committees, but by then it may be too late to initiate the habit of talking about science.
Finally, I believe that students need a rich Wunderkammer of key phenomena, or an armature on which to construct their future minds. You’ll find an unusually large number of such examples called out in the margins and chapter headers.

Format

My department offers a one-semester course on electrodynamics, for PhD students who have taken a full-year intermediate-level course as undergraduates. The opening chapters do review some elementary topics, but I present them mainly to grease the rails, so that later, more advanced chapters can proceed more smoothly by analogy to the earlier discussions.

Any book embodies its author’s vision of where to place emphasis, where to go slowly or fast, what to repeat, and so on, and this one is no exception. If I seem to belabor some points, keep in mind that I have repeated or amplified them based on experience with hundreds of real, 21st-century students.

Also, when I was a student, I found books and courses glossed over certain conceptual points, perhaps implicitly saying, “For some students, this will be intuitively obvious, and the others won’t need it.” I personally wish that some of those things had been spelled out for me, and I offer my best attempt here. You can assign material like Chapter 32, or not, as you see fit.

The applications alluded to in the title of this book are central to its message. Instead of being relegated to a separate section, they often interrupt the conceptual development. This is intentional: By presenting an application as soon as we have the needed framework, we make the discussion concrete and illustrate the utility of otherwise dry abstractions. Thus, a lot of physiology becomes available as soon as we have the quasi-static approximation, without requiring the full baggage of electrodynamics. If the applications seem to distract from the “real business” of your course, then this book is probably not the resource you want.

Instructors who teach a one-semester course, including me, will need to be judicious about how many applications to cover. (Two semesters or even three quarters would be needed to discuss them all.) Even so, there is value to having the whole smorgasbord available here: I like to cover a different subset each year. Also, I want students to have a resource that they can go back to when they get interested in one of the topics, which describes it in notation consistent with what they saw in the course.

Condensing the core material to a single semester has also required many difficult trade-offs, for example, a cursory treatment of magnetic materials. Other topics have been omitted or minimized with less regret, for example, the customary, mysterious formulas for div and curl in curvilinear coordinates. Again, every instructor will have their own priorities; if mine don’t make sense to you, another book may be a better fit.

I also assume that students have had some exposure to basic tools from multivariable calculus (including the Gauss and Stokes theorems) and linear algebra (including determi-
nants and the principal axes of a symmetric matrix). Some familiarity with other branches of Physics at the undergraduate level (mechanics, statistical physics, high-energy physics) will also be needed to appreciate some of the applications, which display the vast reach of electrodynamics. However, background from other fields (such as physical chemistry and physiology) is introduced in a brisk but self-contained way.

Conventions

- SI units are used throughout, in part to emphasize that physics is an experimental science. However, Chapter 16 offers a detailed discussion of gaussian units, because older literature often uses them. Also, learning a foreign language always improves your understanding of your mother tongue.
- The metric tensor uses the \((- + + +\) convention, in part for consistency with the student’s later study of general relativity and (some) quantum field theory books.
- Vectors and tensors in four dimensions are indicated by underbars. Although this notation looks busy on the page, it is easy to produce in handwriting, unlike boldface type.
- When a complex quantity is used to convey amplitude and phase of a real sinusoidal field, this is always written in full for clarity: Thus, \(\frac{1}{2}(\hat{A}e^{-i\omega t} + \text{c.c.})\) is always written in full, not abbreviated to just \(\hat{A}e^{-i\omega t}\).

Specific features

I begin by stating the Maxwell equations at the beginning, considering them as a single physical hypothesis whose meaning and implications we wish to study. The approach is then to specialize to the simplest situations, then gradually work up to greater generality.

Once we have seen some successful predictions, we can then explore the invariances of the equations, at first informally and then by constructing 4-tensor notation. As Arnold Sommerfeld put it,\(^3\)

\[ \text{The path taken by Einstein in 1905 was steep and difficult…. The path which we shall take is wide and effortless. It proceeds from the universal validity of the Maxwell equations… and it ends almost inadvertently at the Lorentz transformation and all its relativistic consequences.} \]

Although I might quibble with “effortless,” still I believe that every student needs to see relativity developed in this way. The apparatus of tensor analysis, initially introduced in this book as a notational aid to organizing the multipole expansion, later brings many benefits for understanding relativity.

I offer many problems at all levels, both at the ends of chapters and embedded into each one. Many involve creating computer-generated graphics, which I believe is a core skill for any quantitative career. Keep in mind that computer work is very time intensive—some of these are more like projects than regular problems—so you can only assign a few in a semester. They can be done with whatever computing platform you prefer.\(^4\) For many

\(^3\)Sommerfeld, 1964a.
\(^4\)For introductions to one free, open-source language, see Lin et al., 2022; Guttag, 2021; Hill, 2020; Pine, 2019.
more problems, try Pierrus, 2017.

**Missing but perhaps not missed**

Although this book spans hundreds of pages, many mathematical methods are hardly mentioned, for example, Bessel functions, spherical harmonics, even the Fourier and Laplace transforms. What you will find here is a lot of physics, specifically, a lot of Electromagnetic Phenomena. Mathematical ideas have been ruthlessly suppressed in Track 1, unless I have seen them being used in research to help understand physical phenomena, and believe that they will be useful to students for future applications. (Mathematical ideas that do meet those criteria have been mercilessly included.)

**On rigor**

Harish-Chandra to Dyson: “I am leaving physics for mathematics. I find physics messy, unrigorous, elusive.” Dyson’s reply: “I am leaving mathematics for physics for exactly the same reasons.”

— recalled by Mark Kac

I make no claim to mathematical rigor. I have tried to present a framework that corresponds to how at least some practicing physicists really think about the material. When I sometimes introduce machinery that seems heavier than is needed for the task at hand, it is usually laying the groundwork for more involved situations that will appear later in a student’s education. Even then, proofs that this machinery really has all the properties claimed are often suppressed. (In some cases, Track 2 sections give pointers.)

**Pathways**

[[Not ready]]

**Other disclaimers**

This is a textbook, not a monograph.

- No claim is made to originality. Although there are many gambits and avenues taken that are not easy to find in books for this audience, most are probably just inferior versions of some existing work. However, many of the problems are original; solving them taught me a lot.
- No attempt is made at scholarly attribution of priority. The experiments discussed were chosen because they seemed to embody conceptual points that I wanted to make—not because they were more important than the experiments not discussed.
- This is not a work on history. I mention historical points mainly because I find it inspiring how the founders accomplished so much with so little.

---

Kinder & Nelson, 2021 gives a brisk treatment specifically adapted to problems in this book.
I trust that the younger public may find these romances of interest, and that here and there one of the older generation may recapture an ancient thrill.

— Sir Arthur Conan Doyle, preface to a collection

It is a pleasure to recall brilliant courses on electrodynamics that I heard from Robert Zacur, Gary Sanders, and David Wilkinson. I hope that, while adding a lot of material that was not in their courses, I have remained true to their ideals of clarity and physical intuition. If my tone ever veers over from dignified to exuberant, well, some of these ideas made me insanely happy when I first heard them.
Electromagnetic Phenomena
In 1774, N. Maskelyne found that the gravitational attraction of Mount Schiehallion displaced a hanging mass by $12\text{arcsec}$.
Prologue

If one asks the older generation about the writings of Maxwell, there appears in their eyes something like the reflection of the love of their youth; but at the same time, they admit that especially Maxwell’s *Treatise* was a kind of intellectual jungle, almost impenetrable in its uncultivated fertility.

— Paul Ehrenfest, 1923

0.1 IN THEIR GLORY

You have already encountered the basic equations of electrodynamics, and the symbols in which they are formulated, in previous classes. This short chapter will mostly just establish some notation. Later chapters will:

- Motivate the form of each equation based on simple Electromagnetic Phenomena;
- Explore less simple phenomena that can be understood on the basis of these equations;
- Reformulate them in ways that for some purposes are more powerful; and
- Extend their reach by incorporating some idealized forms of macroscopic media.

0.1.1 The Maxwell equations

Maxwell did not write them in this form. Each equation is named for somebody prior to Maxwell; besides systematizing everything, Maxwell also made a crucial modification to “Ampère’s” law (Chapter 18).

\[
\vec{V} \cdot \vec{E} = \rho / \varepsilon_0 \quad \text{electric Gauss} \quad (0.1)
\]

\[
\vec{V} \cdot \vec{B} = 0 \quad \text{magnetic Gauss} \quad (0.2)
\]

\[
\vec{V} \times \vec{E} + \frac{\partial \vec{B}}{\partial t} = \vec{0} \quad \text{Faraday} \quad (0.3)
\]

\[
\vec{V} \times \vec{B} - \mu_0 \varepsilon_0 \frac{\partial \vec{E}}{\partial t} = \mu_0 \vec{J} \quad \text{Ampère} \quad (0.4)
\]

These equations can be solved for the fields \( \vec{E} \) and \( \vec{B} \) if we know the motions of charged particles.

---

1We owe today’s formulation to O. Heaviside and J. W. Gibbs. Although Einstein’s original relativity paper came later, he and others still used different names for each cartesian component, today considered horrible.
0.1 In Their Glory

The constants have numerical values \( \mu_0 \approx 4\pi \cdot 10^{-7} \text{ m kg coul}^{-2} \) (the magnetic permeability of vacuum), and \( \varepsilon_0 \approx 8.85 \cdot 10^{-12} \text{ coul}^2\text{N}^{-1}\text{m}^{-2} \) (the electric permittivity of vacuum).\(^2\)

Later chapters will define the charge density \( \rho_q \) and charge flux \( \vec{j} \) in terms of the positions and motions of charged particles.\(^3\)

The official name for \( \vec{E} \) is “electric field intensity”; \( \vec{B} \) is the “magnetic induction.” We’ll just call them the electric and magnetic fields, or even more succinctly “\( \vec{E} \) and \( \vec{B} \)” fields. Some formulas are neater when expressed in terms of a related quantity that we’ll call \( \vec{B} \equiv c\vec{B} \), because this quantity has the same dimensions as \( \vec{E} \).

0.1.2 The Lorentz force law

Reciprocally, the Lorentz force law describes the possible motions of an isolated, charged point under the influence of external fields:

\[
\frac{d}{dt} \vec{p} = q \left( \vec{E} + \vec{v} \times \vec{B} \right) + \vec{f}_{\text{other}}. \tag{0.5}
\]

This time, \( d/dt \) represents the ordinary time derivative along a particle’s trajectory. The fields \( \vec{E}, \vec{B} \) are to be evaluated at some time \( t \) and at the position \( \vec{r}(t) \) of the particle at that time; \( \vec{v} = d\vec{r}/dt \) at that time. \( q \) and \( m \) are constants called charge and mass that completely characterize a point charge.\(^4\) \( \vec{f}_{\text{other}} \) represents any non-electromagnetic force acting on the charged bodies in the system.\(^5\) And the momentum \( \vec{p}(t) = m\vec{v}(t) \), at least for speeds much smaller\(^6\) than \( 10^8 \text{ m/s} \).

A test body refers to a limiting case of a point object with charge and mass infinitesimally small, but \( q/m \) a finite constant. In practice, a test body is a point charge so small that does not significantly perturb surrounding fields set up by other charges.

Section 0.1.2’ (page 15) discusses the notion of “charged point particle.”

0.1.3 In words and a picture

Figure 0.1 symbolizes the strategy. In words:

- The electric Gauss law says, “Charges give rise to electric fields with some constant of proportionality \( 1/\varepsilon_0 \).”
- Chapter 36 will interpret the magnetic Gauss law as saying, “No point sources nor sinks for magnetic field lines.”

\(^2\)Chapter 16 will discuss units and explain why the value of \( \mu_0 \) stopped being exact in 2019.

\(^3\)See Section 8.3. Some authors call \( \vec{j} \) the “current density,” but that term risks confusion; we will use “density” to mean only volume density (a quantity per m\(^3\)). When a quantity per area or per length must be discussed, we will specify “areal density” or “linear density,” respectively.

\(^4\)In particular, we will later see that a point charge is also an object whose multipole moments apart from \( q \) are immaterial in a particular situation.

\(^5\)Sometimes it’s appropriate to instead introduce a constraint. For example, we can imagine a situation in which a static charge is fixed onto on a spinning disk.

\(^6\)Chapter 31 will generalize this relation.
The Faraday law says, “Time-dependent magnetic fields also give rise to electric fields.”

The Ampère law says, “Currents give rise to magnetic fields with some constant of proportionality \( \mu_0 \). Time-dependent electric fields also give rise to magnetic fields.”

The Lorentz force law says, “A charged particle experiences a position-dependent electric force per charge, as well as a position and velocity dependent magnetic force per charge. The latter force is always directed perpendicular to the velocity.”

### 0.2 EXPLANATION OF SYMBOLS

#### 0.2.1 3-vectors

A point in 3-space can be specified by choosing a “good” coordinate system (in particular, a cartesian system\(^7\)) and stating its components:

\[
\vec{r}_i = \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \quad i = 1, 2, 3, \quad \text{or} \quad \vec{r} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}.
\] (0.6)

The symbol \( \vec{r} \) represents an abstract geometric object (an arrow), but with an index it means a set of three numbers, called \( \vec{r}_1 = x \), \( \vec{r}_2 = y \), and \( \vec{r}_3 = z \). The index-free notation \( [\vec{r}] \) is an abbreviation for those numbers regarded as a column (3 × 1 matrix). Again, an index on a 3-vector indicates that only one of its components (an ordinary number) is meant. Most authors drop the over-arrow when explicitly writing the index on a vector, but in this book we will retain it for clarity. Thus, \( \vec{r}_1 \) is a single number, but it’s not a scalar; it is a component of a vector.\(^8\) This book won’t use the 3-vector notation \( \vec{r}^1 \) (upper index).\(^9\)

Other quantities with an over-arrow are understood to be triples of numbers with the same transformation under rotation of the spatial axes as \( \vec{r} \), that is, 3-vectors. The 3-scalar product (also called dot product) is \( \vec{a} \cdot \vec{b} = \sum_{i=1}^{3} \vec{a}_i \vec{b}_i = [\vec{a}]^T[\vec{b}] \). We denote

---

\(^7\)Thus, curvilinear coordinates such as spherical polar are not “good” in this sense. Why make this restriction? For now, our answer is, “Because these are the coordinate systems in which the Maxwell equations look nice, and we’re studying the Maxwell equations.” Section 14.2 will consider how the representation of a vector changes when we switch from one “good” system to another.

\(^8\)Later, we will sometimes append a sub- or superscript in parentheses to the name of a vector. Such labels don’t refer to a component; they indicate which one of a set of related vectors is meant (see for example Section 1.2).

\(^9\)Such notation may, however, be useful when dealing with curvilinear coordinates. Later, when we define 4-vectors, Chapter 32 will introduce an upper-index notation, which is distinct from lower indices even when we use cartesian coordinates.
\( \vec{r} \cdot \vec{r} \) by \( \| \vec{r} \|^2, r^2 \), or simply \( r^2 \); that is, \( r \equiv \sqrt{r^2} \). Section 14.4 (page 213) reviews why \( \hat{a} \cdot \hat{b} = \| \hat{a} \| \| \hat{b} \| \cos \theta \), where \( \theta \) is the angle between the vectors.

The vector \( \hat{r} = \vec{r} / r \) has length equal to one. More generally, a circumflex (“hat”) instead of an over-arrow implies that a vector has been normalized, that is, divided by its length to convert it to a unit vector. Some standard unit vectors include the coordinate-axis directions \( \hat{x}, \hat{y}, \hat{z} \). Some books call them \( \hat{i}, \hat{j}, \text{ and } \hat{k} \).

The components of a vector field, such as \( \{ \vec{E}_i \} \), are themselves functions on spacetime, that is, \( \vec{E}_i(t, \vec{r}) \) and so on. We differentiate them with the vector of operators:

\[
\vec{\nabla}_i = \left[ \begin{array}{c} \partial / \partial x \\ \partial / \partial y \\ \partial / \partial z \end{array} \right].
\]

The dot product of \( \vec{V} \) with itself is the Laplace operator (or laplacian), written as \( \nabla^2 \).

The dot product of \( \vec{V} \) acting on a vector field is called the divergence operator and denoted \( \vec{V} \cdot \vec{V} \). Note that \( \vec{V} \cdot \vec{V} \) is an ordinary function, whereas \( \vec{V} \cdot \vec{V} \) is an operator that acts on whatever sits to its right and does not involve any derivatives of \( \vec{V} \). In fact, \( (\vec{V} \cdot \vec{V}) f \) is the directional derivative of \( f \) along \( \vec{V} \).

If \( \vec{r}(t) \) is a trajectory parameterized as a function of time, then the 3-velocity is \( \vec{\omega} = d\vec{r} / dt \).

### 0.2.2 Right-hand rules and the Levi-Civita symbol

The two best things in Italy are spaghetti and [Tullio] Levi-Civita.

— Einstein

To finish defining the symbols in Equations 0.3–0.5, suppose that we have chosen a convention for “right hand.” This is the same thing as selecting a reference coordinate system on space whose unit vectors \( \hat{x}, \hat{y}, \text{ and } \hat{z} \) are mutually perpendicular. To see the equivalence, note that with such a choice made, we can say which of your hands should be called “right” by the following procedure (Figure 0.2):

- Hold one hand flat with the fingers initially pointing along \( \hat{x} \).
- Orient the hand so that when you bend your fingers by 90 degrees they now point along \( \hat{y} \).
- If with that orientation, your thumb is pointing along \( \hat{z} \), then that hand will be called “right” according to that coordinate system. If your thumb is pointing along \( -\hat{z} \), then that hand will be called “left.”

Conversely, we could instead start by choosing one particular hand (for example, the one farthest from the heart of a normal human\(^{12}\)), and use it to classify coordinate systems as “right handed” or not.

---

\(^{10}\) Again, one can also set up a curvilinear coordinate system for expanding vectors, and find corresponding vector differential operators, but we’ll rarely use such systems: We are constructing tensor analysis on flat spaces, usually in the restricted class of cartesian coordinate systems.

\(^{11}\) Mathematicians use the symbol \( \Delta \) for the laplacian, but most physicists don’t. It’s too easy to confuse that with \( \Delta \), the physicists’ symbol for a change in some quantity.

\(^{12}\) Less anthropocentrically, we could use the helical structure of the DNA of any (terrestrial) organism.
Figure 0.2: A right hand with respect to the ordered triad of unit vectors $\hat{x}, \hat{y}, \hat{z}$ shown. Equivalently, if we begin by declaring this hand to be “right” then $\hat{x}, \hat{y},$ and $\hat{z}$ shown (in that order) constitute a right-handed coordinate basis.

Figure 0.3: Structure of the Levi-Civita symbol. (a) The 27 numbers that make up the Levi-Civita symbol $\varepsilon_{ijk}$, represented as a stack of balls. Three entries are $+1$ (green balls), three are $-1$ (red balls), and 21 are zero (transparent gray balls). (b) Exploded view detailing individual layers.

The vector operators in Section 0.1 are then defined by their usual formulas in any right-handed, cartesian coordinate system. For example, the cross product can be expressed by saying that $\vec{a} \times \vec{b} = \hat{c} ||\vec{a}|| ||\vec{b}|| \sin |\theta|$, where $\theta$ is the angle between $\vec{a}$ and $\vec{b}$ and $\hat{c}$ is a unit vector perpendicular to each of them. There are two such unit vectors; we choose the one for which $\hat{a}, \hat{b},$ and $\hat{c}$ form a right-handed triad in the sense of Figure 0.2.\textsuperscript{13}

There is an equivalent formulation of the cross product that will be helpful throughout this book, via the formula

$$(\vec{a} \times \vec{b})_k = \sum_{j,k=1}^{3} \varepsilon_{ijk} \vec{a}_j \vec{b}_k.$$ \hspace{1cm} (0.7)

\textsuperscript{13}If $\vec{a}$ and $\vec{b}$ are parallel or antiparallel, then the choice of $\hat{c}$ is ambiguous—but in that case $\sin |\theta| = 0$, so the ambiguity doesn’t matter.
The formula involves the **3D Levi-Civita symbol** \( \varepsilon_{ijk} \), which is shorthand for 27 numbers \((3 \times 3 \times 3)\). Most of those entries equal zero: \( \varepsilon_{ijk} = 0 \) if any two of the indices match, for example, \( \varepsilon_{112} \). If all three indices have different values, then they must be a permutation of 1, 2, 3; \( \varepsilon_{ijk} \) is then defined by using the parity of that permutation. Thus, \( \varepsilon_{123} = +1 \), \( \varepsilon_{231} = +1 \), \( \varepsilon_{132} = -1 \) and so on (Figure 0.3).

The entries \( \tilde{a}_j \) and \( \tilde{b}_k \) in Equation 0.7 refer to the components of the vectors in any right-handed coordinate system, and the formula yields the components of the resulting vector in that same system.

We are not ready yet to prove that Equation 0.7 is independent of which right-handed coordinate system we chose, and indeed equivalent to the preceding geometric definition. But you can readily generate some evidence:

**Your Turn 0A**

a. Use Equation 0.7 to show that \( \tilde{a} \times \tilde{a} = 0 \) for any vector, in agreement with the geometric definition.

b. The geometric definition clearly depends on which hand we declare to be “right.” Show that Equation 0.7 also has this (undesirable) feature. [Hint: Let \( u = x, v = y \), and \( w = -z \), and construct the corresponding unit vectors. Then a vector \( \tilde{a} \) will have components with \( \tilde{a}_1' = \tilde{a}_1, \tilde{a}_2' = \tilde{a}_2, \) and \( \tilde{a}_3' = -\tilde{a}_3 \). Writing \( \times' \) for the alternate version, we find \( (\tilde{a} \times' \tilde{b})_3' = \tilde{a}_1' \tilde{b}_2' - \tilde{a}_2' \tilde{b}_1' \) and so on. Are these the primed components of the vector \( \tilde{a} \times \tilde{b} \) defined in the usual way?]

One advantage of the algebraic formulation, Equation 0.7, is that it will show us how, and in what sense, we may generalize the cross product to spaces with more than three dimensions. The cross product of \( \tilde{V} \) acting on a vector field \( \tilde{V} \) is a new vector field called the **curl** of \( \tilde{V} \), denoted \( \tilde{V} \times \tilde{V} \). It enters in the Faraday and Ampère laws.

### 0.2.3 The Kronecker symbol

There’s also the more familiar **Kronecker symbol** \( \delta_{ij} \), which is defined to be +1 if \( i = j \) and 0 otherwise.

### 0.3 MATHEMATICAL MISCELLANY

#### 0.3.1 Streamlines

Think for a moment about a steady flow of water. At any point in a flow, there is a local average velocity \( \bar{u}(\vec{r}) \). We can ask about the **streamlines** of this vector field. The
streamlines are curves in space that are everywhere tangent to $\mathbf{v}$ (Figure 0.4). No individual water molecule will literally follow a streamline, due to its random Brownian motion; nevertheless, the streamlines give a good impression of what is going on. A small but macroscopic tracer particle suspended in the water really will follow a streamline. We will be more interested in the streamlines of the electric and magnetic fields, which we’ll call electric and magnetic field lines.

The preceding paragraph seems to suggest that field lines have a mathematical meaning but no physical reality; for example, we cannot say that a field line runs through any particular point in space. Still, we will see that they are a useful means for visualizing the fields themselves, which we do regard as real. Moreover, Michael Faraday based much of his powerful intuition on the idea that each field line exerted tension along its length, but a form of lateral pressure on neighboring field lines (somewhat like the individual strands of a stretched, aligned sample of a polymer material). Chapter 36 will explain the physical foundation for that imagery in terms of the energy–momentum flux tensor of the electromagnetic field. Moreover, like the twitching of a spider’s web, Faraday imagined disturbances in field lines as communicating motions of charged objects to their neighbors. Chapter 42 will construct a modern version of this idea, which is helpful for understanding radiation.

### 0.3.2 Index conventions

From now on we will employ the **summation convention**: When a vector index appears exactly twice in a product of factors, we mean for it to be summed over all its values, even if we don’t explicitly write the summation symbol. Thus, we abbreviate Equation 0.7 as

$$(\mathbf{a} \times \mathbf{b})_i = \epsilon_{ijk} \mathbf{a}_j \mathbf{b}_k.$$

A summed index is also called a **dummy index**.

If an index appears just once in such an expression, it’s called a **loose index** and is not summed. An expression with one or more loose indices really means several expressions, one for each set of loose index values.

When two or more terms are added, the summation convention applies to each term separately. Thus, in an expression like

$$\epsilon_{ijk} \mathbf{a}_j \mathbf{b}_k + \epsilon_{imn} \mathbf{c}_m \mathbf{d}_n,$$

the first term is summed over $j$ and $k$, the second is summed over $m$ and $n$, but there is no summation over $i$ even though
it appears twice. Instead, \( i \) is an overall loose index for the complete expression.

A loose index on one side of an equation must match a loose index on the other side (unless the other side is zero, in which case we mean that it’s zero for each value of the index).

A summed pair of indices must each be named with the same letter of the alphabet. We can rename them both if we like, as long as they still agree with each other. When we combine formulas, we sometimes need to rename some index pairs in this way, to avoid ambiguity. Thus, the product of \( \hat{a}_i \hat{b}_i \) times \( \vec{c}_j \hat{d}_j \) should be rewritten \( \hat{a}_i \hat{b}_i \vec{c}_j \hat{d}_j \) (or \( (\vec{a} \cdot \hat{b}) (\vec{c} \cdot \hat{d}) \)).

Two crucial theorems from vector calculus are both beefed-up versions of the Fundamental Theorem of Calculus. Please get reacquainted with these formulas, and with the specific conventions they contain concerning choice of handedness:

### 0.3.3 Divergence theorem

\[
\int_V d^3r \ \vec{\nabla} \cdot \vec{E} = \int_{\partial V} d^2\Sigma \cdot \vec{E}. \tag{0.8}
\]

Here \( d^3r \) is a volume element.

\( V \) is a finite volume and \( \partial V \) denotes the closed surface bounding it. Thus, \( \partial V \) itself has no boundary. (For example, a solid doughnut has a surface, called a torus, that itself has no edge.) Any small element of \( \partial V \), called \( d^2\Sigma \), has two perpendicular directions (sometimes called normal vectors). The surface separates space into regions we may call “inside” and “outside,” so one of the normals is the “outward-pointing normal.” On the right side of Equation 0.8, the vector \( d^2\Sigma \) denotes the product of an area element \( d^2\Sigma \) times the unit outward-pointing normal.

### 0.3.4 Stokes theorem

\[
\int_{\Sigma} d^2\Sigma \cdot (\vec{\nabla} \times \vec{E}) = \oint_{\partial \Sigma} d\vec{c} \cdot \vec{E}. \tag{0.9}
\]

Here \( d\vec{c} \) is a vector line element. \( \Sigma \) is a surface (not necessarily closed), and \( \partial \Sigma \) is its boundary, if it has one (a closed curve in space). Thus, \( \partial \Sigma \) itself has no boundaries. (For example, the boundary of a disk is a circle, but that circle has no endpoints.)

An open patch of surface has no “inside/outside” distinction, so we may choose either face as “outward” when defining the sign of \( d^2\Sigma \). Then the line integral along \( \partial \Sigma \) must be traversed in the direction selected by applying a right-hand rule to our choice of perpendicular vector.\(^{17}\)

If the surface \( \Sigma \) is closed (no boundary), replace the right side of Equation 0.9 by zero.

\[^{17}\]Point the thumb of your right hand along the chosen direction, then traverse the boundary in the sense that follows the curve of your fingers.

---

Contents  Index  Notation  Glossary November 14, 2023
Your Turn 0B

a. Show that, if you instead make the opposite choice of “outward” direction for $\delta \Sigma$, then each side of Equation 0.9 changes sign, and the formula is still valid.
b. Similarly, show that if we change our convention for which hand is “right,” then again we get canceling minus signs on each side.

If $\vec{V} \times \vec{E} = \vec{0}$, we call $\vec{E}$ a **curl-free vector field**. Then Equation 0.9 says that its contour integral around any closed loop equals zero.

0.3.5 Some useful vector identities

Your Turn 0C

Use the product rule of calculus to show that:
a. $\vec{V} \cdot (f \vec{V}) = \vec{V} \cdot \vec{V} f + f \vec{V} \cdot \vec{V}$.
b. $\vec{V} \times (f \vec{V}) = (\vec{V} f) \times \vec{V} + f \vec{V} \times \vec{V}$.

Your Turn 0D

If the cartesian components $\vec{V}_i$ of a vector field each depend on position $\vec{r}$ only via its distance $r$ to the origin of coordinates, then show that

a. $\vec{V} \times \vec{V} = \vec{r} \times d\vec{V}_i/dr$, and

b. $\vec{r} \cdot (\vec{V} \times (\vec{r} \times \vec{V})) = -2\vec{r} \cdot \vec{V}$.

0.3.6 Euler theorem

When studying time-varying quantities, it’s useful to know that $e^{-i\omega t} = \cos(\omega t) - i \sin(\omega t)$. Thus, we can represent both sines and cosines in a unified way: Either can be written as $\frac{1}{2} [be^{-i\omega t} + \text{c.c.}]$, where “c.c.” stands for “complex conjugate.” If we choose $b = 1$, then this expression equals $\cos(\omega t)$; if we choose $b = i$, then it equals $\sin(\omega t)$. If $b$ is complex, we may write it as $|b|e^{i\alpha}$; then

$$\frac{1}{2} [be^{-i\omega t} + \text{c.c.}] = |b| \cos(\alpha - \omega t),$$

which still has frequency $\omega$ but is phase shifted relative to sine or cosine.

0.3.7 Angle and solid angle

A short line element $d\vec{r}$, seen from a great distance, subtends an angle $d\theta \to ||d\vec{r} \times \vec{r}||/r$, where $\vec{r}$ is the vector from the observer to the line element. This expression is dimensionless, so any unit of angle is also dimensionless (a pure number). For example, the “radian” rad is strictly speaking the number 1, but sometimes we state it just to emphasize that we are not using some other unit (such as mrad = 0.001 or deg = $\pi/180$).
Similarly, a small surface element $d^2\Sigma$, seen from a great distance, subtends a **solid angle**\(^\text{18}\) $d\Omega \to d^2\Sigma \cdot \hat{r}/r^2$. This expression is dimensionless, so any unit of solid angle is also dimensionless (a pure number). For example, the "steradian" $sr$ is strictly speaking the number 1, but sometimes we state it just to emphasize that we are not using some other unit (such as $m\text{sr} = 0.001$ or $\text{deg}^2 = (\pi/180)^2$).

Although we mostly use cartesian coordinates, for some problems it's preferable to label points in space via:

- Cylindrical coordinates, consisting of $z$ (distance along the cylinder axis, dimension L), $\rho$ (radius, or distance away from the cylinder axis, dimension L), and $\varphi$ (azimuthal angle in the plane perpendicular to cylinder axis, dimensionless).
- Spherical polar coordinates, consisting of $r$ (radius, or distance away from the origin, dimension L), $\theta$ (polar angle, tilt downward from the $+\hat{z}$ axis, dimensionless), and $\phi$ (azimuthal angle in the plane perpendicular to $\hat{z}$, dimensionless).

Thus, $\rho = \sqrt{x^2 + y^2}$ whereas $r = \sqrt{x^2 + y^2 + z^2}$. Both systems agree that $\tan \varphi = y/x$.

### 0.3.8 Delta “function”

Technically, the **delta function**\(^\text{19}\) is not an ordinary function at all: When $\delta(x)$ is integrated over $x$, it's a machine that eats an ordinary function and returns its value at zero:

$$\int dx \, \delta(x) f(x) = f(0).$$

Thus, the dimensions of $\delta(x)$ are always inverse to those of its argument $x$.

For our purposes, it will usually suffice to regard $\delta(x)$ informally as the limit of a bump function, for example $e^{-x^2/(2\sigma^2)}/(\sqrt{2\pi}\sigma)$, which becomes sharply peaked as $\sigma \to 0$, with the area under the curve fixed to 1. That viewpoint again makes it clear that the dimensions are inverse to those of $x$.

Section 34.9.1 will show that

$$\delta(f(x)) = \frac{1}{|f'(x_*)|} \delta(x - x_*).$$

Here we suppose that the function $f$ has one zero at $x_*$; if there's more than one, the right hand side becomes a sum of terms for each zero. In multiple variables, the denominator of the prefactor gets replaced by the absolute value of the determinant of the jacobian matrix.

---

\(^\text{18}\) A better name for this quantity might be **angular area**.

\(^\text{19}\) Sometimes called "Dirac delta function."
**0.4 WHAT LIES AHEAD**

[Einstein’s first relativity paper] says that the usual formulation of the law of induction contains an asymmetry which is artificial, and does not correspond to facts. According to observation, the current induced depends only on the relative motion of the conducting wire and the magnet, while the usual theory explains the effect in quite different terms according to whether the wire is at rest and the magnet moving or vice versa.

— Max Born

The Maxwell equations are two vector partial differential equations (PDEs), plus two more scalar PDEs. That’s a lot of complexity, even though the equations are linear. We will consider various reduced special cases before we start analyzing them in earnest, and some practical applications that can be understood using those simplified versions.

**0.4.1 Einstein’s critique**

If we know the equations, and accept that they are “true,” aren’t we done? Can’t we in principle just slap them on some big computer and find what they predict? In fact, it’s fair to say that nobody understood the real content of Maxwell (certainly not Maxwell himself), until Albert Einstein demonstrated a key hidden feature, an invariance property (or “symmetry”) that was there all along, buried in poor notation. Unfortunately, nobody understood Einstein, till Minkowski and successors found the appropriate generalization of vector notation to make this invariance manifest.²⁰

One point that everybody could understand,²¹ mentioned right at the start of Einstein’s first paper on relativity, concerned what happens when a bar magnet enters a coil of wire (Figure 0.5).

---

²⁰A good lesson: Teachers should, like Minkowski, stay interested in our former students’ work. By the way, how did Einstein get through peer review, if nobody understood him at first? It’s simple. At that time, peer review was: Planck was the journal editor. He read the manuscript, decided “I don’t understand it, but it looks good,” and that was that.

²¹For example, Heaviside had already noted this disconnect in 1885.
Every first-year physics student gets told how to explain the first of the two setups shown: Charges in the wire are free to move within it, but they are constrained not to leave it. When the wire is pushed to the left, as in Figure 0.5a, these charges are also carried leftward. The Lorentz force law (Equation 0.5) then predicts a force perpendicular to that motion and to \( \mathbf{E} \), so a charge initially in the plane of the page gets pushed out of the page, ultimately creating a current measured on the meter.

When the coil is stationary (Figure 0.5b), then its charge carriers are not required to move by the motionless wires containing them. In this case, however, the \( \mathbf{B} \) field is time-dependent. Faraday’s law (Equation 0.3) then implies an \( \mathbf{E} \) field, which can push charges that were initially at rest, again in the direction running along the wire. Again the meter responds.

Einstein said (paraphrasing), it’s crazy to offer two such totally different explanations of what is obviously just one phenomenon. After all, if you walk alongside the moving magnet, it appears stationary to you and the coil appears to move, and vice versa.²²

In fact, why should we even invoke a dynamical explanation (rooted in equations of motion) for this equivalence, which ought rather to be kinematic in character? It will take us a while to arrive at Einstein’s answer to this question, but for now, suffice to note that relativity was born out of frustration with electrodynamics. We will leave it as a Hanging Question:

**Hanging #A:** Can we eliminate the asymmetry between our explanations of the coil/magnet phenomena?

Einstein’s answer was “yes.” Eventually we’ll extend that answer to say: “We make full Lorentz symmetry manifest in the equations. Then \( \mathbf{E} \) and \( \mathbf{B} \) will be seen as two parts of a single object.”

In a moment of historic chutzpah, Einstein later went still farther. Paraphrasing again,

*Moreover, newtonian gravitation also lacks the invariance that was found to be hiding in electrodynamics; therefore newtonian gravitation is also wrong and must be modified.*

This was one of the most amazing examples of successful lateral thinking in the history of science,²³ so we’ll want to understand how to construct other relativistically invariant field theories, beyond electrodynamics.

### 0.4.2 Some more hanging questions

Section 0.4.1 raised a question that we won’t answer for some time. Here are several more. Keep them in mind as we work through to their resolutions.

**Hanging #B:** Why must the Maxwell equations have exactly that (arbitrary-looking) form, for example, the minus sign in Equation 0.4 but not in Equation 0.3?

---

²²Actually, at the time everybody other than Einstein would have agreed that he was crazy: “Obviously” the two situations were not equivalent, because at most one of them could be at rest with respect to the “luminiferous æther.” We’ll see later what Einstein said about that argument.

²³For another, see Section 21.7 (page 331).
Hanging #C: How can $\vec{E}$ and $\vec{B}$ be “two parts of a single object” when they appear in such non-parallel ways?

Hanging #D: How can we solve the eight Maxwell equations with only six unknown functions $\{\vec{E}_1, \vec{B}_1\}$?

Hanging #E: Our equations are full of cross products, which depend on an arbitrary choice of which is our “right” hand. Can we formulate electrodynamics in a way that doesn’t conceal its invariance under spatial inversions?

FURTHER READING/VIEWING

Semipopular:
This video on divergence and curl is incredibly good: www.youtube.com/watch?v=rB83DpBJqsE.

Intermediate:

---

24Section 0.2.2 proposed “the one farthest from the heart of a normal human,” but that isn’t very universal! Even “the one that describes DNA in all living organisms on Earth” is too Earth-centric to have fundamental significance. And “the one opposite to the helicity of a neutrino emitted in beta decay” goes outside the domain of electrodynamics.
0.1.2′ The notion of point charge

Equation 0.5 introduced the notion of “point charge” as an idealized situation, defined by having no relevant dynamical variables besides its trajectory \( \mathbf{r}(t) \) (for example, no orientation in space) and no relevant characteristics besides charge and mass (for example, no dipole moments). More precisely, if such higher moments are present, point-particle approximation assumes that their effects are negligible because the surrounding fields are slowly varying, just as we can ignore Earth’s mass quadrupole when we compute its orbit around the Sun.

Certainly it can be delicate to decide whether a real system may usefully be regarded as a point charge (or assembly of point charges). Indeed, in a strong enough field gradient, even a single electron cannot be regarded as a point charge, because it has a magnetic dipole moment! Similarly a neutron, although electrically neutral, can also be pushed by a magnetic field gradient, and so on. [[Not ready]]
0.1  All Greek to me
Here are the Greek letters most often used by scientists. The following list gives both lowercase and uppercase (but omits the uppercase when it looks just like a Roman letter):

\[ \alpha, \beta, \gamma/\Gamma, \delta/\Delta, \varepsilon, \zeta, \eta, \theta/\Theta, \kappa, \lambda/\Lambda, \mu, \nu, \xi/\Xi, \pi/\Pi, \rho, \sigma/\Sigma, \tau, \upsilon/\Upsilon, \phi (\text{sometimes written } \varphi)/\Phi, \chi, \psi/\Psi, \omega/\Omega. \]

When reading aloud we call them alpha, beta, gamma, delta, epsilon, zeta, eta, theta, kappa, lambda, mu, nu, xi (English speakers pronounce it “k’see”), pi, rho, sigma, tau, upsilon, phi, chi (pronounced “ky”), psi, omega. Don’t call them all “squiggle.” Sometimes we will use the variant form \( \varphi \) for phi and \( \theta \) for theta.

Practice by examining a quotation by D’Arcy Thompson: “Cell and tissue, shell and bone, leaf and flower, are so many portions of matter, and it is in obedience to the laws of physics that their particles have been moved, moulded, and conformed. They are no exception to the rule that \( \Theta\sigma\zeta \alpha\epsilon \gamma\epsilon\omega\mu\epsilon\tau\rho\epsilon \).” From the sounds made by each letter, can you guess what Thompson was trying to say? [Hint: \( \xi \) is an alternate form of \( \sigma \).]

0.2  By any other name
Fundamental constants can be expressed in whatever units are convenient for the problem at hand. Express the constant \( e^2/(4\pi\varepsilon_0) \) in the units MeV fm convenient for nuclear physics.
CHAPTER 1

Warmup: Newtonian Gravitation

No hammer in the horologe of Time peals through the Universe when there is a change from Era to Era.

— Thomas Carlyle

1.1 FRAMING: INTERPLAY

Newtonian gravitation isn’t the subject of this book, but it’s useful to introduce some themes with a scalar field theory before we move up to a vector field theory.

\[
\frac{d^2\vec{r}_\ell(t)}{dt^2} = -\nabla \phi_N(\vec{r}_\ell(t)) + \vec{f}_{\text{other}}/m_\ell
\]

The cartoon above, and an analogous one for electrodynamics, might playfully be called the “Central Dogma of classical physics.” It can be expressed as the slogan “fields tell particles how to move; particles tell fields what to be.” Let’s unpack that lapidary phrase to see how the interplay works.

Phenomenon: Unbound comets follow hyperbolic trajectories.

Physical idea: The equations of motion can be transformed to look like motion in a quadratic potential, a problem whose solution we know.

1.2 SPACE CARRIES A PHYSICAL FIELD CALLED THE NEWTONIAN POTENTIAL

The Newtonian potential \( \phi_N \) is a function that obeys

\[
\nabla^2 \phi_N = 4\pi G N \rho_m.
\]  

(1.1)

Here \( G_N \) is a universal constant of Nature and \( \rho_m \) is the mass density of matter.

We can think of matter as a collection of \( N \) point masses \( m_\ell \) following trajectories \( \vec{r}_\ell(t) \). Here the mass \( m_\ell \) is a constant characterizing particle number \( \ell \). In the notation \( \vec{r}_\ell(t) \), the particle number \( \ell \) appears in parentheses to avoid confusion with a vector index if we later wish to refer to one component \( \vec{r}_\ell(\ell) \).
With that notation understood, then we can finish specifying Equation 1.1 by constructing the mass density distribution as

$$\rho_m(t, \vec{r}) = \sum_{\ell} m_\ell \delta^{(3)}(\vec{r} - \vec{r}_{(\ell)}(t)). \tag{1.2}$$

In this formula, $\delta^{(3)}$ denotes the product of three delta functions. Notice the big distinction between $\vec{r}$ and $\vec{r}_{(\ell)}$: 

- $\vec{r}$ labels the point where we wish to evaluate $\rho_m$.
- The $3N$ functions of time, $\vec{r}_{(\ell)}(t)$, specify the $N$ particle trajectories.

Often it’s a good approximation to blur the many delta functions together. Then $\rho_m$ becomes a continuous function of position in Equation 1.1.

Solving Equation 1.1 gives us the newtonian potential function if we know what the mass distribution is doing. Conversely, Newton’s second law amounts to $3N$ equations of motion that tell what any point mass will do, given the potential:

$$\frac{d^2}{dt^2} \vec{r}_{(\ell)} = -\nabla \phi_N(\vec{r}_{(\ell)}(t), t) + \vec{f}_{\text{other}}/m_\ell. \tag{1.3}$$

Thus, we get a closed system of equations that, when solved together, tells us the future evolution of the system from initial conditions—the goal of classical physics.

The standard terminology is confusing: The “newtonian potential” is not the potential energy of any test particle. Instead, $\phi_N$ is potential energy of a test particle per unit mass.

The term $\vec{f}_{\text{other}}$ allows us to incorporate non-gravitational forces. Sometimes it’s an adequate approximation to instead introduce a constraint. Here the idea is that internal stresses supply whatever force is needed to maintain that constraint. For example, such stresses prevent the Earth from collapsing to a point, so that we may treat it as a fixed mass distribution. Other constraints ensure that the length of a pendulum remains constant, and so on.

But what is the potential “really?” Newton’s successors eventually gave up fiddling with vortices in the æther and other mechanistic explanations, and just said, “It’s a function on space and time, period. We don’t need a more explicit mechanical explanation to get on with making testable predictions. We don’t need to know if it’s really about quantum coherent states of gravitons, or condensates of superstrings…. All we need to do is tell how to measure it operationally. If every time anybody measures it they find that it obeys the equations, then they are useful equations.”

### 1.3 AN IMMOBILE POINT MASS YIELDS THE FAMILIAR $1/r$ POTENTIAL

If we are given a mass distribution, we can find the solution to Equation 1.1. But in the simplest situation, a point mass $M$, we can take a shortcut: It’s not hard to guess the answer and then check that it does solve the equation. Choose cartesian coordinates centered on
that mass ("the Earth"). We now confirm that the formula $\phi_N(t, \vec{r}) = -MG_N/r$ works, using steps that we'll need again and again in this book.

Equation 1.1 tells us to compute the laplacian of $\phi_N$, that is, the divergence of the gradient. Let's start with the gradient, and drop the prefactor $-MG_N$. So we want to find $\nabla (1/r)$, where $r = ||\vec{r}||$ is the length of the vector $\vec{r}$ from the point mass to the observer. The first component of the gradient is

$$\frac{\partial}{\partial x} (x^2 + y^2 + z^2)^{-1/2} = -\frac{1}{2}(x^2 + y^2 + z^2)^{-3/2} 2x = -x/r^3.$$ 

Notice that $1/r$ has units of inverse meters, as does $\nabla$, so it's right and proper that our answer has units of $m^{-2}$. Proceeding similarly with the other two components, and reinstating the constants, gives

$$\nabla \phi_N = (-MG_N)(-\vec{r}/r^3) = MG_N \hat{r}/r^2,$$

the familiar result for force per mass on a test body. Here $\hat{r} = \vec{r}/r$ is the unit vector pointing to $\vec{r}$.

Now we want to compute the divergence: $\nabla \cdot (\nabla r^{-1}) = -\nabla \cdot (\vec{r}/r^3)$. We use the Leibnitz property of derivatives ("product rule") to write this as

$$-r^{-3} \hat{r} \cdot \nabla - r^{-3} \cdot \nabla (r^{-3}). \quad (1.4)$$

The first term is easy because $\nabla \cdot \vec{r} = \frac{\partial x}{\partial x} + \frac{\partial y}{\partial y} + \frac{\partial z}{\partial z} = 3$. For the second term, adapt the previous result:

$$\nabla (x^2 + y^2 + z^2)^{-3/2} = -\frac{3}{2}(x^2 + y^2 + z^2)^{-5/2} \frac{2z}{2z} = -3\vec{r}/r^5.$$

So Equation 1.4 becomes $\nabla^2 (r^{-1}) = -3r^{-3} - \hat{r} \cdot (-3\vec{r}/r^5) = 0$.

Oops. We succeeded too well. We wanted the laplacian to vanish away from the point mass at the origin, but we seem to have proved instead that it vanishes everywhere. The problem is that everything we've done is invalid right at $r = 0$, where the potential function is singular. To handle that point, consider a spherical surface surrounding it and use the divergence theorem:

$$\int_{\text{surf}} d^2 \vec{S} \cdot \nabla (r^{-1}) = (4\pi r^2 \hat{r}) \cdot (-\hat{r}/r^2) = -4\pi.$$

So the integral of $\nabla^2 (-G_NM/r)$ over any spherical volume containing the origin is always $4\pi G_NM$, even though $\nabla^2 (-G_NM/r) = 0$ everywhere other than the origin. The same things can be said of $4\pi G_N \rho_m$ for a point mass (that is, $\rho_m(\vec{r}) = M\delta^{(3)}(\vec{r})$), so we see that the Newtonian potential (which gives rise to the Newtonian force) really does solve Equation 1.1 for a point mass.

---

1See Your Turn 0C (page 10).

2See Equation 0.8. To be a bit more precise, imagine the mass distribution not as a singular point, but spread over a very small volume. Take the spherical surface to lie outside this occupied region. Then the radius of that region drops out of the formulas, so we can take the limit where it, and the surface, shrink to zero size.
1. NEWTON’S LAW UNIFIES CELESTIAL, TERRESTRIAL, AND EVEN LABORATORY MEASUREMENTS

The $1/r$ potential gives the equation of motion for a test particle (that is, a mass so small that its effect on $M$ is negligible):

$$\frac{d^2}{dt^2} = -\nabla \phi_N(t, \vec{r}) = -MG_N \frac{\vec{r}}{r^2}. \quad (1.5)$$

That’s the familiar formula that gives rise to Kepler’s laws.

Just to find the simplest solution, recall that uniform circular motion has $d^2\vec{r}/dt^2 = -\omega^2\vec{r}$, where $\omega$ is the angular frequency. Taking the value of $\omega$ that corresponds to a month, and $r$ to be the Earth–Moon distance, and substituting into Equation 1.3 gives a rough numerical value for the quantity $G_N M_{\text{Earth}}$.

Newton also knew the acceleration of gravity for an object (for example, an apple) dropped near Earth’s surface. Knowing the radius of the Earth gave him another, independent estimate of $G_N M_{\text{Earth}}$. With historic understatement, Newton wrote around 1712 that these two estimates “answered pretty nearly.” That was the first grand unified theory—of celestial and terrestrial motions.

1.5 EXTENDED OBJECTS CAN BE HANDLED BY COMBINING FUNDAMENTAL SOLUTIONS

It’s true that we only found the solution to the field equation for a point mass, but perhaps surprisingly, that’s all we need. Because the field equation is a linear PDE, and also invariant under spatial translations, we can subdivide any complicated distribution of mass $\rho_m(t, \vec{r})$ into small chunks, apply the fundamental solution to each chunk, then use superposition to assemble all the sub-solutions into the full solution for $\phi_N$:

$$\phi_N(t, \vec{r}) = -G_N \int (\rho_m(t, \vec{r}_*) d^3r_*) ||\vec{r} - \vec{r}_*||^{-1}. \quad (1.6)$$

Later we’ll similarly exploit the linearity of the Maxwell equations, for similar wins. The fundamental solution that must be integrated against the mass density is called the Green function for whatever field equation we are studying. Later we’ll find simplified Green functions for electro- and magnetostatics, then a more elaborate one for the full Maxwell equations.

For example, the fact that Earth is not quite spherical is easy to incorporate into our assumed mass density function. Then we can solve the field equation and find the not-quite-spherical potential surrounding Earth, and from there the not-quite-Keplerian orbits of, say, spy satellites. Chapter 3 will develop this idea in the context of electrostatics.

But the expression in Equation 1.6 has a worrisome feature. If the field point $\vec{r}$ is inside a body, then we seem to have $1/0$ behavior when the source point $\vec{r}_*$ approaches the field

---

3See Problem 1.1. It’s not quite accurate to suppose the Earth stationary, because the Moon’s mass is not negligible. We can account for that by instead using the “reduced mass.”
point! And yet, when we drill a well into the Earth, we don’t see any such catastrophic behavior beneath the surface. To see what’s going on, consider for example evaluating $\Phi_N$ at the center of the Earth. The tricky region of the integral is the part close to that field point; we may suppose that mass density is constant in that region. Thus, we are worried about a possible divergence of the integral

$$\rho_n \int_0^{\text{small}} r_s^2 \, dr_s \, d(\cos \theta) \, d\varphi \, r_s^{-1}.$$ 

But the singularity is more than compensated by the $r_s^2$ from the volume element. The same argument can be used near any point in the interior of the body: If the mass density is a nonsingular function, then so will be the Newtonian potential.  

### 1.6 TWO REMARKS

1. Why introduce the potential function? Why not just work directly with the forces? One practical advantage is that the potential is a scalar. Combined with the preceding point, this means that we can conveniently integrate contributions from a complicated source (the nonspherical Earth, and so on), then at the very end compute the gradient, instead of having to carry around vector quantities throughout the calculation.

2. From this promising start, Newton and his successors proceeded to explain planetary motion, motions of moons around other planets, comet orbits, tides, the shape of the Earth, phase-locking of Mercury and of Earth’s Moon, precession of Earth’s axis, effects of Jupiter on other planets—a fantastic wealth of testable predictions from very few assumptions. 

Once the idea sank in that Nature was governed by laws, on Earth as it is in Heaven, the seeds were sown for the Enlightenment and all that entailed. Newton’s biggest fan in France was Voltaire, who thought that if Nature itself is subject to universal laws, not the whims of a supernatural being, then the divine rights of capricious kings looked absurd. But that is another story.

### 1.7 MORE HANGING QUESTIONS

Before we can claim that Equations 1.1–1.3 make testable predictions, we need to give meaning to all the quantities that they interrelate. Even the very coordinates $\hat{r} = (x, y, z)$ and $t$ require careful interpretation.  

---

4 Although we have disposed of the singularity issue in principle, in practice it can reappear when we attempt numerical evaluations; see Problem 1.4.

5 Notably D’Alembert, Clairaut, Euler and Laplace.

6 [[[Not ready]]]) will follow Einstein as he realized that the equations in this chapter, and in particular their invariances, are not quite correct. We are reviewing them because the cartoon at the start of this chapter is still a good way to think about more advanced theories.

7 Chapter 26 and following chapters will look more closely on coordinate choice.
Newton wrote some mumbo-jumbo about absolute space and time, but a more fruitful attitude emerged slowly. Today we say that what the equations are claiming is merely that there exists a way of labeling events by sets of four numbers, such that any motion of any set of masses, with any initial conditions, corresponds in the chosen system to functions of time that solve the equations.

This may sound like a big loss of predictive power—maybe there’s a physical motion that fails to satisfy the equations, but we could rescue them by merely relabeling the points! But even in this weakened form, the equations have the character of an interlocking web of many predictions: One single coordinate choice is supposed to handle any conceivable set of planets, comets, and apples, any initial conditions we may set on those masses, and so on.

Interestingly, and important for our later discussion, once we find one set of “good” coordinates on spacetime (that is, coordinates for which all phenomena obey the equations in their usual form), then there will also be other such “good” coordinate systems with the same property. You probably won’t be surprised to hear that rigidly shifting or rotating $x, y, z$ (leaving $t$ unchanged) gives a new “good system.” Also, shifting $t' = t + t_0$ works, and so does negating any one or more of $x, y, z, t$. Later, we’ll investigate just how big the set of “good” systems is. For now, we content ourselves with the statement that the content of newtonian physics includes the claim that at least one “good” coordinate system exists.

Certainly there will also be bad coordinate systems, in which the equations as written are not valid (just as with accelerating systems in newtonian physics). What is illuminating, however, is the transformations between the presumed good systems. Chapters 28–30 will describe how Albert Einstein found they were not what everybody had expected.

**Hanging #F:** What physically makes some coordinate systems “good” and others not?

**Hanging #G:** Can we introduce a potential function for electromagnetism analogous to the gravitational potential, and reap benefits analogous to the ones we got in that situation?

### 1.8 PLUS ULTRA

We're now in a position to understand the orbits of objects around the Sun, and specifically the unbound cometary orbit alluded to in Section 1.1. Problem 1.2 will lead you through that discovery; later, Problem 2.2 will show that the same derivation is relevant for an electrodynamics problem with medical significance.

---

8Henry Cavendish even designed a gravitational experiment that fits in a room.
$T_2$

1.7' XX
[[Not ready]]
1.1 The first grand unification
Repeat Newton’s early unification:

a. Look up the radius and period of the Moon’s orbit, calculate its acceleration, and estimate the quantity \( G_N M_{\text{Earth}} \). (Make the approximation that the orbit is circular. You can also ignore the reduced-mass effect, that is, make the approximation that the Moon is much less massive than Earth.)

b. Next, look up the Earth’s radius and again estimate \( G_N M_{\text{Earth}} \), this time based on the terrestrial acceleration of gravity. You can neglect the Earth’s rotation for this.

c. Compare the two values you found for \( G_N M_{\text{Earth}} \).

1.2 Flyby
The text claimed that the birth of Western science was when Newton solved the planetary orbit problem, deriving Kepler’s empirical observations as predictions. Newton then predicted the return of Halley’s Comet (among many other things). Because of the similarity between electrostatics and gravitation, we get to revisit this highlight, as it’s mathematically the same problem as one needed to understand proton therapy.\(^9\) In this problem, assume that everything is moving much more slowly than the speed of light; thus, you may use familiar newtonian mechanics.

A heavy object \( M \) sits at the origin of coordinates. We will neglect any perturbation to its position during this problem, because the other object in the collision, \( m \), is much lighter. The lighter object comes initially along a straight line parallel to the \( \hat{x} \) axis, moving from negative to positive \( x \). If it were not deflected by \( M \), the trajectory would pass within distance \( A \) of \( M \); that is, its initial trajectory is \( x(t) = v_0 t, y(t) = A \) when \( t \to -\infty \). Set up polar coordinates centered on \( M \), in which \( \varphi \) is measured clockwise from the \( -\hat{x} \) axis. Thus, the incoming body starts with \( \varphi = 0 \), and \( \varphi \) increases with time. If \( M \) were not present, then the trajectory would have \( \varphi \to \pi \) at \( t \to +\infty \).

a. Express the angular momentum of \( m \) about the origin, and the kinetic energy, both in terms of \( r(t) \) and \( \varphi(t) \). Use the constancy of the angular momentum to eliminate \( d\varphi/dt \) from the KE.

b. Write the potential energy as \(-K/r\). Thus, \( K = G_N M m \) for celestial mechanics. Find an equation that gives \( dr/d\varphi \), and hence determines the shape of the trajectory \( r(\varphi) \).

c. Change variables from \( r(t) \) to \( u(t) = r(t)^{-1} \); that is, derive an equation for \( du/d\varphi \). In a moment you will solve this equation for \( u(\varphi) \).

d. At time \( t \to -\infty \), we have \( u \to 0 \). Also work out the value of \( du/d\varphi \) at this time from the fact that initially mass \( m \) is moving in uniform, straight-line motion.

e. Solve the equation given the initial conditions. Determine the value of \( \varphi \) at which \( u(\varphi) \) stops increasing and turns around. Double this angle to find the total angular

---

\(^9\)See Problem 2.2. One can also argue that the birth of modern Physics was Geiger, Marsden, and Rutherford’s discovery of the atomic nucleus; this problem is relevant for that discovery as well.
deflection during the complete encounter.

1.3 2D field plot, I
Two stars are orbiting each other. One star’s mass is twice that of the other one. Choose axes such that at some moment, the stars are both on the x axis at x = ±a. Use a computer to create a contour graph of the newtonian potential in a suitable region of the xy plane (that is, the plane z = 0). Also get the computer to draw arrows on your graph representing the gradient of this potential.

1.4 2D field plot, II
A mass distribution takes the form of a rectangular prism with uniform mass density and edges of lengths a = b = 1 m, c = 5 m. Choose coordinates with z parallel to the long edge, x, y parallel to the short edges, and origin at the center of the object.

Use a computer to create a contour graph of the newtonian potential in an interesting region of the xz plane (that is, field points (X, Y, Z) with Y = 0), including both inside and outside of the body.

[Hints: (i) Make a grid of X, Z values where you wish to evaluate \( \phi_N \). For each point, you need to do three integrals, over the source point’s coordinates x, y, z. Of these, you can at least do the z integral analytically, for example by using a table of integrals. So do that first, because evaluating an analytic expression is generally more accurate than numerical integration. (ii) Now make a grid of x, y values and sum your partial integral over that grid. (iii) If your XZ grid has any points in common with your xy grid, then you’ll encounter 1/0 errors. Avoid them by shifting one of your grids relative to the other one. (iv) Finally, use Python’s \texttt{plt.contour} function or something similar.]

1.5 Below the surface
Section 1.5 (page 20) asserted that the apparent singularity in Equation 1.6 (page 20) is not a problem anywhere inside the body, but only verified this at one point. Imagine a sphere of radius R made of a material with uniform mass density \( \rho_m \). By symmetry, we only need to investigate the newtonian potential along the z axis: \( \vec{r} = (0, 0, z) \). Work out the newtonian potential for all z (inside and outside the body) by doing the integrals explicitly, and comment on relevant features.

Robert Hooke intuitively understood this result and communicated it to Newton around 1679.

\[\text{Earth’s gravitational acceleration decreases as we descend a deep mineshaft.}\]
“In the Middle Ages it had been the custom to ring peals of bells in the church steeples to disperse the thunder, as a result of which a high number of bell-ringers had been electrocuted. Indeed as late as 1786 the Parlement of Paris enforced an edict forbidding the practice, because over the previous thirty-three years no fewer than 103 unfortunates had been killed on the ends of their wet bell-ropes…. In 1708 Dr. Wall, in England, wrote that electricity “seems in some degree to represent thunder and lightning….” Concern soon focused on lightning strike and the danger it presented to gunpowder arsenals all over Europe. The row over exactly how to protect them began with the work of a hitherto obscure fifteenth child of a Bostonian soap-boiler, Benjamin Franklin…. He suggested that a church steeple be used to prove his theory. The Royal Society was not interested…. The explosion of an arsenal in Brescia, northern Italy, in 1769 made the rod a political issue. An estimated 175,000 pounds of powder exploded, destroying 190 houses within a radius of 639 feet from the explosion. The Brescia authorities asked the Royal Society for help in preventing a further disaster, and a committee was set up, of which Franklin was a member…. Conductors sprang up all over Europe. There was even a chapeau paratonnerre—an anti-lightning hat—for the ladies of Paris in 1778.” – James Burke (Burke, 2007)
CHAPTER 2

Electrostatics Introduced

Thoroughly conscious ignorance is the prelude to every real advance in science.

— James Clerk Maxwell

2.1 FRAMING: COEQUAL PARTNERS

The Maxwell equations simplify a lot if we consider a static, or nearly-static, situation. That is, all charges are either motionless or slowly moving.\(^1\) We will arrive at a system of equations of the form:

\[
 m_\ell \frac{d^2 \vec{r}(t)}{dt^2} = -q_\ell \vec{\nabla} \psi(\vec{r}(t), t) + \vec{f}_{\text{other}}
\]

\[
 \nabla^2 \psi = -\rho_q/\epsilon_0
\]

The cartoon above looks a lot like the one at the start of Chapter 1, but now each particle is characterized by two items of intrinsic information, called “mass” \(m\) (as before) and “charge” \(q\). The charge defines a corresponding density \(\rho_q\) analogous to Equation 1.2 (page 18).

The cartoon also carries a subtext: We will develop an approach where fields and particles are coequal in importance: Again, “fields tell particles how to move; particles tell fields what to be.” This chapter will mostly focus on the second part of that slogan.\(^2\)

Although the equations in the figure are in principle complete, later we will find it useful to modify them in ways that approximately treat complicated systems in simpler, tractable ways:

- In this chapter and the next, we imagine the “other” forces to be constraints, that is, whatever is required to keep the charges at rest. In that case, the distribution of charge is \textbf{static}, that is, unchanged upon either time shift or time reversal.\(^3\)
- Later, Chapter 6 will introduce dielectric media, containing molecules (distributions of charge that can distort slightly but that cannot separate altogether). Then the

---

\(^1\) Eventually we’ll say more precisely “slowly enough that we may neglect magnetic field effects.”

\(^2\) For the first part, see Problem 2.2.

\(^3\) In contrast, current flowing steadily through a wire is unchanged under time shifts but not under time reversal—that’s called \textbf{stationary}, not static.
Chapter 2 Electrostatics Introduced

“other” force will include an elastic component that opposes deformation, a classical stand-in for a quantum-mechanical effect. We will see that instead of treating these molecular constituents explicitly, we may summarize them with a modified value of the permittivity.

- Chapter 8 will go beyond statics, introducing conductors in which even mobile charges find their motion impeded by the surroundings. Then the “other” force effectively has a dissipative (frictional or ohmic) part.
- Then Chapter 10 will introduce thermal agitation, which changes the equation of motion for the charges by adding a statistical-physics aspect. The situation may still be static, however, if the average velocity of charges in any region is zero (or small).
- Chapter 15 will consider situations in which the average charge velocity is nonzero, beginning with the case where individual charges move slowly, but nevertheless are so numerous that magnetic effects may not be ignored.
- Chapter 18 will begin our study of charges with general motions.

For now, however, we will stick to the most basic situation: electrostatics. Thus, we focus on the lower arrow on the figure.

Electromagnetic phenomenon: As a proton beam penetrates tissue, it suddenly loses most of its energy in a narrow range of depth.

Physical idea: The energy loss per distance itself depends on particle energy, introducing a nonlinearity that modifies the more usual exponential absorption law.

2.2 REPHRASE IN TERMS OF A POTENTIAL

2.2.1 A static electric field can be re-expressed via an integrability lemma

Because charges are not moving, the charge flux \( \mathbf{j} = 0 \). The static condition also implies that \( \frac{d\mathbf{E}}{dt} = 0 \), so Equations 0.2 and 0.4 imply that there are no magnetic fields \( \mathbf{B} = 0 \), and all we have left of Maxwell are Equations 0.1 and 0.3:

\[
\nabla \cdot \mathbf{E} = \rho_q / \varepsilon_0, \quad \nabla \times \mathbf{E} = 0.
\]

(2.1)

Here \( \rho_q \) is electric charge density and \( \varepsilon_0 \) is a proportionality constant. Some such constant is needed for dimensional reasons: Because charge carries a new kind of dimension that cannot be converted to length, time, or mass, and \( \mathbf{E} \) is force per charge, \( \varepsilon_0 \) must among other things cancel two powers of charge units.

Equations 2.1 look much more complicated than the single Equation 1.1 of newtonian gravity! Let’s first address that defect.

Choose any fixed “reference point” \( \mathbf{r}_0 \) in space and define the electrostatic potential as the scalar function

\[
\psi(\mathbf{r}) = -\int_{\mathbf{r}_0}^{\mathbf{r}} \mathbf{d}\mathbf{r}' \cdot \mathbf{E}(\mathbf{r}').
\]

(2.2)

Here the notation denotes the line integral along any path that starts at the reference point \( \mathbf{r}_0 \) and ends at the field point \( \mathbf{r} \). It doesn’t matter which such path we choose. Any
two such paths differ by a closed loop, so switching to a different path changes $\psi$ by the integral $-\int \mathbf{d}\mathbf{r} \cdot \mathbf{E}$ around that closed loop (Figure 2.1). By Stokes’s theorem, this can be written as a surface integral of $\nabla \times \mathbf{E}$ (Equation 0.9, page 9), which is always zero by Equation 2.1.

As in gravitation, the standard terminology is confusing: The electric potential is not the potential energy of a test particle:

- Equation 2.2 shows that, in electrostatics, $\psi$ is potential energy of a test particle per unit charge.
- In non-static situations, we will see later that $q\psi$ has no direct interpretation as potential energy at all.

Note, too, that our construction of the potential depends on an arbitrary choice of reference point, but only in a trivial way: Changing $\mathbf{r}_0$ just adds a constant to $\psi$. We don’t explicitly indicate the dependence on $\mathbf{r}_0$, because we are already accustomed to the fact that potential energy is only well defined up to an additive constant.

Remarkably, the scalar function $\psi$ contains the same information as the vector field $\mathbf{E}$. To prove that, let’s evaluate the gradient of $\psi$. For example, we’ll find $\psi(\mathbf{r} + \epsilon\hat{x}) - \psi(\mathbf{r})$. We can evaluate Equation 2.2 using any path we like, so choose any path from the origin to $\mathbf{r}$, and another that follows the first one but then moves from $\mathbf{r}$ parallel to the x axis a distance $\epsilon$. Both integrals are the same and cancel, except for the last bit of the first one, which contributes minus $\epsilon \mathbf{E}_x$. We conclude that $-\nabla_1 \psi(\mathbf{r}) = \mathbf{E}(\mathbf{r})$, a generalization of the Fundamental Theorem of Calculus. More generally,

$$
\mathbf{E} = -\nabla \psi.
$$

Then the first of Equations 2.1 becomes the Poisson equation:

$$
\nabla^2 \psi = -\rho_q/\varepsilon_0.
$$

In a region of space with no net charges, the right hand side is zero and the equation is often rechristened the Laplace equation.
2.2.2 Force law

The Lorentz force law with no magnetic field nor external force becomes:

\[
\frac{d}{dt} \vec{P}_\ell = -q_\ell \vec{V} \psi(\vec{r}_\ell(t)).
\] (2.5)

Here \(q_\ell\) is the electric charge, a fixed quantity that is attached to particle \(\ell\).

The SI unit for the electrostatic potential is joule per coulomb, which is the definition of “volt.” The SI officially abbreviates this unit “V,” but that could lead to confusion with volume or something, so we will write \(V\).

The electric field \(-\vec{V} \psi\) therefore has units of newtons per coulomb, or equivalently volts per meter.

2.2.3 An integrability lemma underlies the success of the potential method

We have transformed electrostatics from a set of four linear PDEs in the three unknown functions \(\vec{E}\) (Equation 2.1) to one linear PDE in one unknown function \(\psi\) (Equation 2.4), a considerable simplification. Indeed, it’s the same equation as in newtonian gravitation.

Our success relied on establishing an integrability lemma: While clearly any gradient has zero curl, we found that conversely any curl-free vector field can be written as a gradient via Equation 2.2. We will upgrade this argument when it’s time to find a potential for magnetostatics (Chapter 15), and then again when it’s time to find a 4-vector potential for electrodynamics (Chapter 37).

2.3 BASIC SOLUTIONS

2.3.1 Point charge

One solution of the Poisson equation is the one we found in gravitation is that a point charge of strength \(q\) located at the origin gives \(\psi(\vec{r}) = q/(4\pi \varepsilon_0 ||\vec{r}||)\), or more generally \(5\)

\[
\psi(\vec{r}) = \frac{q}{4\pi \varepsilon_0 ||\vec{r} - \vec{r}_s||}
\] (2.6)

if the charge is located at \(\vec{r}_s\).

**Your Turn 2A**

Find the negative gradient of this function,\(6\) then go back via Equation 2.2 to see how it all fits together.

---

\(4\)See Equation 0.5 (page 3).

\(5\)Placement of the \(4\pi\) factor is just a convention. In gravity we put it into the Poisson equation; in electrostatics, it’s conventional to bury it in the definition of the constant \(\varepsilon_0\), but then it pops up in Equation 2.6.

\(6\)H. Cavendish discovered experimentally that the force between a pair of electrical charges varies inversely to the square of the distance between them. As usual, Cavendish didn’t publish, so this result is now known as Coulomb’s Law.
One big difference between newtonian gravity and electrostatics is that the mass density \( \rho_m \) must always be nonnegative (everything attracts everything), but charge density \( \rho_q \) need not be nonnegative (some pairs of things attract but others repel).

The minus sign in the Poisson equation says that a + charge creates a +1/r potential, that is, a potential energy hill for another + charge. Hence similar charges repel, unlike in gravity.

### 2.3.2 Continuous charge distribution

As in Section 1.5 (page 20), the Poisson equation is linear in \( \psi \), so we can quickly generalize our point-charge solution to the case of a continuous distribution with charge density \( \rho_q(\vec{r}_s) \). Simply subdivide charge into small elements \( dq = \rho_q(\vec{r}_s) d^3 r_s \) and add up their contributions. We’ll call \( \vec{r}_s \) the **source point**, to distinguish it from the field point \( \vec{r} \) where we wish to know the potential. Thus, the potential at the field point becomes an integral over source points:

\[
\psi(\vec{r}) = \int d^3 r_s \frac{\rho_q(\vec{r}_s)}{4\pi \varepsilon_0 ||\vec{r} - \vec{r}_s||} .
\]  

(2.7)

This expression gives the general solution to the Poisson equation. It is called a **Green function solution**, and \( 1/(4\pi||\vec{r} - \vec{r}_s||) \) is called the **Green function** of the Laplace operator.

The apparent singularity in Equation 2.7 is integrable, just as in Section 1.5.

### 2.4 CONDUCTORS

Another difference from gravity concerns “conductors.” These are a class of macroscopic bodies for which it’s a good approximation to say that charges (eventually) arrange themselves freely inside the body, without leaving it.\(^7\)

It may seem a nightmare to handle problems of this sort—we can’t find the fields until we know where the charges go, and vice versa. In practice, however, the method of potentials gives an elegant approach: The free charges in a conductor just scoot around till they no longer feel any net force, that is, until \( \vec{E} = 0 \) everywhere inside the conducting body (and hence \( \psi = \text{const} \)). Because \( \psi \) is a potential energy per test charge, it cannot change discontinuously across the conductor’s surface. Thus, we get a boundary condition on the potential’s gradient: The derivatives of the potential parallel to the surface equal zero.

\[
\vec{E}_\parallel = 0 . \quad \text{just outside a conductor, static}
\]  

(2.8)

The perpendicular component \( \vec{E}_\perp \) need not be zero at the surface; by the Gauss law, \( \vec{E}_\perp \) tells us about the areal charge density.

Often we don’t even need to know the surface charge distribution. But if we do, we can find it by computing \( -\varepsilon_0 \vec{V}_\perp \psi \) once we have solved the boundary-value problem for the potential.

---

\(^7\)“Eventually” because charges may only rearrange slowly, due to friction.
Chapter 10 will modify the preceding comments, acknowledging that they are true only at zero temperature. At nonzero temperature, thermal fluctuations are constantly knocking surface charges away from the surface, so there will be a thin layer with nonzero interior electric field even in equilibrium. That’s called a depletion layer in semiconductors, or “electric double layer” in soft matter (Chapter 10).

2.5 UPCOMING

2.5.1 Reality of electric field

“But what is the electric field really?” This question turned out for many practical purposes to be as unnecessary as the similar one about the newtonian gravitational potential. In this book, $\vec{E}(t, \vec{r})$ is a set of three functions on spacetime, period.

But another kind of “reality” question deserves comment. We could imagine saying, “there’s no such thing as the electric field, just action at a distance between charges via Coulomb’s law.” Today physicists find such nonlocal hypotheses to be repugnant, but that could be prejudice. Must we attribute independent reality to $\vec{E}$? Occam’s Razor would say, “not if you can avoid doing so.” (Especially we should avoid introducing entities that you cannot see, hear, feel, smell, or taste.)

Let’s look ahead a bit. When we graduate to full electrodynamics, we’ll find wave solutions that are “real” (for example, they transport real energy) even after the charges that generated them have stopped moving or even ceased to exist. For example, dipping into quantum phenomena for a moment, consider the atom-like bound state of an electron and a positron. At some moment the electron and positron annihilate each other, as for example in positron emission tomography (PET) imaging. Now nothing remains of them, nothing that could be exerting forces on distant charges—and yet, distant detectors eventually receive any radiation that the electron and positron gave off when they formed that bound state. It would be contrived at best to attempt to represent this situation as action at a distance from charges that no longer exist at the time of detection! Occam says don’t add new entities unnecessarily. But this example shows that the field concept is unavoidable, if we want to live in a world in which energy is locally conserved.

Of course, “wanting” isn’t enough. Eventually we’ll need to prove some mathematical result about local conservation.

**Hanging #H:** Where is the energy in between emission and absorption of radiation? What continues to carry that energy even after the source no longer exists? Is there even a useful concept of “electromagnetic energy,” and for that matter, what does “useful” mean?

Chapter 35 will show that there is indeed a way to attribute energy to fields in such a way that the total energy (particles plus fields) is locally conserved. As a bonus, we’ll also get similar results for momentum and angular momentum.
2.5.2 Quasi-static

We’ll see in Section 8.6 that many situations of interest are not precisely static, but may nevertheless be approximately treated as such when charges are moving slowly.

2.5.3 Beyond static

When things are moving fast, so that we’re not even approximately static, it may seem that we can’t get to first base: The electric field won’t be curl-free, which seems to preclude introducing a potential. Luckily that’s not true—later we’ll construct a set of potential functions applicable in this case as well. None of them will have any interpretation as potential energy per unit charge, however.
2.2’ Falsifiable content of the equations
Similarly to the discussion in Section 1.7, Equations 0.1–0.5 simultaneously give operational meaning to the electric and magnetic fields, and to the charge/mass ratios of the charged bodies, and to the choice of good coordinates on spacetime. Yet in addition to defining the quantities they contain, they also make falsifiable predictions about relations between those quantities! The way this works is that the formulas have the character of an interlocking web of many predictions:

(i) Suppose that we have reproducible classes of test bodies (for example protons, muons…), and an apparatus that creates repeatable situations. Then there exists at least one coordinate system on spacetime, and a number \( q_i / m_i \) characterizing each test body \( i \) (but independent of the apparatus and the test body’s motion), and a set of six functions \( \vec{E}(t, \vec{r}) \), \( \vec{B}(t, \vec{r}) \) characterizing the apparatus but independent of the test body and its initial conditions, such that any physically realizable trajectory of any test body is a solution to Equation 0.5.

(ii) If the apparatus consists of charges executing specified motions, then the functions \( \vec{E} \) and \( \vec{B} \), measured as described in (i) above, are not arbitrary, but are solutions to the partial differential Equations 0.1–0.4 with sources determined by the charges.

(iii) If the apparatus consists of point charges which are themselves free (other than being influenced by EM fields and known forces \( \vec{f}_{\text{other}} \)), then the combined history of the fields and charges is a self-consistent solution of Equations 0.1–0.5, with sources given by formulas in Section 8.3 and Section 34.6.1 (specifically Equation 34.9, page 534).

Similarly to the situation in newtonian gravity, once we find one set of “good” coordinates on spacetime (that is, coordinates for which all phenomena obey the equations in their usual form), then there will also be other such “good” coordinate systems with the same property. The same example transformations mentioned on page 22 work in electrostatics: rigidly shifting or rotating \( x, y, z \) (leaving \( t \) unchanged); shifting in time, and negating any or all of \( x, y, z \), or \( t \) all work. Later, we’ll enlarge this catalog further, but for now, just note that, in parallel with gravitation, the content of the Maxwell/Lorentz equations includes the assertion that at least one “good” coordinate system exists.

Einstein called any “good” coordinate system on spacetime inertial. Later chapters will discuss this notion in detail, but for now, note that all “good” coordinate systems in the above sense are, in particular, cartesian in \( x, y, \) and \( z \) and non-accelerating. One can extend the definitions of the vector operators, dot product, and so on to accommodate curvilinear or accelerated coordinates, but the very fact that those formulas look different from the usual cartesian form means that the Maxwell and Lorentz equations are not form-invariant under arbitrary change of coordinate systems. There is something special about inertial coordinate systems.
2.1 Statics basics
A static charge distribution produces a radial electric field \( \vec{E} = Ar^{-2}e^{-br} \hat{r} \), where \( A, b \) are constants. \( \hat{r} \) is the unit vector in the radial direction.

a. What is the total charge \( q_{tot} \)?

b. What is the charge density? Let \( g(r)dr \) denote the amount of charge located in a spherical shell between radius \( r \) and \( r + dr \), and sketch a graph of \( g(r) \).

2.2 Proton therapy
This problem continues Problem 1.2. In that problem, you found a formula for the deflection angle when a small mass flies by a large mass at rest with “impact parameter” \( A \).

a. Adapt your solution to apply to the electrostatic interaction between two point charges. Be sure that your answer is reasonable for both the attractive and repulsive cases.

b. An electron flies past a (much heavier) proton at rest, with \( A = 100 \text{ pm} \) and \( v_0 = 3 \cdot 10^6 \text{ m/s} \). (A picometer is \( 10^{-12} \text{m} \).) What is the total deflection?

c. This time, the proton flies past an electron initially at rest. First, relate this situation to the one you just solved. The proton’s path is approximately unaffected by the electron, but the electron gains kinetic energy \( W \). Derive an expression for \( W \). Evaluate your answer for the illustrative case in (a).

Your formula involved the quantity
\[
Y \equiv \frac{K}{(m_e A v_0^2)},
\]
where \( K = e^2/(4\pi\varepsilon_0) \), \( e = \) proton charge = \(-\)electron charge, \( m_e = \) electron mass, and \( v_0 = \) magnitude of initial velocity. \( A \) is the perpendicular distance from the proton to the electron’s initial trajectory.

d. Do a little trigonometry to express the electron’s final kinetic energy \( W \) in the form
\[
W = \text{(stuff)}/(A^2 + (\text{more stuff})).
\]
The factors in parentheses don’t depend on \( A \); you are to find them.

e. When a proton flies through a gas of many electrons, all initially at rest\(^8\), it occasionally encounters one with a small value of \( A \) and gives it a significant kick. So far, we’ve pretended that during that encounter the proton is unaffected. But over many collisions, the proton will lose energy, about equal to the sum of all the \( W \) values for each encounter.

Suppose that the medium has a uniform number density of electrons, \( n_e \). Initially the incoming proton is at depth \( x = 0 \) within the tissue, and has kinetic energy \( T_0 \). After passing through to depth \( x \), its velocity has fallen to some value \( v(x) < v_0 \) due to

\(^8\)You may neglect screening in this problem.
many encounters, and so its kinetic energy has also fallen to \( T(x) \). Neglect the fact that the proton’s direction will also change; suppose that it is always moving in the same direction.

In the next \( dx \), there are electrons at various values of \( A \). Of these, \( (2\pi A \ dA)(dx) \rho_q \)

have impact parameter values lying between \( A \) and \( A + dA \). Write a formula for the total energy loss of the proton due to these electrons, and integrate it over \( A \) to get the energy loss per depth, \( dT/dx \).

f. Uh-oh. You found an infinite result; the integral is divergent. But wait. The electrons in human tissue aren’t free; they are bound into molecules. If the energy transfer exceeds the binding energy, then maybe it’s OK to neglect that fact, as we have done. But otherwise, the passage of the proton just deforms the molecule temporarily without necessarily any net loss of energy; your formula from (d) is not applicable in this case.

We’ll take this complication into account crudely by just cutting off the integral in (e) at the value \( A_{\text{max}} \) at which \( W \) equals the ionization energy \( I \) of a molecule. Find a formula for \( A_{\text{max}} \) in terms of \( m_e, v_0, I \), and constants of Nature.

g. Now do the integral over \( A \), to find \( dT/dx \) in terms of \( T, I, n_e \), and constants. Note that \( T = \frac{1}{2} M_p v^2 \), where \( M_p \) is the proton mass (that is, it’s not \( \frac{1}{2} m_e v^2 \)).

h. Simplify your expression by defining a suitable length scale and expressing \( x \) in terms of it. Also substitute some numbers:

You know \( K = e^2/(4\pi\varepsilon_0) = 1.4 \text{ eV nm} \).

You know the electron and proton masses.

Suppose that \( I \approx 10 \text{ eV} \).

Tissue is mostly water. You know how to compute the electron density \( n_e \) of water. (Assume that all the electrons have the same ionization energy.)

Suppose that the proton initially has \( T_0 = 100 \text{ MeV} \).

i. Now you can find the relation between \( x \) and \( T(x) \). This will involve solving the differential equation \( dT/dx = (\text{expression you found}) \). Luckily that equation can be solved just by doing an integral. But you may not know how to do that integral. Get a computer to evaluate it numerically, and hence find the \( x \) values corresponding to a set of \( T \)’s starting at \( T_0 \) and decreasing to, say, \( T_0/50 \). Plot your answer as a graph of remaining kinetic energy \( T \) versus \( x \).

j. Actually, we are more interested in the deposition of energy as a function of depth. Make a second plot showing \( dT/dx \) as a function of \( x \), and comment on its general form.
CHAPTER 3

Electrostatic Multipole Expansion

3.1 FRAMING: DISTILLATION

Chapter 2 showed that electrostatics is straightforward if you are told where the charges are (fixed charge distribution). That’s often a reasonable approximation when we study molecules, for example H₂O or CO₂. The charge distributions on these molecules come from quantum mechanics, but given that, we can ask what electrostatic fields they create, and what qualitative conclusions we can draw.¹ Moreover, often we are only interested in the fields far from a molecule or other localized charge distribution. It’s useful to be able to distill just a few numbers from the distribution that characterize the most significant features of its far fields. This chapter will systematize that procedure for electrostatics.

Besides bringing technical and conceptual benefits, Chapter 15 will extend the ideas to get a similarly useful magnetic multipole expansion. Then it will come around a third time, when we study radiation in Chapter 43. It’s a powerful method.

Electromagnetic phenomenon: Molecular symmetry gives some quick, qualitative predictions about molecular interactions.

Physical idea: Symmetry can force some multipole moments to be zero.

3.2 THE ELECTROSTATIC MULTIPOLE EXPANSION

Consider an isolated, static charge distribution confined to a region of size \( a \), viewed from far away; that is, at a field point \( \mathbf{r} \) with

\[ r \gg a. \quad \text{far field (static)} \quad (3.1) \]

¹It’s true that a molecule is not quite fixed—it can deform, for example—but for some purposes we don’t need that level of detail.
We’ll choose a fixed reference point somewhere inside that region and use it as an origin of coordinates; thus, charge \# \epsilon \text{ sits at a position } \vec{r}(\epsilon) \text{ with } r(\epsilon) \lesssim a \ll r. \text{ The goal is to show that the electrostatic potential at } \vec{r} \text{ can be expanded as}

\[ \psi(\vec{r}) = q_{\text{tot}} \psi_{\text{[0]}}(\vec{r}) + \vec{D}_E \cdot \vec{\psi}_{\text{[1]}}(\vec{r}) + \sum_{ij} \left[ \bar{\Omega}_{\epsilon ij} \vec{\psi}_{\text{[2]} i j}(\vec{r}) \right] + \cdots. \tag{3.2} \]

Before proving this daunting formula, let us define all its symbols.

$q_{\text{tot}}$ is a scalar constant called **electric monopole moment** or **zeroth moment** of charge. The three constants \( \bar{D}_E \) form a vector called **electric dipole moment** or **first moment** of charge. The constants \( \bar{\Omega}_{\epsilon ij} \) are called the **electric quadrupole tensor** or “traceless part of the **second moment** of charge.” Later chapters will develop a general definition of tensors,\(^2\) but for now think of \( \bar{\Omega}_E \) as a \( 3 \times 3 \) matrix.

The moments just mentioned are defined by\(^3\)

\[ q_{\text{tot}} = \sum_{\epsilon} q_\epsilon, \quad \vec{D}_E = \sum_{\epsilon} q_\epsilon \vec{r}(\epsilon), \quad \bar{\Omega}_{\epsilon ij} = \sum_{\epsilon} q_\epsilon (3 \vec{r}(\epsilon) \vec{r}(\epsilon) - ||\vec{r}(\epsilon)||^2 \delta_{ij}). \tag{3.3} \]

Although the indices on the quadrupole tensor each run from 1 to 3, so that it has nine entries, only five of these have independent values. That’s because \( \bar{\Omega}_{E \epsilon ij} \), regarded as a matrix, is always symmetric and traceless. More explicitly, the first term of \( \bar{\Omega}_E \) is the sum over all charges of

\[ 3q_\epsilon \begin{bmatrix} x_\epsilon^2 & x_\epsilon y_\epsilon & x_\epsilon z_\epsilon \\ y_\epsilon x_\epsilon & y_\epsilon^2 & y_\epsilon z_\epsilon \\ z_\epsilon x_\epsilon & z_\epsilon y_\epsilon & z_\epsilon^2 \end{bmatrix}. \]

The matrix just given is symmetric. Its trace (sum of diagonal entries) is \( 3 \sum q_\epsilon (x_\epsilon^2 + y_\epsilon^2 + z_\epsilon^2) \). In the second term of \( \bar{\Omega}_E \), we get the symmetric matrix \( \delta_{ij} \), whose trace is 3, times \(- \sum q_\epsilon ||\vec{r}(\epsilon)||^2 \). When combined, these terms form a symmetric matrix whose trace equals zero.

For a continuous charge distribution, we have analogously

\[ q_{\text{tot}} = \int d^3r * \rho_q(\vec{r}_*), \quad \vec{D}_E = \int d^3r * \rho_q(\vec{r}_*) \vec{r}_*, \]

the **zeroth** and **first moments** of the charge distribution with respect to the chosen reference point, and similarly for \( \bar{\Omega}_E \).

Continuing to unpack Equation 3.2, the functions \( \psi_{\text[\alpha]} \) are called **multipole potentials**; they are universal functions of observer position (independent of the nature of the charge distribution):\(^4\)

\[ \psi_{\text[0]}(\vec{r}) = \frac{1}{4\pi \epsilon_0 r}; \quad \vec{\psi}_{\text[1]}(\vec{r}) = \frac{1}{4\pi \epsilon_0 r^2} \vec{r}; \quad \vec{\psi}_{\text[2] i j}(\vec{r}) = \frac{1}{8\pi \epsilon_0 r^3} \left( \vec{r} \vec{r}_j - \frac{1}{3} \delta_{ij} \right). \tag{3.4} \]

---

\(^2\) Chapters 13–14 and 32–34

\(^3\) Beware that some authors move a factor \( 1/2 \) from the quadrupole field \( \vec{\psi}_{\text[3]} \) into the definition of the moment \( \bar{\Omega}_E \); others instead use the convention given here. Still others use the phrase “quadrupole tensor” to mean the second moment of charge, and “traceless quadrupole tensor” to mean our \( \bar{\Omega}_E \).

\(^4\) The \( \delta_{ij} \) terms in Equations 3.3 and 3.4 are redundant: Omitting either (but not both) leaves \( \psi \) unchanged. Both are included to emphasize that: (a) The potential at order \( r^{-3} \) has a traceless character, no matter what the charge distribution; and (b) the trace of the second moment of charge cannot contribute at all to the parts of the field that are of order \( r^{-3} \).
These formulas define a single monopole field, a set of three dipole fields, and a set of five independent quadrupole fields, respectively. Finally, the ellipsis in Equation 3.2 denotes corrections that fall off with distance faster than the ones shown, specifically as $(r^{-1})^4$ or higher.

Note that each successive multipole moment contains an additional factor of order the system size $a$, whereas each successive multipole potential contains an additional factor of $1/r$; thus Equation 3.2 is an expansion in powers of the small dimensionless parameter $a/r$.

### 3.3 SOME TAYLOR EXPANSIONS

We need to prove Equation 3.2. First recall some useful facts.

We will often use the series expansions for the functions $(1 + \varepsilon)^{1/2}$ near $\varepsilon = 0$:

$$\sqrt{1 + \varepsilon} = 1 + \frac{1}{2} \varepsilon - \frac{1}{8} \varepsilon^2 + \cdots$$

$$\frac{1}{\sqrt{1 + \varepsilon}} = 1 - \frac{1}{2} \varepsilon + \frac{3}{8} \varepsilon^2 + \cdots.$$  \hspace{1cm} (3.5)

It is good to know how to get these from Taylor’s theorem.

**Your Turn 3A**

You may wonder *how good* those approximations are, how small $\varepsilon$ must be, and so on.

a. Use a computer to make a graph of the residuals: $f_0(\varepsilon) = \sqrt{1 + \varepsilon} - 1$, $f_1(\varepsilon) = \sqrt{1 + \varepsilon} - (1 + \varepsilon/2)$, $f_2(\varepsilon) = \sqrt{1 + \varepsilon} - (1 + \varepsilon/2 - \varepsilon^2/8)$ and comment.

b. Repeat for the function $(1 + \varepsilon)^{-1/2}$.

Now suppose that the small quantity $\varepsilon$ is itself given in terms of another small quantity: $\varepsilon = \delta + A\delta^2$, and we wish to rearrange our result as a series in $\delta$. Substituting gives

$$\frac{1}{\sqrt{1 + \delta + A\delta^2}} = 1 - \frac{1}{2} (\delta + A\delta^2) + \frac{3}{8} (\delta + A\delta^2)^2 + \cdots$$  \hspace{1cm} (3.6)

$$= 1 - \frac{1}{2} \delta + \delta^2 (-\frac{1}{2} A + \frac{3}{8}) + O(\delta^3).$$  \hspace{1cm} (3.7)

Note that:

- We chose to stop the expansion at some fixed order in $\delta$ (here second order).
- Part of the term that was first order in $\varepsilon$ in Equation 3.6 has entered into the term that is second order in $\delta$ in Equation 3.7.
- Some but not all of the order $\varepsilon^2$ term was needed. (The terms $\frac{3}{8} (2A\delta^3 + A^2\delta^4)$ were not.)
- There was no need to write down any term of order $\varepsilon^3$ or higher, because anything contained in such a term would be at least order $\delta^3$, and we chose to only keep explicit track of terms up to order $\delta^2$. 
3.4 DERIVATION OF THE FORMULA

Now that we have unpacked our claim (Equations 3.2–3.4), it’s time to prove it starting from the basic solution for the potential around a point charge, by making a Taylor expansion:

\[
\psi(\vec{r}) = \sum_{\ell} \frac{q_{\ell}}{4\pi\varepsilon_0} \left( \frac{4}{r} - \frac{\vec{r} \cdot \vec{r}(\ell)}{r^2} \right) - \frac{1}{4\pi\varepsilon_0 r} \sum_{\ell} q_{\ell} \left( \frac{\vec{r} \cdot \vec{r}(\ell)}{r^2} + \frac{\vec{r}(\ell)}{r^2} \right)^{1/2}
\]

The small quantity in this expansion is itself the sum of two terms, of which the second is even smaller than the first. Following Section 3.3, we therefore reorganize in powers of \(1/r\) keeping terms up to \(r^{-3}\):

\[
= \frac{1}{4\pi\varepsilon_0 r} \sum_{\ell} q_{\ell} \left( \frac{1}{r^2} \left( 1 - \frac{1}{2} \frac{\vec{r} \cdot \vec{r}(\ell)}{r^2} \right) + \frac{3}{8} \left( -\frac{1}{2} \frac{\vec{r} \cdot \vec{r}(\ell)}{r^2} + \ldots \right)^2 + \ldots \right)
\]

This result is nearly the one announced earlier (Equations 3.2–3.3). We only need to note that the difference between the last formula and Equation 3.2 is \(1/(8\pi\varepsilon_0 r^3)\) times \(-1/3\vec{Q}_E \cdot \delta_{ij}\), which is zero because \(\vec{Q}_E\) is traceless.

3.5 MULTIPOLe MOMENTS ORGANIZE THE FEATURES OF A DISTRIBUTION ACCORDING TO IMPORTANCE

Now that we’ve proved the result, it’s worthwhile to ask if it was worthwhile.

The virtue of Equation 3.2 is that each term has been written as a sum of products of:

- a universal, archetypal field (a component of one of the \(\psi^{[p]}\)'s), times
- a number (one of the moments).

The “multipole fields” \(\psi^{[p]}(\vec{r})\) have nothing to do with the source object—they just catalog possible solutions of the Laplace equation. In contrast, the moments have nothing to do with observer position \(\vec{r}\)—they are constants that state how much of each field type is present.

Thus, the first few moments are a convenient summary of the relevant aspects of the source for purposes of finding its far fields. Specifically, keeping up to order \(p\) (the “2\(p\)-pole approximation”) tells us the distant potential up to order \((a/r)^{p+1}\), or equivalently the electric field up to order \((a/r)^{p+2}\). It can be more convenient and insightful to work...
with just a few moments than to include all the irrelevant other details of the full charge distribution.

Section 3.7.3 will show that the multipole expansion also lets us connect symmetry of, say, a molecule to the character of its long-range forces.

3.6 MORE REMARKS

3.6.1 Summary so far

Starting from the humble $1/r$ solution, we got Equations 3.2–3.4, which show that:

- If a static, localized charge distribution has any term in its potential that falls as $1/r$, that contribution to $\psi$ must be spherically symmetric.
- If $\psi$ has any term of order $1/r^2$, that term cannot be spherically symmetric; instead, it will have a specific angular dependence (it must be dipolar). Everything about this contribution is fixed once we specify its strength and orientation via a vector $\vec{D}_k$.
- If it has any $1/r^3$ term, that part also cannot be spherical. To get an angle-independent $A/r^3$ dependence would require the quadrupole tensor to be a constant times the identity matrix, but any distribution whatever will have a traceless quadrupole tensor.

These are powerful and general results, which we obtained without much work.

To get the helpful decomposition into (few things about source) $\times$ (few universal fields), we were obliged to introduce a new kind of entity $\vec{Q}_k$, which we called a “tensor.”

3.6.2 From potentials to fields

This derivation would have been a nightmare had we worked directly with the electric field. So the potential method has practical advantages. After finding the quadrupole potentials from the moments, then we can take a negative gradient and find the electric field, if we wish that.

**Your Turn 3B**

Derive expressions for the contributions to the electric field coming from the dipole and quadrupole potentials $\psi^{[1]}$ and $\psi^{[2]}$ appearing in Equation 3.4. Note how the units work out.

3.6.3 Apparent singularity

Every term in the multipole expansion of $\psi$ is singular at $r = 0$. The corresponding singularities in the electric field are worse still. Is that a problem? No: The expansion

---

Chapter 13 will point out that this observation is reminiscent of something in mechanics: To express the angular momentum of a rigid body as a product of (few things characterizing the body) $\times$ (angular velocity imposed on body), we must also introduce a “moment of inertia tensor.”
Chapter 3 Electrostatic Multipole Expansion

is a power series in $a/r$, so it breaks down (becomes inaccurate) at $r \to 0$. (Similarly, the Earth’s gravitational potential looks like $1/r$ outside the Earth, but that doesn’t imply there’s a black hole at the center!) A smooth distribution of charge will have nonsingular potential and field.

3.6.4 All moments after the first nonzero one depend on choice of base point

Our expansion of $\psi$ depends implicitly on our choice of the origin of coordinates. If we choose a different origin, then $q_{\text{tot}}$ won’t change, but in general $D_e$ will, and so will $\Phi_e$, and so on. It’s not really about coordinate choice: We could alternatively have defined moments relative to any reference point $\vec{h}$ via

$$\tilde{D}_e^{\text{alt}} = \sum_{\ell} q_\ell (\vec{r}_{\ell} - \vec{h}), \quad \text{and so on.} \tag{3.8}$$

Your Turn 3C

a. Get formulas for the changes in $\vec{D}_e$ and $\vec{\Phi}_e$ under change of reference point. That is, compare $\vec{h} = 0$ to a general value in Equation 3.8.
b. Show that $\vec{D}_e$ won’t depend on $\vec{h}$ if $q_{\text{tot}} = 0$.
c. Show that $\vec{\Phi}_e$ won’t depend on $\vec{h}$ if both $q_{\text{tot}} = 0$ and $\vec{D}_e = 0$.

Your result implies that if net charge is nonzero, then we can always arrange that $\vec{D}_e = 0$ just by choosing an appropriate reference point: Adjusting the three components of $\vec{h}$ suffice to set the three components of $\vec{D}_e$ to desired values.\(^6\)

Your Turn 3D

a. So can we forget about electric dipole fields? Why or why not?
b. Can we use a similar argument to eliminate $\vec{\Phi}_e$?

3.6.5 Spherical distributions

Any spherically-symmetric distribution of charge trivially has $\vec{D}_e = 0$, and not so trivially $\vec{\Phi}_e = 0$ also.\(^7\) In fact all moments beyond the 0th are zero: $\psi = q_{\text{tot}}/(4\pi\varepsilon_0 r)$ outside any such distribution (Birkhoff’s theorem).\(^8\)

---

\(^6\)There is an analogous gravitational multipole expansion in newtonian gravity. After working Your Turn 3C, you’ll understand why you never hear about a “gravitational dipole moment.”

\(^7\)See Problem 3.3. We assumed that the reference point is taken to be the central point.

\(^8\)Robert Hooke intuitively understood this result and communicated it to Newton around 1679. Newton proved a version of it in 1685.
3.6.6 Symmetry may dictate that some moments equal zero

Even without spherical symmetry, we sometimes have a shortcut to seeing that some moments must equal zero.

Any static charge distribution with an inversion symmetry through a point will have \( \vec{D}_E = 0 \) when evaluated with respect to that point. Any distribution with a plane of reflection symmetry will have \( \vec{D}_E \cdot \vec{n} = 0 \) where \( \vec{n} \) is the perpendicular to that plane. Hence an axially-symmetric distribution will have \( \vec{D}_E \) aligned with its axis.

Next, suppose that +q is located at \((0, 0, a)\) and -q is at \((0, 0, -a)\). Then \( \vec{D}_E = (2qa)\hat{z} \).

You should show that \( \vec{E}_E = 0 \) directly from the definition, but here is a more insightful, and generalizable, argument.

Consider any arbitrary static charge distribution. Create a new charge distribution obtained from the given one by the recipe:

- Invert all positions, \( \vec{r}'(\vec{r}) = -\vec{r}(\vec{r}) \), and also
- Reverse the signs of each charge, \( q'_E = -q_E \).

Then note that:

- The new distribution has \( q'_\text{tot} = -q_\text{tot} \).
- The new distribution has two minus signs in the dipole moment, so \( \vec{D}'_E = \vec{D}_E \).
- The new distribution has three minus signs in the quadrupole moment, so \( \vec{Q}'_E = -\vec{Q}_E \).

If transformation \( T_1 \) leaves the charge distribution unchanged, then every multipole moment is also unchanged. We can then conclude without detailed calculation that in this situation:

- \( q_\text{tot} = q'_\text{tot} = -q_\text{tot} \), so \( q_\text{tot} \) must equal zero.
- \( \vec{D}_E = \vec{D}'_E = \vec{D}_E \), which is a tautology, so there is no restriction on the dipole moment.
- But \( \vec{Q}_E = \vec{Q}'_E = -\vec{Q}_E \), so the quadrupole moment equals zero. In fact, every \( 2^p \)-pole moment with \( p \) an even integer must be zero.

(3.9)

Returning to the specific distribution with +q located at \((0, 0, a)\) and -q at \((0, 0, -a)\), we see that it is indeed unchanged under \( T_1 \). You can check the validity of the claims (3.9) by explicit calculation. But octupole, for example, is not constrained; we cannot conclude it's zero in this situation.\(^9\)

**Your Turn 3E**

a. Think up a charge distribution that becomes *minus* itself under \( T_1 \). [**Hint:** Try placing four point charges all in the xy plane.]

b. Explain why, for any such distribution, every \( 2^p \)-pole moment with \( p \) an *odd* integer must equal zero.

c. Check your general conclusion in (b) for your specific example in (a).

---

\(^9\)See Problem 3.2.
3.6.7 Pure dipole is an idealization arising as a limiting case

The two-charge distribution discussed in the previous subsection must have vanishing quadrupole moment, but as mentioned, nothing prevents it from having octupole and higher odd-\(p\) moments.

If we want a purely dipole field, then we must consider a limiting case, in which the separation \(2a\) between the two point charges is sent to zero while increasing the charges so as to hold the dipole moment fixed. Thus, in this limit the charge \(q = D_{ij}/(2a) \to \infty\). That idealized limit is called the pure dipole or point dipole distribution.

3.6.8 The quadrupole moment may be uniaxial or biaxial

A symmetric matrix like \(Q_{\alpha_\beta}^{\alpha_\beta}\) has three real eigenvalues, each of which is unchanged by rotation. (Only two of these values are independent because the quadrupole moment is always traceless.) If any two eigenvalues match, we call quadrupole moment uniaxial; otherwise it is biaxial (no two match).

**Your Turn 3F**

a. Why didn’t we include the case where all three eigenvalues match?

b. Try to find concrete examples of each case.

3.7 FORCE AND TORQUE ON A FIXED CHARGE DISTRIBUTION

3.7.1 Potential energy depends both on position and on orientation

Imagine a localized charge distribution (subsystem \(1\), for example, a molecule) sitting in an externally created, static electric potential \(\psi^{\text{ext}}\) (from subsystem \(2\), for example, a macroscopic lab apparatus). Choose a reference point somewhere inside distribution \(1\). Then describe the distribution by stating the location \(\vec{r}\) in space of that point, the constituent charges \(q_\ell\) and their offsets \(\vec{r}_\ell(\ell)\) from the reference point. We assume the distribution to be rigid: That is, it may only change by overall translation (change \(\vec{r}\)). (Later, Section 3.7.2 will also consider rigid rotation.)

Suppose that subsystem \(2\) is not significantly distorted by the presence of \(1\). Then the potential energy of charge distribution \(1\) in the external potential is\(^{10}\)

\[
U(\vec{r}) = \sum_\ell q_\ell \psi^{\text{ext}}(\vec{r} + \vec{r}_\ell(\ell)) + \text{const}.
\]

The last term includes the mutual potential energies of the constituent charges. It is constant because rotation and translation don’t alter the distances between those charges, so we will drop it.

Next, add the additional condition that the external potential varies slowly over the size of subsystem \(1\). Mathematically, this means that it is characterized by some length

---

\(^{10}\)See Section 2.2.1.
scale $L$, the $n$th derivatives of $\psi^{\text{ext}}$ are smaller than their predecessors by roughly a factor of $1/L$, and $L$ is much larger than the size of charge distribution $1: L \gg ||\vec{r}_1||$. Certainly a macroscopic apparatus will be much bigger than any individual molecule that we choose to study. In this situation, we may Taylor expand the external potential about the reference point:

\[ U(\vec{r}) = \sum_{\ell} q_{\ell} \psi^{\text{ext}}(\vec{r}) + \sum_{\ell} q_{\ell} \frac{\partial \psi^{\text{ext}}}{\partial \vec{r}} \bigg|_{\vec{r}_\epsilon} \cdot \vec{r}_\epsilon + \sum_{\ell} q_{\ell} \frac{1}{2} \frac{\partial^2 \psi^{\text{ext}}}{\partial \vec{r}_\epsilon \partial \vec{r}_{\ell}} \bigg|_{\vec{r}_\epsilon \vec{r}_{\ell}} + \cdots \quad (3.10) \]

\[ = q_{\text{tot,1}} \psi^{\text{ext}}(\vec{r}) + \vec{D}_{\text{b(1)}} \cdot \vec{\nabla} \psi^{\text{ext}} \bigg|_{\vec{r}} + \cdots. \quad (3.11) \]

Equation 3.10 uses the summation convention in the third term (Section 0.3.2, page 8); the ellipsis denotes terms with third and higher derivatives. Equation 3.11 truncates further, by dropping everything beyond dipole order.

**Your Turn 3G**

Even if the net charge and dipole moment of charge distribution 1 are both zero, nevertheless in general there will be some interaction: Continue the Taylor expansion, Equation 3.10, to the next order and describe what new force you get in that case.

As an application, consider the interaction of two neutral dipoles. That is, suppose that $q_{\text{tot,1}} = 0$, and that subsystem 2 is itself a fixed charge distribution localized near some distant point, which we take to be the origin (Figure 3.1). Its net charge is zero and its dipole moment evaluated at that reference point is $\vec{D}_{\text{b(2)}}$. 

---

**Figure 3.1**: Two interacting, distant, rigid charge distributions.
Your Turn 3H

a. Find the leading-order contribution to the interaction potential energy, $U(\vec{r})$. How does it depend on the separation distance $r$?

b. Holding $r$ fixed, consider four possible orientations of the dipoles:

$(\uparrow \cdots \uparrow); \quad (\rightarrow \cdots \rightarrow); \quad (\uparrow \cdots \downarrow); \quad (\rightarrow \cdots \leftarrow)$.

In each case, the separation vector $\vec{r}$ is horizontal (indicated by the ellipses). Rank-order these four cases according to their interaction potential energy, and say which feel attractive and which feel repulsive forces. In each case, start by stating your physical expectation, then see how it is borne out in the math.

3.7.2 Force and torque arise as derivatives of potential energy

**Force**

We can now compute the negative gradient of $U$ to find the net force on the charge distribution. In addition to the expected $q_{\text{tot}} \vec{E}^\text{ext}_{i}$ (from the first term of Equation 3.11), the next term is

$$\vec{F}_{ij}^\text{ext} = (-\vec{V}_{j} \vec{V}_{j} \psi^\text{ext}) \vec{D}_{i,j}^\text{ext} = \vec{D}_{i}^\text{ext} \cdot \vec{V}_{j} \vec{E}_{i}^\text{ext}.$$

That is, even a neutral charge distribution will feel a net force if it has a dipole moment and is immersed in a nonuniform field.

**Torque**

Until now, we have allowed the charge distribution to translate (that is, to change its position $\vec{r}$) but not rotate. If its potential energy changes upon rotation about some point, then our charge distribution will experience a net torque about that point. To be concrete, consider rotation by $d\theta$ about an axis parallel to $\hat{x}$ and passing through the reference point we used to define the multipole expansion. Then $-dU/d\theta$ is the $z$ component of torque, $\vec{r}_{3}$.

To find it, we displace each constituent charge from $\vec{r}_{(\ell)}(\theta)$ to $S\vec{r}_{(\ell)}$, where the infinitesimal rotation matrix $S$ is defined by

$$S = \begin{bmatrix} 1 & -d\theta & 0 \\ d\theta & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \cdots = \mathbb{1} + d\theta \mathbf{T} + \cdots. \quad (3.12)$$

The ellipses denote terms of second and higher order in $d\theta$. The matrix $\mathbf{T}$ is called the **generator** of the rotation $S$. To first order in $d\theta$, the potential energy is then

$$U = \sum_{\ell} q_{\ell} \psi^\text{ext}(\vec{r} + \vec{r}_{(\ell)} + d\theta \vec{r}^\prime \cdot \vec{r}_{(\ell)})$$

$$= q_{\text{tot}} \psi^\text{ext}(\vec{r}) + \sum_{\ell} q_{\ell} \vec{V}_{\ell} \psi^\text{ext} \bigg|_{\vec{r}} \cdot (\vec{r}_{(\ell)} + d\theta \vec{r}^\prime \cdot \vec{r}_{(\ell)}) + \cdots.$$
The crossed-out terms are constants; the change as we rotate is

$$dU = -\vec{E}^{\text{ext}} \cdot d\vec{T} \cdot \vec{E}.$$  

Notice that the antisymmetric matrix in Equation 3.12 can be written $d\vec{T}_{ij} = -\epsilon_{ij3} d\vec{T}.$ 

So

$$\vec{\tau}_3 = -dU/d\vec{T} = -\epsilon_{ij3} \vec{E}^{\text{ext}} \cdot \vec{D}_{k,l} = (\vec{D}_k \times \vec{E}^{\text{ext}})_3.$$  

More generally, $\vec{\tau} = \vec{D}_k \times \vec{E}^{\text{ext}}.$

In short, a neutral dipole free to rotate in an external field tends to align with that field: It feels a torque that vanishes when $\vec{D}_k \parallel \vec{E}^{\text{ext}}.$ When aligned, we already found in Equation 3.11 that it further feels a force directed toward a region of stronger $||\vec{E}^{\text{ext}}||$.

### 3.7.3 Several intermolecular forces are dipolar in character

Physical chemists tell us that:

- Sodium chloride is just a lot of ions (electric monopoles).
- Water and HCl consist of molecules that are neutral but that have net dipole moments.
- CO$_2$ has no dipole moment but nonzero quadrupole moment.
- Methane is a tetrahedron.
- Neon does not form molecules; it is a perfectly spherical charge distribution. So all of its multipole moments vanish.

That list was written in a particular sequence, based on the rank of the leading multipole interaction (0, 1, 2, > 2, and $\infty$). Interestingly, however, it is also ordered in terms of boiling points! For example, the dipole-dipole attraction of water molecules for each other gives them a strong cohesive force that discourages them from separating (vaporizing). As we go down the list, the intermolecular forces fall faster with distance and the boiling point goes down.

The reasoning just given is a bit glib, and may not seem applicable to molecules with a permanent dipole moment but in liquid state. For example, in water at room temperature the dipoles are thermally randomized, so the average $\langle \vec{D}_k \rangle = 0.$ Remember, however, that the expectation of a product is not equal to the product of expectations. Indeed, the random thermal fluctuations of neighboring molecules will be partially correlated, leading to nonzero $\langle \vec{D}_{k(1)} \cdot \vec{D}_{k(2)} \rangle \neq 0,$ and hence decreased energy via Equation 3.11: Each of the dipoles can be thought of as partially aligning the others. So there will be a net attraction after all, in this context sometimes called **Keesom interaction**.

Even neon, with no dipole moment at all, does liquefy, albeit at a low temperature. So its atoms do develop some attraction, despite being perfectly spherical in the ground state! To understand this qualitatively, remember that even though the dipole moment’s expectation is zero, still its instantaneous value will have spontaneous quantum fluctuations. And these quantum fluctuations again have an energetic tendency to correlate with

---

11See Problem 3.4.

12Actually, carbon dioxide goes directly from solid to vapor at atmospheric pressure, so here we should consider the sublimation point.
those of a neighboring atom. This source of electrostatic attraction is sometimes called London interaction or dispersion interaction.\(^\text{13}\)

Together, the quantum and the statistical correlation attraction effects are sometimes called the van der Waals interaction. Such effects play a key role in some soft matter systems.

### 3.7.4 Dipole moment can be induced by an external field

Moreover, real atoms and molecules are not perfectly rigid; they may deform in the presence of an external field, acquiring an induced dipole moment that does not average to zero. For example, a CO\(_2\) molecule can bend slightly. Much larger objects, such as micrometer-scale particles, can also gain dipole moments in this way. Once such a moment exists, it can lead to net force and torque as computed earlier.

Thus, for example, the resulting moment can align with the external field, and then experience a force pushing it toward regions of higher field strength, even if the atom or molecule is neutral and had no dipole moment to begin with. For example, a hairbrush that is charged after running through your cat’s fur will attract small neutral objects.

To get intuition, imagine the molecule as two charges on a Hooke-law spring. Then the induced dipole moment is linearly proportional to the imposed electric field: \(\hat{D}_E = \alpha \hat{E}\), where \(\alpha\) is a constant called the molecular polarizability.\(^\text{14}\) That induced moment in turn feels a force \(\alpha \hat{E} \nabla (\hat{E}_0) = \frac{1}{2} \alpha \nabla (E^2)\) directed toward the region of higher field strength.

Note that the electric field appears squared in the preceding formula. If we change its sign, that doesn’t affect the force. So even the rapidly-varying electric field of a laser beam will create a net force pulling a polarizable object into the beam. This observation is one way to think about optical tweezers, which can exert precisely controlled, piconewton-scale forces on micrometer-scale objects. Typically the object is not in vacuum, but what matters is the difference between its polarizability and that of the surrounding water (at optical frequency).

### 3.8 PLUS ULTRA

Pursuing the quadrupole term may seem like hairsplitting—it’s subleading in powers of the small quantity. But:

- Sometimes the dipole moment of a neutral atom or molecule is zero for symmetry reasons, for example, in CO\(_2\). In that case, the quadrupole term is the dominant one.
- There is also a multipole expansion for electromagnetic radiation.\(^\text{15}\) Here, too, even if the transition dipole moment is zero, still the atom or molecule can radiate via its...

---

\(^{13}\)Dispersion interactions increase with increasing number of electrons, which partially explains why carbon tetrachloride has a higher boiling point than methane, despite having the same tetrahedral symmetry.

\(^{14}\)Later, we’ll account for the possibility that the polarizability may not be isotropic (Section 13.3.1 and Chapter 52).

\(^{15}\)See Chapters 43–44.
quadrupole moment. But that radiation is weaker in classical electrodynamics (the emission rate is smaller), a reflection of its higher-multipole character, just as we found that the static quadrupole field falls faster than any dipole field.

- In *gravitational* radiation, there’s *never* a dipole component; the leading order behavior involves the time-dependence of the quadrupole moment of mass (or higher if the quadrupole moment is zero).

**FURTHER READING**

*Intermediate:*
Van der Waals interactions: Butt & Kappl, 2018; Israelachvili, 2011.

*Technical:*
Almost all about multipole expansions: Raab & de Lange, 2005.
3.2’a Counting moments
There’s only one kind of monopole field, characterized by only one overall constant of proportionality, \( q_{\text{tot}} \). There’s essentially only one kind of dipole field: You can convert any of the \( \psi[^1_j] \) into any other just by rotating and rescaling, or in other words you can place any dipole in a standard orientation, normalize its overall strength and it then resembles any other.

Quadrupole fields are more diverse. Even if we choose a standardized normalization, the quadrupole tensor \( \tilde{Q}_E \) has \( 5 - 1 = 4 \) independent degrees of freedom, too many to be reduced to a standard form by the action of just three rotations.

Indeed, Section 3.6.8 (page 44) pointed out that some quadrupoles have more symmetry than others; there is a rotationally invariant distinction between those for which two eigenvalues match (uniaxial object) and those for which no two match (biaxial object).\(^{16}\)

3.2’b Connection to spherical harmonics
This book won’t say much about the spherical harmonic functions \( Y^\ell m(\theta, \phi) \), but take a moment to examine the quadrupole fields (Equation 3.4), and show that:

- The angular dependences of the dipole potentials \( \tilde{\psi}[^1_j] \) are simple linear combinations of \( Y^1m \).
- The angular dependence of \( \tilde{\psi}[^2_{x,y}] \) is the same as that of \( Y^2 \); Both are \( -\frac{1}{3} + \cos^2 \theta \).
- The angular dependence of \( \tilde{\psi}[^2_{x,z}] \) is the same as the linear combination \( Y^{22} + Y^{2,-2} - Y^{20} \).
- The angular dependence of \( \tilde{\psi}[^2_{y,z}] \) is the same as the linear combination \( Y^{22} + Y^{2,-2} + Y^{20} \).
- The angular dependences of \( \tilde{\psi}[^2_{x,y}] \) and \( \tilde{\psi}[^2_{y,z}] \) are the same as the linear combinations \( Y^{21} \pm Y^{2,-1} \).
- (You think about \( \tilde{\psi}[^2_{x,x}] \).

If you’ve studied spherical harmonics, you may have found them at the end of a tortuous derivation in spherical polar coordinates, involving Legendre polynomials, raising/lowering operators, and so on. So it’s remarkable to see them just pop out automatically when we apply Taylor’s theorem to a superposition of \( 1/r \) potentials in cartesian coordinates.

In particular, we found the famous results that for \( \ell = 0 \) there is just one (the monopole potential), for \( \ell = 1 \) there are three (the dipole potentials), and for \( \ell = 2 \) there are five (the quadrupole potentials).

3.3’a Electric dipole moments of fundamental particles
Interestingly, no fundamental particle is known to have a permanent electric dipole moment. A nonzero moment would break “CP” symmetry, and although the Standard Model predicts such breaking, it does so very weakly. For example, the predicted moment for the electron is \( \approx (10^{-38}) \) cm, whereas in 2018 the experimental bound was \( D_E \lessapprox (10^{-29}) \) cm. (In contrast, many fundamental particles, such as electrons and neutrons, have readily measurable magnetic dipole moments, which do not violate CP symmetry.)

3.3’b Nuclear quadrupole moments

\(^{16}\)Note that a uniaxial object may not actually have axial symmetry, that is, continuous symmetry under rotations about an axis.

November 14, 2023; Contents Index Notation Glossary
Nuclei with spin $\geq 1$, such as $^{14}\text{N}$, $^{17}\text{O}$, $^{35}\text{Cl}$ and $^{63}\text{Cu}$, have an electric quadrupole moment. The nuclear quadrupole moment is a measure of the degree to which the nuclear charge distribution deviates from that of a sphere; that is, the prolate or oblate shape of the nucleus. Nuclear quadrupole resonance is a direct observation of the interaction of the quadrupole moment with the local electric field gradient (EFG) created by the electronic structure of its environment. – https://en.wikipedia.org/wiki/Nuclear_quadrupole_resonance
PROBLEMS

3.1 Behind the curtain
Figure 3.2 represents the electric field lines outside a static charge distribution that is overall neutral. (The gray disks cover up singular regions.)

a. Sketch a charge distribution that could result in such a field.

b. If the electric field's magnitude falls with distance as \( \vec{E} \sim r^p \), what is the exponent \( p \)?

3.2 Electrostatic multipole

a. Find the electrostatic potential far away from two point charges, \( q \) and \(-q\) fixed on the \( z\)-axis at \( z = a \), \(-a\) respectively. Give only the first two nonzero terms in the expansion of the potential in powers of \( r/a \). Comment on why your answer “had to” behave this way.

b. Consider point charges \( q \), \(-2q\), \( q \) located on the \( z\)-axis at \( z = a \), \(0\), \(-a\) respectively.
   Find the term in the electrostatic potential at \( r \gg a \) that falls off as \( r^{-4} \) and comment.

3.3 Just a moment
Consider a spherically symmetric charge distribution: \( \rho_q(\vec{r}) = f(r) \) is independent of the polar and azimuthal angles.

a. Such a distribution must have vanishing dipole moment, because no vector other than \( \vec{0} \) can be rotationally invariant. But work this out explicitly from the definition of dipole moment as an integral over the distribution.

b. More precisely, the dipole moment computed about the point of symmetry must be zero.
   Repeat your calculation but this time suppose that the distribution, while spherically

Figure 3.2: See Problem 3.1.
Problems

53

3.4 Tetrahedron
Consider four identical point charges \( q \) rigidly fixed at the vertices of a tetrahedron, and \(-4q\) fixed at its center. The distance from the center to any vertex is \( a \).

a. Find the dipole moment and quadrupole tensor for this distribution. What do these results imply about the behavior of the electric field to leading nontrivial order in \( a/r \)?

[Remark: Figure 3.3 shows a convenient construction of a tetrahedron, beginning with a cube centered on the origin (eight vertices \((\pm \ell, \pm \ell, \pm \ell)\) where \( \ell \) is a length scale related to \( a \)). Four of the cube’s eight vertices are shown as solid dots; they form the vertices of the desired tetrahedron. Since the center of the cube is the only point that is equidistant from all 8 vertices, and the center of the tetrahedron is the only point that is equidistant from all 4 vertices, we can conclude that the center of the cube is also the center of the tetrahedron.]

b. Connect your result to the discussion in Section 3.7.3 (page 47).

c. Find the leading nonzero multipole potential outside this charge distribution.

3.5 Benzene I
We can idealize an isolated aromatic molecule, such as benzene, as follows. Charge \(-q\) is spread uniformly throughout a thin ring (annulus) in the \( xy \) plane, that is, the region \( w < \sqrt{x^2 + y^2} < 2w \). A point charge \(+q\) is all concentrated at the center of the ring. Find the static electric potential far from this charge distribution to leading nontrivial order in powers of \( r^{-1} \) for \( r \gg w \). Also find the static electric field \( \vec{E} \) in the same approximation.

3.6 Benzene II
To improve on Problem 3.5, this time idealize the benzene molecule as six positive point charges \( q \) in the \( xy \) plane at the vertices of a regular hexagon, each a distance \( a \) from the origin. There is also neutralizing point charge \(-6q\) at the origin.

a. Find the dipole and quadrupole moments of this charge distribution in terms of \( q \) and \( a \) and comment. For example, maybe your result has something to do with the fact that benzene is more volatile than water, despite being a more massive molecule.
b. Start over by writing an exact expression for the electric field $\vec{E}$ created by these seven point charges. Show that, when evaluated in the $xy$ plane, the electric field must always itself lie in the $xy$ plane, and hence you can conveniently display it graphically. Get a computer to evaluate it and create an arrow plot.

3.7 Discuss discuss
a. An ellipsoid is defined by the equation $(x/a)^2 + (y/b)^2 + (z/c)^2 \leq 1$. Suppose that it has net charge $q$ uniformly distributed throughout its volume, balanced by a point charge $-q$ at the center. Find the quadrupole tensor of this charge distribution. [Hint: Take the reference point to be its center.]

Suppose further that the ellipsoid in (a) has $a = c = 1$ m and $b = 0.5$ m—it’s “oblate” (squashed). The center of the ellipsoid is placed at the origin of coordinates, in an external electrostatic potential $\psi(\vec{r}) = \vec{\alpha} \cdot \vec{r} + \beta yz + \gamma(x^2 - y^2)z$. Here $\vec{r} = (x, y, z)$ and $\vec{\alpha}$, $\beta$, and $\gamma$ are constants with appropriate dimensions.

b. Under what conditions may we use the multipole expansion to calculate the force on this charge distribution?

c. Assuming the condition in (b) is met, find the force on the ellipsoid exerted by this field to leading order in the multipole expansion.

3.8 Derive a formula for each of the six functions $\nabla^2 (r^{-5} \vec{r}_i \vec{r}_j)$ where $\nabla^2$ is the Laplace operator and the indices $i$ and $j$ each can be 1, 2, or 3. For example, one of these is $\nabla^2 (r^{-5} x^2)$; simplify it and the others. If any of your answers is nonzero, explain how the expression $r^{-5} \vec{r}_i \vec{r}_j$ is admissible as a term in the multipole expansion of the electrostatic field (which is supposed to give us solutions to the Laplace equation).

3.9 Pure versus composite quadrupole

Four point charges are placed in the $xy$ plane as follows:

1,2: Charges $+q$ are placed at points $(0, \pm a, 0)$.

3,4: Charges $-q$ are placed at points $(\pm a, 0, 0)$.

An observer sits at a position $\vec{r}$, with $r \gg a$.

a. Work out the monopole, dipole, and quadrupole moments of this distribution. Is it uniaxial (two eigenvalues are equal) or biaxial (no two are equal)?

b. Substitute the nonzero moment(s) into the general formula to find the far-potential of this static distribution to leading nontrivial order in $1/r$.

c. Differentiate your answer to (b) to get an analytic formula for the electric field (again, to leading nontrivial order). Simplify by evaluating only on the plane $z = 0$.

d. Use a computer to display this vector field, after first normalizing it to unit length. On the same axes, but in a different color, display the exact answer for the electric field of the charge distribution (1–4) and comment.

3.10 Pictures at an exhibition

In this problem, you are to make graphical representations of electrostatic fields corresponding roughly to charge distributions encountered in simple molecules. Section 3.6.7
Problems

55

described a limiting charge distribution whose potential consists of only the dipole term of Equation 3.2 (page 38). By computing minus the gradient of such a function, you can find the corresponding electric field. In this problem, you are to find and display exact expressions for the fields outside pairs of pure dipoles that are not located at the origin.

a. Figure 3.4a represents two dipoles of equal strength, both directed along $+\hat{z}$, located at $(0, 0, \pm a)$. Write an exact expression for the electric field. Use a computer to evaluate this vector field on a suitable grid of points in the $xz$ plane covering the region $-3a < x < +3a$ and $-3a < z < +3a$.\(^{(1)}\) (Arrange your grid so that the two singular points $(0, 0, \pm a)$ are not themselves grid points.) Normalize the vector field to a constant length, to make it easier to see each arrow, and display it. Then get your computer to find and show some representative streamlines\(^{(18)}\) of the vector field in a separate plot. You don’t need a specific value for the length scale $a$ (why not?), but for a molecule it could be, say, 0.2 nm. You also don’t need a specific value for the strength of the dipoles (why not?).

b. Repeat for the situation in panel (b): two dipoles directed along $+\hat{z}$ located at $(\pm a, 0, 0)$.

c. Repeat for panel (c): two dipoles tilted $\pm 60$ deg away from $\hat{z}$ towards the $\pm x$-axis and located at $(\pm a, 0, 0)$. What familiar molecule might this model?

d. Repeat for panel (d): similar to (a), but the dipoles oppose each other. What familiar molecule might this model?

e. Repeat for panel (e): similar to (b), but the dipoles oppose each other. This might model two familiar molecules electrostatically sticking to each other (like what?)

f. The fields in examples (a–c) all fall into one group, and examples (d–e) into a different group, based on some common characteristic. What is it and what does it mean physically?

3.11 3D field line plot

a. Learn how to use a computer to create 3d streamplots, and show them for a pure (point) electric dipole field. One way to do this is to solve a system of ordinary differential equations numerically, then plot the result (for various starting points) as a parametric curve. Look at various viewing angles till you find one that is most informative.

\(^{(1)}\)One way to approach the problem is to evaluate the potential first, then compute its gradient numerically. You’ll get numerically better results, however, if you instead evaluate the electric field symbolically before evaluating it numerically.

\(^{(18)}\)See Section 0.3.1. Changing the normalization does not change the streamlines.
b. Repeat, but this time for a superposition of a pure dipole plus a monopole of charge $Q$. For concreteness, take the ratio $Q/D_E = 1.1/m$.

3.12
[[Not ready]]

3.13 Mystery potential

a. Suppose that far from a source, we measure the electrostatic potential

$$\psi(\vec{r}) = \frac{K}{r^5}(2x^2 - y^2 - z^2),$$

where $\vec{r} = (x, y, z)$, $r = \sqrt{\vec{r}^2}$, and $K$ is a constant. Working in cartesian coordinates, derive a formula for the electric field $\vec{E}(\vec{r})$.

b. Compute $\vec{V} \cdot \vec{E}$ for the field you found in (a) and comment. To what class of potential functions does this one belong?

c. Could this function describe the newtonian gravitational potential far from a localized distribution of mass?

3.14 Animate equipotentials

Consider a pure dipole with $\vec{D}_E = q \left[ \begin{array}{c} 1 \\ 0 \\ 0 \end{array} \right]$. Use a computer to make an animation that serially displays the intersections of the equipotentials with the $xy$ plane. That is, define the dimensionless function $f(\vec{r}) = \frac{4\pi\varepsilon_0}{(1 m)q} \psi(\vec{r})$. Then each frame of your animation should show a curve in the $xy$ plane, a single level set $\{ \vec{r} : f(\vec{r}) = A \}$, for an interesting range of $A$ values (positive and negative).

3.15 Visualize equipotentials in 3D

One way to visually display a function of two variables is to make a contour plot. But often we wish to display a function of three variables, for example, an electrostatic potential.

One approach is sometimes called “z-stack”: We prepare a lot of video frames that successively display contour plots of the function in planes of constant $z$, then present them as an animation. In short, $z$ is represented as time. But that approach can make it difficult to appreciate the overall 3D structure.

In this problem, you’ll take a different approach. The analog of a contour line in 3D is an isosurface, for example, an equipotential: $\{ \vec{r} : \psi(\vec{r}) = A \}$. The problem is that the isosurfaces are nested, so the outer ones hide the inner ones. So try preparing a sequence of video frames that successively display the isosurfaces one at a time in a fixed 3D axes, then present them as an animation. In short, the level $A$ is represented as time.

Some computer math systems offer specialized functions for plotting general isosurfaces, but we won’t need them because of a special circumstance: Specifically, think about an equipotential surface of a pure dipole field:

$$A' = r^{-2} \vec{r} \cdot \vec{D}_E,$$

Generally, a level set of a function is the set of all points where the function has a specified value. An isosurface is a level set that is two-dimensional; a contour line is one that is one-dimensional.

Problem 3.14 may be a useful warmup for this.
where \( A' \) is a constant related to \( A \). The right side has the property that it is a function only of \( r \) (independent of angle) multiplied by a function only of angle (independent of \( r \)). So we can just solve it for \( r \) as a function of \( \theta, \varphi \). For example, if the dipole moment lies along the \( \hat{z} \) direction, then \( r^2 = B \cos \theta \), where \( B = \|D_o\| / A \). For each value of \( B \), we can set up a grid of \( \theta, \varphi \) values, evaluate \( r \), drop the points that have no solution (\( r^2 < 0 \)), and create a 3D surface plot.

a. Carry out the steps just mentioned for one interesting nonzero value of \( 1/B \). Then repeat over an interesting range of evenly spaced \( 1/B \) values and make the animation described earlier.

b. Repeat for an axisymmetric (uniaxial) pure quadrupole field, for example, the one with

\[
\tilde{Q}_{i,j} = \text{(const)} \begin{bmatrix} 1 & 1 \\ 1 & -2 \end{bmatrix}_{ij}
\]

c. Repeat for a nonaxisymmetric (biaxial) pure quadrupole, for example, the one with

\[
\tilde{Q}_{i,j} = \text{(const)} \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}_{ij}
\]

d. Try some more generic quadrupole.

e. Do your four animations have any visual features that correspond to physics ideas? [Hint: It may not be easy to “triangulate” your surfaces, because they’re not given in the form \( z = f(x, y) \). It’s perfectly adequate to simply generate a lot of \( xyz \) triplets and make a 3D scatterplot of them instead. Make your grid dense enough so that the dots appear to merge into a surface.]
CHAPTER 4

Vista: Fluorescence Resonance Energy Transfer

A paradox is only the truth standing on its head to attract attention.

— G. K. Chesterton

4.1 FRAMING: A PRIVATE CHANNEL

We are already in a position to harvest a nontrivial payoff. For many reasons, it is good to be able to observe a macromolecule going about its daily business. Some macromolecules “walk” along “tracks,” carrying a “load.” Others transmit information by sensing conditions and binding or unbinding from partners based on what they have “learned,” and so on. But optical microscopy seems hopeless for the task of observing nanometer-scale movements in molecules that may themselves be just ten nanometers wide—vastly smaller than the wavelength of light.

For decades, the key technique for macromolecular structure was x ray crystallography. However:

- It requires forming a macroscopic crystal. That’s a very different state from the milieu of a macromolecule in a living cell. Moreover, many macromolecules cannot be crystallized.
- Crystallization also immobilizes the molecules, typically forcing them all into a single conformation. It would be better to watch conformational changes, in real time, in order to assess kinetics.

Other high-resolution techniques have other drawbacks. (Electron microscopy rapidly destroys whatever it’s examining, and so on.) Each of these methods has strengths, but it would be great if we could observe macromolecular association and conformational change, in real time, in solution, possibly even inside living cells. Is that asking too much?

Electromagnetic phenomenon: Resonance energy transfer creates a “private communication channel” between two fluorophores with a characteristic dependence on distance, and on the relative orientations of the two molecules and the vector separating them. Physical idea: Electrostatic dipole-dipole interaction explains these dependences.

4.2 FRET HAPPENS
4.2 FRET happens

Figure 4.1: [Experimental data.] Spectral overlap. Curves on left: Excitation and emission spectra of fluorescein, a fluorophore sometimes used as a FRET donor (and in some highlighter pens). Curves on right: Corresponding spectra of Texas red, a fluorophore sometimes used as an acceptor for fluorescein. When a solution containing both molecules is illuminated with light of wavelength shorter than 500 nm (blue bar), fluorescein molecules will be directly excited, but not those of Texas red. Nevertheless, excitation can be passed from donor to acceptor, resulting in acceptor fluorescence, due to the overlap between the donor’s emission spectrum and the acceptor’s excitation spectrum (shaded). To measure the fraction of donor excitations that get transferred, the system can be observed through a filter that eliminates the exciting light but passes light in one or the other of the emission bands. [Data from Johnson et al., 1993.]

4.2.1 Fluorescence microscopy is a versatile tool to image specific molecular actors

Some molecules are fluorescent: They can capture energy from light, wait a long time (typically a nanosecond), and then re-emit the energy as new light. Even a single atom can do this, but for medium-size molecules, there is an interesting twist. A fluorescent molecule, or fluorophore, has a characteristic excitation spectrum, the probability per time per incoming energy flux of getting excited as a function of incoming wavelength. Each fluorophore also has a characteristic emission spectrum of the re-emitted light. The twist is that these spectra are offset, with the emission spectrum peaking at longer wavelengths than the excitation spectrum, a difference called the Stokes shift (Figure 4.1). The energy loss implied by a Stokes shift can be thought of as intramolecular “friction”; like the absorption and emission themselves, its origin is quantum mechanical, and hence this book will treat it as a black-box observed phenomenon.

Certainly the Stokes shift is convenient for microscopists. After attaching a fluorophore to a molecule of interest, a cell can be illuminated with monochromatic light with wavelength in the fluorophore’s excitation peak. Then it can be observed with a filter that passes only light near the fluorophore’s emission peak. Besides eliminating light that was merely scattered from the incoming beam, this fluorescence microscopy technique shows only objects that make a very specific conversion of light—in practice, only the fluorophore of interest, hence showing only the objects to which that fluorophore binds.

---

1 One elegant method genetically encodes a protein resembling a natural protein, but with an extra fluorescent group. Cells with this gene will therefore express (manufacture) a fluorescent version of the protein (a chimera). Another method joins the fluorophore to an antibody that attaches specifically just to the objects under study. The resulting “probe” can be introduced into a living cell (or used in vitro).
4.2.2 Resonant energy transfer defies naïve expectations

A puzzle emerged long before the advent of fluorescence microscopy, however. Starting in the 1920s, experiments began to reveal something odd. Suppose that we dissolve some fluorophores of type 1 (a donor species), with excitation spectrum peaking around $\lambda_{1,\text{ex}}$ and emission spectrum peaking around $\lambda_{1,\text{em}} > \lambda_{1,\text{ex}}$. Now we add a second fluorophore species 2 to the solution (an acceptor), chosen to have excitation spectrum peaking around $\lambda_{2,\text{ex}} \approx \lambda_{1,\text{em}}$ and emission peaking around some longer $\lambda_{2,\text{em}}$. Imagine illuminating with a spectrum that includes $\lambda_{1,\text{ex}}$ but that completely misses the acceptor’s excitation spectrum, so that only the donor gets directly excited (Figure 4.1). We could imagine that some light emitted from 1 would excite 2, instead of leaving the sample or getting absorbed. That two-step process would yield emitted fluorescence around $\lambda_{2,\text{em}}$.

In fact, the process just described is observed. Moreover, in 1996 T. Ha and coauthors managed to capture the faint light from single fluorophores, documenting the effect at the single-molecule level. Excitation transfer of this type is now called fluorescence resonance energy transfer or FRET, and it is a workhorse tool in labs around the world. Other modified versions of FRET have names like bioluminescence resonance energy transfer (BRET) and lanthanide based luminescence resonance energy transfer (LRET).

What’s puzzling is that we’d expect this transfer to be nearly impossible at low concen-

---

2Some authors drop the first word and instead say “RET.”
4.3 Dipole-mediated Transfer

4.3.1 Molecular near fields can be strong

To summarize, well-separated fluorescent molecules can transfer energy efficiently, either in gas phase (separated by vacuum) or in aqueous solution (separated by water molecules). The transfer is highly specific: a “private channel” between a donor and its acceptor that bypasses the many other surrounding molecules. It does not involve direct contact (collision), and we have argued that the mechanism also cannot involve radiation—it is “nonradiative.” What could it be?

The earliest experiments actually did not directly observe excitation transfer; they observed an unexpectedly large loss of incoming polarization, consistent with the two-stage process. Later experiments did show the effect described here (Figure 4.2), with its more direct implication of transfer.
From the very earliest days, researchers had an idea that neutral molecules could be coupled by their surrounding electrostatic fields. A later chapter will show that when an object radiates, it creates an oscillating electric field whose amplitude falls with distance as $r^{-1}$. But Chapter 3 pointed out that dipole and higher fields fall faster with distance, as $r^{-3}$ or higher powers. If you’re far away, this makes the nonradiative fields subleading, suppressed by $r^{-2}$. But the obverse of that statement is that as you approach a molecule, the static fields grow faster than the radiation field. An oscillating dipole can therefore be surrounded by a zone of electric field that is far stronger than we may have expected from radiation. Already in 1925, L. Mensing had incorporated dipole interactions into her theory of spectral line broadening, and many others followed.\(^4\)

Let us boldly hope that, although light absorption and emission are quantum mechanical, perhaps the transfer of energy that interests us may be understood via ideas in Chapter 3. As always, we ask whether this hypothesis leads to quantitative, falsifiable predictions.

### 4.3.2 FRET as a “spectroscopic ruler”

Imagine the initial state as one in which the donor’s electrons form an oscillating dipole, with dipole moment vector depending on the donor’s orientation in space. The resulting near fields then include a dipole component, which in turn applies force to every electron in the vicinity. Most molecules are not resonant with this oscillating field, so the shaking transfers little energy; in particular, the ubiquitous water molecules in solution are hardly affected. But acceptor fluorophores can absorb lots of energy over time, because they do have an excitation at the appropriate resonant frequency. Suppose that the acceptor has a preferred direction, in which its electrons are more free to respond than in other directions. Then the relevant oscillating force is the component of the donor’s dipole field along $\hat{d}$.

We know from mechanics that energy will be transferred from one oscillator to another at a rate proportional to the amplitude of the force squared.\(^5\) The amplitude of the electric force from a dipole on a test particle falls with distance as $r^{-3}$, so we expect this rate to be $\propto r^{-6}$. In addition, the donor has the option to emit its own fluorescence, a process with rate independent of the separation, because that process doesn’t involve the acceptor at all.

We can now ask, what fraction of the donor’s energy loss goes to resonant transfer, relative to the total? From the above discussion, this ratio (the FRET efficiency) must be $\frac{A r^{-6}}{B + A r^{-6}}$, where $A$ and $B$ are constants. Rephrasing gives a prediction for the FRET efficiency:

$$\mathcal{E}_{\text{FRET}} = \left(1 + r^6 \left(\frac{B}{A}\right)\right)^{-1}. \quad (4.1)$$

The constant $B/A$ characterizes the given donor/acceptor pair (in a given solvent) and is typically expressed in terms of a single quantity, the Förster radius $r_F = (A/B)^{1/6}$.

---

\(^4\)For example, the Keesom and dispersion interactions in Section 3.7.3 (page 47) were of dipole-dipole origin.

\(^5\)You’ll recall the details in Problem 4.1.
4.3 Dipole-mediated Transfer

Figure 4.4: [Experimental data.] **FRET efficiency as a function of the separation between donor and acceptor.** $E_{\text{FRET}} = 1$ corresponds to 100% probability that an excitation from a donor will be transferred to an acceptor. Here, the experimenters prepared a series of short DNA molecules each with a donor fluorophore at one end, but with its acceptor at various distances down the chain (horizontal axis). Circles show single-molecule measurements of $E_{\text{FRET}}$. The curve shows the value of $E_{\text{FRET}}$ given by Equation 4.1 for each separation. The value of the Förster radius $r_F$ in that formula was obtained by fitting the data. The quantitative success of the fit supports the hypothesis of FRET mediated by dipole–dipole interaction. [Data from Lee et al., 2005.]

In aqueous solution, fluorophores are distributed with random separations, complicating attempts to test the quantitative prediction of Equation 4.1. However, it is now possible to synthesize “spacers,” molecules of precisely known and adjustable length, and to attach fluorophores to each end. Figure 4.4 shows an experimental test of this sort, with one fitting parameter. Chemical supply catalogs that sell donor/acceptor fluorophores will quote their $r_F$ value.

Thus remarkably, in addition to giving a qualitative explanation of how anything like FRET is possible at all, the dipole-dipole interaction model offers a tool for the quantitative measurement of distances on the nanometer scale, with good time resolution—better than a few seconds for the conformational changes observed in Figure 4.2.

4.3.3 FRET depends on donor and acceptor orientation

Data like those in Figure 4.4 make the dipole-mediated transfer hypothesis look promising. Can we make a more detailed, and hence more falsifiable, prediction?

So far, we have ignored the dependence of dipole-dipole coupling on orientation. Really, however, dipole fields have an angular structure, and moreover we pointed out that the ability of the acceptor to respond to the donor’s field is also anisotropic in general. Free fluorophores in solution undergo rotational brownian motion, which averages these angular dependences and leads to a transfer rate with a single effective Förster radius in Equation 4.1. Something similar may also happen even if the fluorophores are chemically linked to specific points on a single object by flexible connections. More generally, however, the Förster radius does depend on orientation.\(^6\)

---

\(^6\)You’ll work out more aspects in Problem 4.1.
Figure 4.5: [Cartoon; experimental data.] Orientation dependence of FRET. (a) See text. (b) FRET efficiency as a function of spacer length. The curve shown is a fit to data. The molecule is not rigid; instead, thermal motion constantly flexes it. To account for this effect, the fitting function was averaged over fluctuations in the relative angle of donor and acceptor, assumed to have standard deviation of about 55°. The other fit parameters were an overall magnitude factor and the value of $r_F$, which was found to be about 5.3 nm. [Data from Iqbal et al., 2008.]

Your Turn 4A

Use your result in Your Turn 3B (page 41) to make a prediction for the orientation dependence of the transfer rate. Then specialize to the particular case in which both the donor’s dipole moment $\vec{D}_D$ and the acceptor’s polarizability $\hat{d}$ are perpendicular to the vector joining them.

A. Iqbal and coauthors tested the prediction you just made. They used a series of spacers that were short chains of DNA. Short DNA is stiff, with a helical structure, so there was a relative rotation $\varphi$ of the basepair at one end relative to the other that depended on length. Specifically, this angle swept through a full circle each time the DNA length was increased by about 10.5 basepairs. The chemical details of the construct implied that both donor and acceptor’s preferred directions were perpendicular to the long axis of the DNA, and at a definite orientation relative to the terminal basepair (Figure 4.5a). Therefore, we expect that the generic $r^{-6}$ falloff should be periodically modulated as $\cos^2 \varphi$, and hence periodic, repeating each time $\varphi$ increased by $\pi$. The data shown in the figure indeed show such modulation, partially washed out by orientational fluctuations (the molecular construct was not perfectly rigid).

4.4 PLUS ULTRA

A “FRET pair” (donor and acceptor) can be used to report on conformational changes in a single macromolecule, for example, when a molecular motor steps. That conformational change can itself be a report on some other condition, such as the presence of calcium,
yielding “FRET-based calcium reporters” and so on

Alternatively, each member of the pair can be attached to its own macromolecule, perhaps an internal signaling molecule and its target, to give real-time reports on the location and timing of their binding. The ensuing time series can also be correlated with environmental changes read out by the signaling molecule, in order to tease out both the control network mechanism and its kinetics. The clever applications are endless.

FURTHER READING

Intermediate:

Many more applications of FRET: Nadeau, 2018, chap. 8; Nelson, 2017, chap. 2.
en.wikipedia.org/wiki/Single-molecule_FRET.
Quantum theory of FRET, and when the classical treatment gives a good approximation:

Technical:
The quantum theory of FRET was first presented in a very condensed form by J. R. Oppenheimer: Oppenheimer, 1941.
Spectroscopic ruler: Sindbért et al., 2011.

PROBLEMS

4.1 Classical model of FRET

We can get some insight into fluorescence resonance energy transfer by using ideas from newtonian mechanics. Imagine an oscillator representing the charge cloud (electric dipole moment) of a donor fluorophore. The donor gives rise to an electrostatic force on a second oscillator, which represents the acceptor fluorophore. Suppose that this force \( f_D(t) \) has fixed angular frequency \( \omega_D \) (determined by the donor’s excited state), and amplitude \( J \) (determined by the donor’s state and the distance to the acceptor):

\[
f_D(t) = J \cos(\omega_D t) .
\]

(4.2)

We model the acceptor’s electron cloud as a point object with mass \( m \). It’s attached to a fixed object (representing the molecule’s heavy nuclei) by a spring, with spring constant \( k \). Moreover, the acceptor slowly dissipates energy to “friction,” which represents energy loss from the acceptor, including by fluorescence. Calling the friction constant \( \eta \), Newton’s law \( f_{tot} = ma \) states that the donor’s position \( x(t) \) obeys

\[
m \frac{d^2x}{dt^2} = -kx - \eta \frac{dx}{dt} + f_D.
\]

(4.3)

To simplify this equation, define new symbols \( \omega_L = \sqrt{k/m}, \eta = \eta/m \), and \( L = J/m \), and eliminate \( k, \eta, \) and \( J \) by writing them in terms of the new quantities.
a. After a short transient, the solution \( x(t) \) will oscillate at angular frequency \( \omega_d \). So consider the trial solution \( x(t) = A \cos(\omega_d t) + B \sin(\omega_d t) \). Find the constants \( A \) and \( B \) in terms of \( L, \eta, \omega_d, \) and the acceptor’s resonant frequency \( \omega_A \).

b. In the steady state that we are studying, the rate at which the acceptor gets energy from the donor must equal the rate at which it loses energy to dissipation, which is \( \mathcal{P} = \eta \frac{dx}{dt}^2 \). Evaluate this for your solution.

c. The quantity you found in (b) is always positive, but it oscillates. We only need its time-average \( \langle \mathcal{P}(\omega_d, \omega_A) \rangle \), which is given by a simpler expression than the answer to (b). Derive an expression for that.

d. Actually, the donor and acceptor are not in precisely known states: Rather, each is a molecule that moves with a distribution of possible states, with varying values of \( \omega_d \), \( \omega_A \). The average rate of energy transfer is then the average of the quantity you found in (c), weighted by the corresponding probability density functions \( \varphi_d(\omega_d) \) and \( \varphi_A(\omega_A) \):

\[
\langle \mathcal{P} \rangle = \int d\omega_d \varphi_d(\omega_d) \int d\omega_A \varphi_A(\omega_A) \langle \mathcal{P}(\omega_d, \omega_A) \rangle.
\]

To simplify this expression, suppose that the damping \( \eta \) is very small. Then your expression from (c) is very sharply peaked near \( \omega_d = \omega_A \). Exploit this fact by letting

\[
\omega_d = \bar{\omega} - \frac{1}{2}\Delta \omega, \quad \omega_A = \bar{\omega} + \frac{1}{2}\Delta \omega,
\]

and changing integration variables from \( \omega_d, \omega_A \) to \( \bar{\omega}, \Delta \omega \). Then approximate your answer to (c) by replacing \( \Delta \omega \) by 0 everywhere, except for the one term in the denominator responsible for making the sharp peak. With this approximation, you can readily do the integral over \( \Delta \omega \).

e. The donor creates a dipole field, which shakes charges on the acceptor. Imagine the acceptor dipole as having a fixed axis \( \hat{D}_A \) and a charge \( q \) that is only able to move along that axis. Then the force driving that charge’s motion is the product of the charge times the component of the donor’s electric field along \( \hat{D}_A \). From this information, the behavior of dipole fields, and your calculations, comment on how the energy transfer rate depends on the separation and relative orientation of donor and acceptor.

f. In the experiment sketched in Figure 4.5, the donor and acceptor dipoles are both oriented perpendicular to the separation vector, but at various angles to each other. Specialize your answer in (e) to this situation.

---

7 If you prefer, you can use complex exponential notation (Section 18.7, page 278).
[Abbé Nollet] speaks as if he thought it presumption in man to propose guarding himself against the thunders of Heaven! Surely the thunder of Heaven is no more supernatural than the rain, hail or sunshine of Heaven, against the inconvenience of which we guard by roofs and shades without scruple.
— Benjamin Franklin

5.1 FRAMING: LEVEL SETS

Chapter 2 gave a general solution to Poisson’s equation. Doesn’t that say everything there is to say about electrostatics?

Unfortunately, Equation 2.7 (page 31) only tells us the potential if we know the locations and magnitudes of every charge. Frequently, however, we deal with multitudes of mobile charges, for example, in a conductor, so we don’t know up front where each one is, even when they come to static equilibrium. We may nevertheless have some boundary conditions to guide us, for example, the one that says the electrostatic potential is constant throughout a conductor. Hence we often need to go back to Poisson’s equation, and solve it with specified boundary conditions.

Electromagnetic phenomenon: The tip of a nearfield scanning optical microscope generates huge electrostatic fields localized to nanometer regions.

Physical idea: The choice of a curvilinear coordinate system with appropriate level sets reduces the Laplace equation to manageable ordinary differential equations.

5.2 SEPARATION OF VARIABLES IN THE LAPLACE EQUATION

Poisson’s equation is a partial differential equation, and hence not as easily solved as ordinary differential equations. Numerical solution can be useful, but it can also break down in singular situations, such as a sharply pointed conductor. Yes, there are advanced numerical methods, but whenever there’s an exact solution available, we should cherish that case and add it to our (short) catalog of analytically tractable situations. One good trick is separation of variables, which can effectively bring us down to ordinary differential equations.

Section 0.2.1 said that cartesian coordinates are “good” because the Maxwell equations look exactly the same in any cartesian coordinate system. We will say much more on that subject later. But you already know that some non-cartesian coordinate systems are
also “pretty good,” and that such a system can be extremely convenient for certain kinds of problems, for example, those with certain symmetries. Here we will sharpen the notion of “pretty good” to introduce systems for which the Laplace operator may be solved by the method of separation of variables.

Cartesian coordinates have the property that the surfaces with one coordinate fixed and the other two varying (level sets for the first coordinate) form a stack of parallel planes. That’s convenient for boundary-value problems in which a boundary, for example a conductor, is a plane. But many problems have boundaries that don’t look so simple in cartesian coordinates. So we will find some other coordinate systems, collectively called \textbf{curvilinear}, in which the Laplace operator can be simplified in a way analogous to cartesian, but the surfaces with one coordinate constant are not necessarily planar. Specifically, we’ll find useful examples where those surfaces are spheres, cylinders, or ellipsoids.

5.3 **FAMILIAR EXAMPLES**

5.3.1 **Cartesian coordinates**

The Laplace operator is the sum of a term not involving \(x, z\), plus a term not involving \(x, y\), plus a term not involving \(r, \varphi\). The payoff is that we can find many solutions of the form \(A(x)B(y)C(z)\), where each factor is a function of one variable and obeys an ordinary differential equation: \(A'' = \kappa A, B'' = \lambda B, C'' = \nu C\), where \(\kappa + \lambda + \nu = 0\). This strategy is called \textbf{separation of variables}.

5.3.2 **Plane polar coordinates**

For simplicity, let’s warm up with just two dimensions. Let \(x = r \cos \varphi\) and \(y = r \sin \varphi\) as usual. You already know what the Laplace operator looks like in these coordinates, but let’s redo that derivation in a way that will generalize easily.

Define two vector fields \(\vec{e}_{(r)}(r, \varphi)\) and \(\vec{e}_{(\varphi)}(r, \varphi)\) as the motions we make when we vary one or the other of the new coordinates:

\[
\vec{e}_{(r)} = \frac{\partial \vec{r}}{\partial r}, \quad \vec{e}_{(\varphi)} = \frac{\partial \vec{r}}{\partial \varphi}.
\] (5.1)

It will soon be convenient that these two vector fields are everywhere perpendicular to each other. Note that the first one is the same as the unit vector \(\hat{r}\), but the second is not the same as \(\hat{\varphi}\). Instead, \(\vec{e}_{(\varphi)} = r \hat{\varphi}\), as one might guess on dimensional grounds.

We want to formulate the Laplace operator in terms of the new variables. Let \(f\) be a function on the plane, and abbreviate \(\vec{\nabla} f = \partial f / \partial r, \vec{\nabla}_{\varphi} f = \partial f / \partial \varphi\), and so on. The cartesian components of the gradient can be written via the Chain Rule as

\[
\vec{\nabla} f = J \begin{bmatrix} \vec{\nabla}_{(r)} f \\ \vec{\nabla}_{(\varphi)} f \end{bmatrix}, \quad \text{where} \quad J = \begin{bmatrix} \frac{\partial r}{\partial x} & \frac{\partial r}{\partial y} \\ \frac{\partial \varphi}{\partial x} & \frac{\partial \varphi}{\partial y} \end{bmatrix}.
\] (5.2)

Here’s a useful trick to get an expression for the Laplace operator re-expressed in terms of our new coordinates. Let \(g\) be any function and let \(f\) be a function that is zero...
everywhere except in some small region. Then Equation 5.2 gives that

\[ \int d^2 r \, \nabla f \cdot \nabla g = \int d^2 r \, \left[ \nabla_r f, \nabla_f g \right] J^1 \left[ \nabla_r g \right]. \quad (5.3) \]

However, we also have\(^1\)

\[ \int d^2 r \, \nabla f \cdot \nabla g = \int d^2 r \, \left[ \nabla \cdot (f \nabla g) - f \nabla^2 g \right] = - \int d^2 r \, f \nabla^2 g. \quad (5.4) \]

In the last step, we used the divergence theorem to express the first term as an integral over the boundary. That term is zero because of our assumption about \(f\).

We have found two expressions that must agree for any choice of \(f\) and \(g\). To derive a formula for \(\nabla^2\), then, we will just manipulate the right-hand side of Equation 5.3 until there are no more derivatives on \(f\), then compare to the right-hand side of Equation 5.4.

First we need an explicit formula for the \(2 \times 2\) matrix \(J^1\). It’s messy to compute \(J^1\) directly, because after computing \(\partial \hat{s}/\partial x\) and so on, we must then re-express everything as functions of \(r\) and \(\varphi\). Luckily, there’s a shortcut to make that step unnecessary. Note that \(J^{-1}\) is the matrix

\[ \begin{bmatrix} \partial x/\partial r & \partial y/\partial r \\ \partial x/\partial \varphi & \partial y/\partial \varphi \end{bmatrix}. \]

(Proof: The stated matrix transforms cartesian derivatives to polar, the opposite of what \(J\) does.) The nice property about \(J^{-1}\) is that its rows are the components of \(\vec{e}_i(\varphi)\) and \(\vec{e}_j(\varphi)\) defined by Equation 5.1. Thus, we may write

\[ J^{-1}(J^1)^\dagger = \begin{bmatrix} ||\vec{e}_i(\varphi)||^2 & \vec{e}_i(\varphi) \cdot \vec{e}_j(\varphi) \\ \vec{e}_j(\varphi) \cdot \vec{e}_i(\varphi) & ||\vec{e}_j(\varphi)||^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & r^2 \end{bmatrix}. \]

We need \(J^1J\), which is the inverse of the preceding result. But the inverse of a diagonal matrix is easy. That’s the benefit we get from the fact that the \(\vec{e}_i(\varphi)\)’s are everywhere perpendicular to each other.

Now we can return to Equations 5.3–5.4. For the first of these, we use the result just found for \(J^1J\), then integrate by parts:\(^2\)

\[ \int d^2 r \, \nabla f \cdot \nabla g = \int (\partial r \, d\varphi) \frac{\partial f}{\partial \varphi} \hat{\nabla}_r f \hat{\nabla}_g + (\hat{\nabla}_r f) r^{-2} \hat{\nabla}_g \hat{\nabla} \varphi g = - \int (dr \, d\varphi) \left( f \frac{\partial}{\partial r} (r \hat{\nabla}_r g) + f r^{-1} \frac{\partial}{\partial \varphi} \hat{\nabla}_g \hat{\nabla} \varphi g \right). \]

We wish to recast this expression in a form resembling the right side of Equation 5.4, so multiply and divide by \(r\):

\[ \int d^2 r \, \nabla f \cdot \nabla g = - \int (dr \, d\varphi) f \left[ r^{-1} \frac{\partial}{\partial r} (r \hat{\nabla}_r g) + r^{-2} \frac{\partial^2 g}{\partial \varphi^2} \right]. \quad (5.5) \]

\(^1\)See Your Turn 0C (page 10).

\(^2\)Again, boundary terms are zero: For the first term, we consider the \(\varphi\) integral first and apply parts integration; a complete circle of \(\varphi\) has no boundary. For the second term, we consider the \(r\) integral first and apply parts integration; but \(f\) was assumed to be zero at the outer boundary.
Equation 5.4 says that the last expression equals $- f \, d^2 r \, f \, \nabla^2 g$ for any function $f$ that vanishes outside a small region. For example, $f$ could be a bump function localized anywhere. The only way that these expressions could be equal for any such $f$ is if the terms in square brackets of Equation 5.5 are equal to $\nabla^2 g$. This is a familiar formula:

$$\nabla^2 g = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial g}{\partial r} \right) + r^{-2} \frac{\partial^2 g}{\partial \phi^2}.$$

5.3.3 Plane polar payoff

If we have a circularly-symmetric, 2D problem, we can entertain trial solutions of the form $\psi(\vec{r}) = A(r)B(\phi)$. Then $\nabla^2 \psi = 0$ becomes

$$0 = \frac{1}{B} \frac{\partial^2 B}{\partial \phi^2} + r \frac{\partial}{\partial r} \left( r \frac{\partial A}{\partial r} \right).$$

The first term is completely independent of $r$. The second term is completely independent of $\phi$. Their sum is the constant 0, so each term must separately be a constant. That reduces our problem to two decoupled ordinary differential equations.

If moreover our boundary conditions can be stated simply in these coordinates, for example as $A = 1$ on some sphere, then we win.

5.3.4 Another hint about general relativity

It is definitely not the case that the Laplace operator can be written as $\partial^2 / \partial r^2 + \partial^2 / \partial \phi^2$! Einstein asked himself, “What’s special about some coordinate systems (such as cartesian) that makes the Laplace operator look simpler in them than in others (such as polar)?” Following that road led him into general relativity.

For now, we just notice that in polar coordinates the Laplace operator still looks fairly simple, whereas in completely general coordinates it does not.

5.3.5 Three dimensions

**Your Turn 5A**

Run through all these steps for cylindrical and spherical polar coordinates, to see how they yield the rather mysterious formulas for gradient and laplacian found on the inside cover of any E&M textbook.

5.4 A SPHERICAL CONDUCTOR IN A UNIFORM FIELD

Consider a spherical conductor of radius $R$ between two distant, infinite, flat, parallel, charged plates. We choose an origin of coordinates centered on the center of the sphere and set up spherical polar coordinates with axis along $\hat{z}$, which is perpendicular to the planes.
At the sphere, \( \psi(r = R) \) must be independent of \( \theta \) and \( \varphi \); by adding a constant we may take its value to be zero. Far from the sphere, we get the same uniform electric field that we’d have had from the charged plates alone (without the sphere), so

\[
\psi \to Cz \quad \text{at } r \gg R, \tag{5.6}
\]

where \( C \) is a constant related to the areal charge density on the plates. Now we want \( \psi \) everywhere (not just far from the sphere).

Our problem isn’t spherically symmetric, but at least it’s axially symmetric, so we get a shortcut: \( \psi \) will be independent of azimuthal angle \( \varphi \). The boundary condition at the sphere is simple in spherical polar coordinates (the sphere is a surface of constant \( r = R \)), so let’s seek a \( \varphi \)-independent solution of the form \( A(r)B(\theta) \). Your answer to Your Turn 5A, combined with the same reasoning as was used in Section 5.3.1, then implies that in order to solve the Laplace equation in the space between sphere and plates, we need functions that satisfy

\[
A^{-1}(r^2 A')' = \lambda \quad \text{for } r \geq R \quad \text{and} \quad B^{-1} \frac{1}{\sin \theta}(\sin \theta B')' = -\lambda \quad \text{for } 0 \leq \theta \leq \pi.
\]

In the first equation, prime means \( d/dr \); in the second one, prime means \( d/d\theta \). Now change variables from \( \theta \) to \( \mu = \cos \theta \), so \( d\mu = -\sin \theta d\theta \). Thus, the second equation becomes the **Legendre equation**:

\[
B^{-1} \frac{d}{d\mu} \left( (1 - \mu^2) \frac{dB}{d\mu} \right) = -\lambda. \tag{5.7}
\]

One solution is \( B = \text{const} \), which has eigenvalue \( \lambda = 0 \). But that’s a spherically symmetric solution, and our distant boundary condition Equation 5.6 is not spherically symmetric. The next most complicated solution to Equation 5.7 is \( B(\mu) = \mu \), which has eigenvalue \( \lambda = 2 \). Put that back into the equation for \( A \):

\[
(r^2 A')' = 2A.
\]

This equation is homogeneous, so we look for a power-law solution: \( A(r) = r^p \). Substituting shows that \( p = 1 \) or \(-2\) both work, so we try an unknown linear combination of those solutions:

\[
\psi(r, \theta) = (\alpha r + \beta r^{-2}) \cos \theta.
\]

Any expression of this form does approach \( \alpha z \) at \( r \to \infty \), as required by Equation 5.6.\(^3\)

And we can satisfy the inner boundary condition by choosing \( \beta = -R^3 \alpha \). That exhausts our remaining freedom, so we have found a unique solution.

We’re done. The second term is familiar (it’s an electric dipole potential), but the first term is new: A multipole expansion would have missed it because it does not drop off with distance.\(^4\)

---

\(^3\)The apparent singularity at \( r \to 0 \) is not a problem because this solution is only to be used outside the sphere.

\(^4\)Physically, the charged plates at infinity violate the multipole expansion’s assumption that all charges are confined to a small zone.
5.5 LIGHTNING ROD VIA ELLIPSOIDAL COORDINATES

Benjamin Franklin was not the first to discover that electric discharges tend to occur at sharp points. It’s not at all clear that he even did the dangerous and stupid kite experiment that he almost, but not quite, claimed to have done. (Others actually did it, and not all survived.) Franklin’s breakthrough was to connect the abstractions of natural philosophy to the urgent practical matter of saving lives.\(^5\)

We can think of a sharply pointed spike as a limit of a family of ellipsoidal conductors. But how shall we find the electric field just outside an ellipsoid? The multipole expansion only gives the potential far away from an object, and even then requires that we know the charge distribution in advance. Spherical harmonic expansion goes bad in the limit of interest, where the ellipsoid is very pointy.\(^6\) Finite-grid numerical solution also loses accuracy in that limit. Conformal transformation only works for 2D problems.

Really, we’d like an exact solution. We saw earlier that spherical polar coordinates enable that goal for spherical conductors. In Problem 5.1, you’ll find a different curvilinear system in which the level sets of one of the coordinates are ellipsoids. By following the steps in this chapter, you’ll find solutions to the Laplace equation by separation of variables in those coordinates as well, and hence get an exact solution for the lightning-rod problem almost as readily as in Section 5.4.

\(^5\)Some people objected—lightning strikes were manifestations of divine will, which humans would defy at their peril. Franklin was persistent (see the epigraphs on pages 26 and 67).

\(^6\)It’s also bad in the opposite limit, where the ellipsoid is squashed very flat to a thin conducting pancake with a sharp edge (Problem 8.3).
5.6 OTHER VECTOR OPERATORS

So far, we have restricted attention to the Laplace operator, but the rest of vector calculus can be cast into curvilinear coordinates when that’s helpful. Just remember that if you use someone else’s formulas, you must be sure you know how they work.

For example, we will later need the divergence of the following vector field:

\[ \vec{V}(r) = \frac{1}{r} \hat{x}_r e^{ikr}. \] (5.8)

Here \( r \) is distance from the origin, and \( k \) is a scalar constant. We will first do this the hard way, just to highlight how much easier our second approach is.

5.6.1 Hard way

The hard way at first seems easier: Just look in any (other) book and find the formula

\[ \vec{V} \cdot \vec{V} = r^{-2} \frac{\partial}{\partial r} (r^2 \vec{V}_r) + \frac{1}{r \sin \theta} \left( \frac{\partial}{\partial \theta} (\sin \theta \vec{V}_\theta) + \frac{\partial \vec{V}_\phi}{\partial \phi} \right). \] (5.9)

But it’s tricky to apply this formula to Equation 5.8 correctly! Note that \( \vec{r} = \begin{bmatrix} \sin \theta \cos \phi \\ \sin \theta \sin \phi \\ \cos \theta \end{bmatrix} \), so

\[ \hat{x}_r = \vec{r} \cdot \hat{x} = \sin \theta \cos \phi, \quad \hat{x}_\theta = \hat{\theta} \cdot \hat{x} = \cos \theta \cos \phi, \quad \hat{x}_\phi = \hat{\phi} \cdot \hat{x} = -\sin \phi. \]

After you substitute these into Equation 5.8 to get \( \vec{V}_r, \vec{V}_\theta, \) and \( \vec{V}_\phi \), then apply the black-box formula Equation 5.9, you still must do a lot of algebra to find

\[ \vec{V} \cdot \vec{V} = \sin \theta \cos \phi (r^{-2} + ik/r) e^{ikr}. \]

5.6.2 Easy way

Instead of using a black-box formula, let’s do it from scratch. Use the product rule, \( (f \hat{x}) = \hat{x} \cdot \vec{V} f + f \vec{V} \cdot \hat{x} \) and take \( f = r^{-1} e^{ikr} \). The second term is zero because the cartesian components of \( \hat{x} \) are all constants (0 or 1). Thus,

\[ \vec{V} \cdot \vec{V} = \hat{x} \cdot \vec{r} \frac{\partial}{\partial r} (r^{-1} e^{ikr}) = \sin \theta \cos \phi (r^{-2} + ik/r) e^{ikr}. \]

5.7 PLUS ULTRA

There are a total of 11 coordinate systems in which the 3D Laplace equation can be solved by separation of variables, plus two more that are almost as good. See Further Reading.

FURTHER READING

Semipopular:

7See Your Turn 0C (page 10).

*Intermediate:*
About the famous list of 13 good coordinate systems: See Weisstein, Eric W. ‘Laplace’s Equation’: mathworld.wolfram.com/LaplacesEquation.html. Also see books: Landau & Lifshitz, 1981, §48; Arfken et al., 2013; and Morse & Feshback, 1953.
Ben Franklin: Cohen, 1990; Franklin, 1941.
5.1 **NSOM probe**

The chapter motivated the study of a long, thin metal probe in a uniform background electric field, which is relevant to apertureless nearfield scanning optical microscopy. Near-field scanning optical microscopes generate huge electrostatic fields localized to nanometer regions.

We can define an ellipse as the locus of points in the \(xz\) plane that solve

\[
\left(\frac{x}{\alpha}\right)^2 + \left(\frac{z}{\beta}\right)^2 = 1,
\]

where the constants \(\alpha\) and \(\beta > \alpha\) are called the “seminiminor” and “semimajor” axes, respectively. Thus, \(2\beta\), the major axis length, is the distance between the two most distant antipodal points (the “poles”), and \(2\alpha\), the minor axis length, is the distance between the two least distant antipodal points.

Consider two points \(P_\pm\) on the \(z\) axis, located at \(z = \pm \sigma\). For any other point, let \(r_\pm\) be the distances from that point to \(P_\pm\). We can specify that point either by its \(x\), \(y\), \(z\) values, or its cylindrical polar coordinates \(\rho\), \(\varphi\), \(z\), or by new coordinates \(\xi\), \(\eta\), and \(\varphi\). Here \(\varphi\) is the same as in cylindrical coordinates and

\[
\xi = \left(\frac{r_+ + r_-}{2\sigma}\right) \quad \eta = \left(\frac{r_+ - r_-}{2\sigma}\right).
\]

(a) Show that the surface \(\{\xi = \xi_0\}\) is what you get by rotating an ellipse about its axis (see Figure 5.1). Find its major and minor axis lengths in terms of \(\sigma\) and \(\xi_0\).

We wish to find the field outside a conductor whose surface is the one in (a). The conductor carries zero net charge, but there is a background electrostatic field that’s uniform at infinity. To get started, we need some more math.

(b) Express \(\rho\) and \(z\) in terms of \(\xi\) and \(\eta\). [Hint: Express \(\xi\eta\) and \((\xi^2 - 1)(1 - \eta^2)\) in terms of \(\varphi\) and \(\sigma\), then think.]

(c) Thus, express \(x\), \(y\), \(z\) in terms of \(\xi\), \(\eta\), and \(\varphi\). Differentiate to find the vector \(\vec{e}_{(\xi)}\), \(\vec{e}_{(\eta)}\), and \(\vec{e}_{(\varphi)}\). These three vectors have a convenient property similar to the one found in Section 5.3.2 for plane polar coordinates—what is it?

(d) Use (c) to express the volume element \(d^3r\) in terms of \(d\xi\) \(d\eta\) \(d\varphi\). Find the region in the \(\xi\eta\) plane corresponding to the region outside the surface in (a).

(e) Use (c,d) to express the integral \(\int \vec{\nabla} \cdot \vec{\nabla} \psi\) in the coordinates \(\xi\), \(\eta\), and \(\varphi\). Here \(\psi\) is any function independent of \(\varphi\), whereas \(\nabla\), also independent of \(\varphi\), is nonzero only in some small region of \(\xi\) and \(\eta\).

(f) Use integration by parts to work out the Laplace operator \(\nabla^2\psi\) in these coordinates, for the case where \(\psi\) is independent of \(\varphi\).

You’re ready to begin the problem, which is to find the electrostatic potential in the region outside the conductor, subject to the same boundary conditions used in Section 5.4:

\[
\psi = 0 \text{ on the surface,} \quad \psi \to -E_\infty z \text{ far away.}
\]

We seek an exact solution \(\psi = A(\xi)B(\eta)\) by separation of variables.
g. Translate the boundary conditions above into conditions on \( A \) and \( B \). Find a solution to the equation for \( B \) meeting those conditions.

h. Now that you know the dependence on \( \eta \), write the required ordinary differential equation and boundary conditions on the function \( A \).

i. The equation is second order, so it has two independent solutions. You can readily guess one of them from the boundary condition at infinity, and substitute to confirm that it works.

j. But we need the other solution too, in order to satisfy the surface boundary condition. You may not remember how to find the other solution, but symbolic mathematical systems like the free Wolfram Alpha do. So ask one of them (unless you know all about special functions).

k. Finish the problem: Work out the magnitude of the electric field just outside the conductor at its two poles, and compare this value to the applied \( E_\infty \).

l. Consider a conductor with major axis length 100 \( \mu \text{m} \) and minor axis length 0.5 \( \mu \text{m} \) and evaluate your expression in (k) for the field ratio numerically. Then make a contour plot of the normalized electrostatic potential \( \psi/E_\infty \) in the \( xz \)-plane.

Here is a related problem that’s easy after you construct the above formalism:

m. Now consider a metal ellipsoid carrying nonzero net charge \( q \) but totally isolated, that is, the electric field approaches zero at infinity. Adapt the procedure of parts (a–j) to find the exact solution for the potential. Then make a contour plot of the electrostatic potential \( \psi/q \) in the \( xz \)-plane for the same geometry as in (l).

[Remark: At optical frequencies, most metals are not really well described by our assumption of perfect conductors. Moreover, the geometry of a probe approaching a surface is probably closer to a hyperboloid near a plane than to the geometry assumed in this problem. Nevertheless, \( \xi \eta \) coordinates are still useful in realistic treatments of NSOM probes and their field-focusing properties.]

5.2 Razor’s edge

A thin metal plate in vacuum is placed in the half-plane \( y \approx 0, x < 0 \) for all \( z \). Thus, the edge of the plate is the \( z \) axis. The electrostatic potential \( \psi \) is constant everywhere on the plate, but the plate may be charged.

We can seek a solution by using separation of variables in cylindrical coordinates, for which the plate occupies the half-plane with \( \varphi = \pm \pi \):

\[
\psi(\rho, \varphi, z) = f(\rho) \cos(\varphi/2).
\]

This trial solution for the angular dependence satisfies the boundary condition \( \psi(\rho, \pm \pi, z) = 0 \). Write and solve the equation satisfied by the radial function \( f \). Comment on how your solution behaves near the edge.
CHAPTER 6

Capacitors

6.1 FRAMING: DIELECTRICS

Section 2.1 pointed out that what makes electrodynamics physics is that we must constantly seek idealizations of systems that are too complex to handle explicitly. Thus, in an electron beam we may be able to apply Newton’s laws of motion with electrostatic forces to each electron individually, but many other situations involve condensed (solid or liquid) matter, which is packed with too many charges to handle explicitly. Section 2.4 already introduced one such idealized element: a good conductor. This chapter will introduce another one that is useful in many real situations: a dielectric material.¹

Electromagnetic phenomenon: A charged capacitor will pull dielectric material into its gap.
Physical idea: Polarization of the medium acts as a “spring in parallel with” the vacuum field energy, reducing total energy at fixed charge.

6.2 PARALLEL PLATES IN VACUUM

Charge $q$ is placed on a flat planar conductor with area $\Sigma$. Charge $-q$ is placed on another such conductor, parallel and a distance $w$ away from the first in the $+x$ direction. Both conductors are much bigger in $y$ and $z$ than $w$, so we will neglect edge effects. By symmetry, the electric field must point along $\hat{x}$. Let $\sigma_q = q/\Sigma$ be the areal charge density on the left plate.

Use the electric Gauss law (Equation 0.1, page 2) to find that between the planes, $\vec{E}_x = \sigma_q/\varepsilon_0$. Integrate $-\vec{E}$ along $x$ to find the potential throughout the gap, and its total change $\Delta\psi = \psi(0) - \psi(w) = \sigma_q w/\varepsilon_0$. We define the capacitance as the constant of proportionality relating charge and potential:

$$C = q/\Delta\psi. \quad (6.1)$$

Mnemonic: If you have large “capacity,” you can store lots of charge without developing a big potential. That’s why $q$ is in the numerator and $\Delta\psi$ is in the denominator of the definition. The natural SI unit for capacitance is therefore one coulomb per volt, which is called the farad: $1 \text{F} = 1 \text{coul}/\text{V}$.

For the system just discussed, $C = C/\Sigma = \varepsilon_0 \Sigma/w$. Thus, the capacitance per plate area is $C = \varepsilon_0/w$.

¹Just don’t confuse “dielectric material” with “dialectical materialism.”
6.3 THE ENERGY STORED IS PROPORTIONAL TO VOLUME

We can now imagine pulling a charge $dq$ away from the negative plate and depositing it on the positive plate. If $dq$ is positive, then we must do work against the electric field to accomplish this: $(dq)\Delta\psi = dq(q/C) = d(\frac{1}{2}q^2/C)$. If we wish to build up charge starting from zero, then we must do a total amount of work

$$E = \int_0^Q dE = \int_0^Q d(\frac{1}{2}q^2/C) = \frac{1}{2}Q^2/C.$$ 

Rephrasing using Equation 6.1 gives the stored electrostatic potential energy as

$$E/(\text{volume}) = \frac{\varepsilon_0}{2}||\vec{E}||^2.$$

That’s interesting: The total energy is proportional to the volume, as though it were stored in empty space with a density depending quadratically on the field:

*The equations of electrostatics appear to be compatible with energy conservation if we attribute energy density to fields in empty space.*

(6.3)

We’ll need to do a lot more work before we can be confident about this suspicion, however. It is useful to work out that $e/(4\pi\varepsilon_0) \approx 2.3 \cdot 10^{-28}$ J m, where $e$ is the proton charge.

**Your Turn 6A**

a. Adapt the preceding argument to find the work that must be done to bring total charge $q$ from infinity onto a spherical shell of radius $R$.

b. Then evaluate the expression in Equation 6.2 everywhere outside the shell, integrate it over that region, and compare the result to (a).

c. A heavy atomic nucleus may contain charge of around $100e$ confined to a sphere of radius $10$ fm. Suppose that nucleus fissions into two equal-size fragments, each a sphere with about half the charge, and with radius smaller by a factor of $2^{1/3}$. Approximate by supposing that the charge sits on the surfaces of the spheres. Compute the change in electrostatic self-energy and compare to the typical scale involved in chemical reactions.

Electrostatic self-energies can be huge, so it’s normally a good approximation to suppose that macroscopic objects are neutral:

**Ex.** Consider a raindrop of radius $R = 1$ mm suspended in air. How much work would be needed to remove just one electron from just 1% of the water molecules in the drop?

**Solution:** Removing an electron leaves some water molecules electrically charged. These charged water molecules (ions) migrate to the surface of the drop to get away from one another, thereby forming a shell of charge of radius $R$. Your solution to Your Turn 6Aa gives the electrostatic potential energy of such a shell as $\frac{1}{2}q\psi(R)$, or $q^2/(8\pi\varepsilon_0R)$. The charge $q$ on the drop equals the number density of water molecules, times the drop
volume, times the charge on a proton, times 1%. Squaring gives

\[
\left( \frac{q}{e} \right)^2 = \left( \frac{10^3 \text{ kg} \cdot 6 \cdot 10^{23}}{0.018 \text{ kg} \cdot \frac{4\pi}{3} (10^{-3} \text{ m})^3 \times 0.01} \right)^2 = 1.9 \cdot 10^{36}.
\]

Multiplying by \(2.3 \cdot 10^{-28} \text{ J m} \) and dividing by \(2R\) yields about \(2 \cdot 10^{11} \text{ J}\).

Two hundred billion joules is a lot of energy! And indeed, macroscopic objects really are electrically neutral (they satisfy the condition of “bulk electroneutrality”) to avoid this stiff energy penalty. Later, however, we’ll see that things are different in the nanoworld.

### 6.4 CYLINDRICAL CONDUCTORS IN VACUUM

Consider a long, straight metal cylinder (“wire”) of radius \(R_0\) carrying linear charge density \(\rho_q^{[ID]}\), with units coulombs per meter. Inside any good conductor the electric field must equal zero, so the potential must be a constant. Outside, the potential must obey the Laplace equation: \(\nabla^2 \psi = 0\). Cylindrical coordinates make this problem straightforward:

\[\psi(r, \varphi, z) = B \ln(r/R_0)\]

outside the cylinder (and uniform inside). Here \(B\) is a constant related to \(\rho_q^{[ID]}\) and to \(R_0\).

**Your Turn 6B**

Find that relation.

We have used our freedom of adding any constant to the potential to arrange that \(\psi = 0\) on the wire surface.

Next, consider two long, parallel cylinders with charge densities \(\pm \rho_q^{[ID]}\). We can superpose two solutions of the above form. The result will again solve the Laplace equation outside each cylinder. It won’t be exactly constant on the two cylinders’ surfaces, but it will be approximately so if their radii \(r_0\) are much smaller than their separation \(d\).

**Your Turn 6C**

Work out the electrostatic potential difference in this approximation between the two wires as a function of \(\rho_q^{[ID]}\), the radii, and the separation. From this, work out an approximate formula for the capacitance per unit length of this “twinlead” cable.

### 6.5 PARALLEL PLATES WITH MEDIUM

#### 6.5.1 Dielectric susceptibility describes the response of a material in linear approximation

Now imagine filling the gap between conductors with nonpolar atoms or molecules, maybe liquid argon, or more prosaically some kind of oil. What matters is that there

\[2\text{See Your Turn 5A (page 70).}\]
Figure 6.1: Dielectric medium. Polarizable “molecules” with density \( \rho_{\text{stuff}} \) fill the gap between parallel conducting plates, creating a density of dipole moment \( \vec{P} = q_1 a_1 \rho_{\text{stuff}} \hat{x} \). On the left, a layer of thickness \( a_1 \) contains uncanceled \(-q_1\) per molecule, so the total bound charge near that plate is \( a_1 \rho_{\text{stuff}} \Sigma (-q_1) \), partially canceling the free charge \(+q\) on the plate. The bound area charge density can be expressed as \( \sigma_b = \vec{n} \cdot \vec{P} \), because the outward-pointing perpendicular is \( \vec{n} = -\hat{x} \).

Similarly, on the right side there is again a partial cancellation of free and bound charges.

be no freely mobile charges, so that the material is an insulator. In this context such a material is generically called a dielectric.

Each atom/molecule has no dipole moment in isolation, but nevertheless it can deform under the influence of an external field, and so develop an induced dipole moment. Figure 6.1 suggests that the resulting uniform polarization density will lead to canceling net charge density in the interior (see the dashed red lines in the figure), but not on the two boundaries of the medium. Suppose that each molecule separates charge \( q_1 \) by distance \( a_1 \), and that they are packed with volume density \( \rho_{\text{stuff}} \). Then the uncanceled net charge forms a thin “bound” layer at the interface, with areal density

\[
\sigma_b = (a_1 \rho_{\text{stuff}} \Sigma q_1) = (\rho_{\text{stuff}})(a_1 q_1) = \hat{n} \cdot \vec{P},
\]

where \( \vec{P} \) is the volume density of induced dipole moment (polarization) and \( \hat{n} \) is the unit vector perpendicular to the surface and directed outward from the medium. We will refer to \( \sigma_b \) as the **bound area charge density**, because it can’t escape from the medium, nor even move freely within it; in contrast, the **free charge** on either plate could be moved elsewhere by connecting a wire to the plate. We’ll call the areal density of free surface charge \( \sigma_f \).

On the left side of Figure 6.1, \( \vec{P} \) and \(-\hat{n}\) point rightward, so the bound charge on the left plate is negative and indeed partially cancels the free charge we put there.

Applying the electric Gauss law to the total charge at the left plate gives

\[
\vec{E}_x = \left( \sigma_f + \sigma_b \right) / \varepsilon_0 = \left( \sigma_f + \hat{n} \cdot \vec{P} \right) / \varepsilon_0 \quad \text{where} \quad \hat{n} = -\hat{x}
\]

\[
= (\sigma_f - \vec{P}_x) \quad (6.7)
\]

---

3 Recall Section 3.7.4 (page 48).
4 Section 6.6 will look at the general situation, where the polarization density may be nonuniform.
The electric displacement is defined as

\[ \vec{D} = \varepsilon_0 \vec{E} + \vec{P} \]  \hspace{1cm} (generally).  \hspace{1cm} (6.8)

Solving Equation 6.7 gives

\[ \vec{D}_s = \sigma_f. \]  \hspace{1cm} (6.9)

The bound charges do not appear explicitly in Equation 6.9.

Most dielectric materials have zero polarization in the absence of an externally applied field. So it’s natural to suppose that \( \vec{P} \) will have a Taylor expansion, whose leading term is \( \vec{P} \propto \vec{E} \). The constant of proportionality is called the bulk polarizability of the medium.\(^6\) It is traditionally expressed as \( \varepsilon_0 \chi_e \), where the dimensionless constant \( \chi_e \) is called the dielectric susceptibility. The relation

\[ \vec{P} = \varepsilon_0 \chi_e \vec{E} \]  \hspace{1cm} (6.10)

is our first example of a linear response function. Unlike laws of Nature, it is approximate (for example, we truncated the Taylor expansion in the field strength) and nonuniversal (different materials will have different values of \( \chi_e \)).\(^7\)

Substituting Equation 6.10 in to Equations 6.7–6.9 then yields

\[ \vec{D} = \varepsilon \vec{E}, \]  \hspace{1cm} (linear, isotropic medium) \hspace{1cm} (6.11)

where the permittivity \( \varepsilon \) of the medium\(^8\) is

\[ \varepsilon = (1 + \chi_e)\varepsilon_0. \]  \hspace{1cm} (6.12)

Thus, the effect of the medium is simply to replace the vacuum permittivity \( \varepsilon_0 \) by a larger effective value \( \varepsilon \). Instead of accounting explicitly for every charge in the medium, we can simplify by forgetting it and making this one substitution. Equation 6.11 is a particular example of response function, called the constitutive relation for the material in the capacitor.

The same argument as earlier now gives capacitance as

\[ C = \varepsilon \Sigma / \omega, \]  \hspace{1cm} (6.13)

which is greater than the vacuum value.

\[ \text{Section 6.5.1' (page 92) mentions some generalizations of the phenomena discussed here.} \]
6.5.2 An energy puzzle

Can we still maintain our idea of energy as being stored in the space between the plates? At first it looks bad: Our previous formula gave $\frac{1}{2} \varepsilon_0 E^2$. We could minimize this expression by assuming enough polarization to completely neutralize the applied charge, resulting in $E = 0$ and hence zero energy storage! That doesn’t seem right.

To see what went wrong, remember that the polarization surface charge arose from deformation of molecules (or atoms) throughout the gap. The molecules will resist this deformation. They therefore store “elastic” energy; the actual polarization will be whatever minimizes the total energy (field plus deformation).

To keep things simple, this section will temporarily make the approximation\(^9\) that $\vec{E}$ is uniform (whereas actually it is disturbed by many point dipoles), and that each dipole in turn responds to this average field. Following Section 3.7.4, again imagine an individual molecule as a pair of charges $\pm q_1$, with a Hooke-law spring constant $k_1$ controlling their separation $a_1$. Thus, $a_1 = \frac{q_1 \vec{E}_x}{k_1}$. Again suppose that the polarizable objects are distributed with density $\rho_{\text{stuff}}$.

Molecular polarizability leads to bulk susceptibility.

**Your Turn 6D**

Show that in this model, $P = q_1^2 \varepsilon_0 \rho_{\text{stuff}} / k_1$, and so $\varepsilon_0 \chi = q_1^2 \rho_{\text{stuff}} / k_1$. Hence by Equation 6.12,

$$\varepsilon = \varepsilon_0 + q_1^2 \rho_{\text{stuff}} / k_1$$

for the low-density medium we are studying.

To understand this result from an energy viewpoint, let’s write down the total stored energy (electric field plus elastic deformation energy):

$$\varepsilon_{\text{tot}} / (\text{volume}) = \frac{1}{2} \varepsilon_0 E^2 + \frac{1}{2} k_1 a_1^2 \rho_{\text{stuff}} = \frac{1}{2} \left( \varepsilon_0 + \frac{q_1^2}{k_1} \rho_{\text{stuff}} \right) E^2 = \frac{1}{2} \varepsilon E^2. \quad (6.14)$$

But Equations 6.9–6.11 give $\vec{E}_x = \sigma_l / \varepsilon$, so

$$\varepsilon_{\text{tot}} / (\text{volume}) = \frac{1}{2} \frac{\sigma_l^2}{\varepsilon}. \quad (6.15)$$

We see that, for fixed free charge introduced on the plates, the system finds an equilibrium: a compromise between minimizing the two kinds of energy. The net energy is smaller than it would have been with no polarization at all (because the denominator contains $\varepsilon > \varepsilon_0$). Had the material polarized enough to eliminate the electric field altogether, as suggested at the start of this section, then the electric field energy would have been zero but the total energy would have been greater than Equation 6.15.

Another key point about Equation 6.15 is that once again, stored energy is proportional to volume. Nobody is surprised that the elastic part of the energy has this property—the

---

\(^9\)Section 6.5.3 will justify this when the dielectric is of low density, and will give an improved derivation for dense matter.
polarizable objects are spread through space at fixed density—and we already showed that the electric term also has it.

Instead of using the Hooke law, we could have left \( \alpha_1 \) arbitrary. Then Equation 6.14 has two terms that are analogous to a mechanical system: two springs connected in series. We know that that system will minimize its total energy by distributing overall deformation between the springs, rather than assigning all of it to just one of them. Similarly, our capacitor will minimize total energy by canceling some, but not all, of its imposed free charge with bound charge, again resolving the paradox at the start of this section.\(^{10}\)

The reduction of total energy when we introduce a dielectric material at fixed free charge implies a force that pulls that material into the gap. For example, a fluid dielectric will be pulled into the space between charged plates, even if it must overcome gravity to do so.\(^{11}\)

### 6.5.3 Dense media roughly follow the Clausius–Mossotti relation

The preceding section warned that it is not really justified to assume that each polarizable molecule responds to the spatially-averaged field. This may be surprising: Often, when a medium is uniform on macroscopic length scales, we may work with spatially averaged quantities, such as the local velocity in fluid mechanics. This section will make some more ad hoc assumptions, but we will at least see why this reasoning breaks down in the presence of long-range forces such as electrostatics.

We again imagine a parallel-plate capacitor with a uniform, polarizable medium between the plates. This time, however, we will single out one particular molecule for study, and set up polar coordinates centered on it. This dipole of interest responds to the net electric field created by all charges except itself. Those charges include the free charge on the distant plates, as well as bound charges in the medium. To improve, if only slightly, on our previous derivation, we now suppose that we may treat the medium as continuously and uniformly polarized, except in a spherical cavity surrounding the dipole of interest, with volume equal to \( \frac{1}{\rho_{\text{stuff}}} \). After all, surely it is foolish to insist on a continuous distribution below the molecular scale.\(^{12}\)

Figure 6.2 illustrates our idealization. An induced dipole of unknown moment \( \vec{D}_E \) sits at the center of a spherical cavity. It feels a local field \( \vec{E}_{\text{loc}} \) with three contributions: from free charge with areal density \( \pm \sigma_f \) at the plates, from bound charges \( \pm \sigma_{b,p} \) at the plates, and from bound charge \( \sigma_{b,c} \) on the surface of the cavity. The free charge density is given, but we must self-consistently find all of the bound charge densities and the dipole moment density \( \vec{P} \).

Because \( \vec{P} \) points to the right in the figure, we define \( b \) as its magnitude via \( \vec{P} = b \hat{x} \). The same reasoning as in Section 6.5 gives the bound charge density at the left plate as

---

\(^{10}\)You’ll work out details in Problem 6.8.

\(^{11}\)See Problem 6.4 and Media 2.

\(^{12}\)More sophisticated treatments consider a spherical hole that is much larger than the molecular scale, but in the end they still make assumptions, and still give only rough answers except for extremely special media such as liquid helium. For a much more sophisticated treatment see Zangwill, 2013, chap. 6.
Figure 6.2: Origin of the Clausius–Mossotti relation. A spherical surface has been drawn surrounding one polarizable molecule in a medium. We regard the interior of this surface as a “cavity” containing only a point dipole representing the molecule.

\[ \sigma_{b,p} = (-\hat{x}) \cdot \vec{p} = -b. \]

Let \( \hat{r}_s \) be the unit vector from the molecule of interest to a point on the surface of the cavity and let \( R \) be the cavity’s radius. Then the unit vector perpendicular to the surface and “outward” (away from the bulk material) is \( -\hat{r}_s \), and Equation 6.4 gives the bound surface charge at the cavity is \( \sigma_{b,c} = (-\hat{r}_s) \cdot \vec{p} = -b \cos \theta_s \), where \( \theta_s \) is polar angle measured from \( \hat{x} \). The figure illustrates why the cosine factor is needed: For example, at \( \theta_s = \pi/2 \) the molecular distortion is parallel to the surface and no net bound surface charge arises.

We wish to find the electric field at the center of the cavity, \( \vec{E}_{\text{tot}}(0) \), because that is what acts on the molecule we are studying. It receives a contribution from the charges on the plates:

\[ \vec{E}_{\text{plate}} = \frac{\sigma_f + \sigma_{b,p}}{\varepsilon_0} \hat{x} = \frac{\sigma_f - b}{\varepsilon_0} \hat{x}. \]  \hfill (6.16)

The other contribution, \( \vec{E}_{\text{cav}} \), comes from \( \sigma_{b,c} \). To find it, first use the potential formula Equation 2.7 (page 31)

\[ \psi_{\text{cav}}(\vec{r}) = \frac{1}{4\pi\varepsilon_0} \int_{\text{sphere}} d^2 \Sigma \frac{\sigma_{b,c}(\vec{r}_s)}{||\vec{r} - \vec{r}_s||}. \]

\[ \vec{E}_{\text{cav}}(\vec{r}) = -\vec{\nabla} \psi_{\text{cav}} = \frac{-1}{4\pi\varepsilon_0} \int_{\text{sphere}} R^2 d(\cos \theta_s) d\varphi_s \frac{-b \cos \theta_s}{(-2)||\vec{r} - \vec{r}_s||^3} 2(\vec{r} - \vec{r}_s) \]

\[ \vec{E}_{\text{cav}}(\vec{0}) = \frac{-b}{4\pi\varepsilon_0 R^3} \int_{\text{sphere}} d(\cos \theta_s) d\varphi_s \cos \theta_s (-\vec{r}_s). \]

Only the \( \hat{x} \) component of this vector will survive averaging over \( \varphi_s \), so

\[ \hat{x} = \frac{b 2\pi}{4\pi\varepsilon_0} \int_{-1}^{1} d(\cos \theta_s) \cos^2 \theta_s = \hat{x} b/(3\varepsilon_0). \]

The induced dipole moment equals the total field times the molecular polarizability \( \alpha \):

\[ \vec{D}_e = \alpha (\vec{E}_{\text{plate}} + \vec{E}_{\text{cav}}(\vec{0})) = \hat{x} \alpha \left( \frac{\sigma_f - b}{\varepsilon_0} + \frac{b}{3\varepsilon_0} \right). \]
6.6 Nonuniform Polarization Leads to a Bound Charge Distribution

We have now established a connection between the induced moment \( \vec{D}_l \) and the strength \( \vec{b} \) of the dipole moment density \( \vec{P} \). But the same connection applies to every molecule, so we also have

\[
\vec{P} = \rho_{\text{stuff}} \vec{D}_l.
\]

Combining the last two displayed equations and recalling that \( \vec{P} = b \hat{x} \) gives

\[
\frac{b}{\rho_{\text{stuff}}} = \frac{\alpha}{\varepsilon_0} \left( \sigma_f - \frac{2b}{3} \right).
\]

Solving for \( b \) gives

\[
\vec{P} = \hat{x} \sigma_f \left( \frac{2}{3} + \frac{\varepsilon_0}{\alpha \rho_{\text{stuff}}} \right)^{-1}.
\]

Now compare the last formula for \( \vec{P} \) to Equation 6.16 to find

\[
\vec{P} = \vec{E}_{\text{plate}} \frac{\varepsilon_0}{\varepsilon_0 - \alpha \rho_{\text{stuff}} / 3}.
\]

Writing this as \( \varepsilon_0 \chi_{\text{e}} \vec{E}_{\text{plate}} \) at last gives the dielectric susceptibility in terms of molecular polarizability:

\[
\chi_{\text{e}} = \frac{\alpha \rho_{\text{stuff}}}{\varepsilon_0 - \alpha \rho_{\text{stuff}} / 3},
\]

Clausius–Mossotti relation \( \text{(6.17)} \)

Many materials conform to approximate versions of this formula, although with other factors of order unity in place of the factor of \( 1/3 \) that came from our simplified approach.

Returning to the start of this section, consider subdividing a substance more and more finely, \( \rho_{\text{stuff}} \rightarrow \infty \) while holding \( \alpha \rho_{\text{stuff}} \) fixed. Even in this continuum limit, Equation 6.17 disagrees with the simplified version (Your Turn 6D, page 82 with \( \alpha = q_1^2/k_1 \)). However, in the limit of low density the denominator of Equation 6.17 becomes just \( \varepsilon_0 \) and we recover our earlier, approximate, relation Equation 6.10.

6.6 NONUNIFORM POLARIZATION LEADS TO A BOUND CHARGE DISTRIBUTION

We are not always so lucky as to have the polarization density \( \vec{P} \) spatially uniform. Figure 6.3 illustrates what can happen if it is not.

Section 6.5.1 argued that an interface, for example between a medium and vacuum, will develop a layer of bound surface charge with areal density \( \sigma_b \) given by

\[
\sigma_b = \hat{n} \cdot \vec{P}, \quad \text{[6.4, page 80]}
\]

where \( \hat{n} \) is the perpendicular unit vector directed away from the medium. At a sharp interior interface between two media with \( \vec{P} \) uniform on either side, we can substitute the difference in \( \vec{P} \) values into this formula, but what should we do for an arbitrary (nonuniform) \( \vec{P} \)?
Figure 6.3: Creation of interior bound charge. A collection of electrically polarizable “molecules” with nonuniform polarization (magnitude increasing as we move to the right). Net bound charge appears that is minus the divergence of the polarization density, in this case \(-\partial \hat{P}_x / \partial x < 0\).

If \(\hat{P}\) is spatially nonuniform, then the cancellation seen in the interior region of Figure 6.1 will be incomplete. Figure 6.3 shows a simple example of this effect. More generally, charge density is a scalar quantity, and according to the figure, the net bound charge density \(\rho_b\) involves spatial gradients of the polarization density. The general formula

\[ \rho_b = -\nabla \cdot \hat{P} \]  

(6.18)

is rotationally invariant, dimensionally consistent, and agrees with the figure in the special case shown there.

We can now write the Gauss law including both free and bound charges:

\[ \nabla \cdot \bar{E} = (\rho_f - \nabla \cdot \hat{P}) / \varepsilon_0. \]

Combining the divergence terms gives

\[ \nabla \cdot (\bar{E} + \hat{P} / \varepsilon_0) = \rho_f / \varepsilon_0. \]

When phrased in terms of the electric displacement (Equation 6.8), this becomes

\[ \nabla \cdot \bar{D} = \rho_f, \quad \text{electric Gauss law with medium} \]  

(6.19)

We can now specialize to the case of a linear, isotropic medium (Equations 6.11–6.12), to find \(\nabla \cdot \bar{E} = \rho_f / \varepsilon\). This formula is the same as the vacuum case, except that the free charge has been effectively reduced by a factor \(\varepsilon_0 / \varepsilon < 1\).

We can now quickly generalize earlier formulas. Consider a spherical charge distribution with total free charge \(q_f\), immersed in a dielectric medium. In the medium, the solution of Equation 6.19 is \(\bar{D}(r) = r q_f / (4\pi r^2)\), and the corresponding electric field is that function divided by \(\varepsilon\). We can find the line integral of \(dq_f \bar{E}\) from infinity to \(R\), obtaining the work that must be done to increase \(q_f\) by \(dq_f\):

\[ d\mathcal{E} = q_f dq_f / (4\pi \varepsilon R). \]
Then integrating again, from \( q_f = 0 \) to \( Q_f \), gives the total energy of a charged sphere in an infinite dielectric medium:

\[
\mathcal{E} = \frac{Q_f^2}{8\pi\varepsilon R}.
\]  
(6.20)

This expression has the same form as the vacuum result (Your Turn 6Aa), but reduced by \( \varepsilon_0/\varepsilon \).

Section 6.6' (page 92) discusses corrections to Equation 6.18.

### 6.7 CHARGE NEUTRALITY BREAKS DOWN ON THE NANOSCALE

Individual ions, and even much bigger objects (such as proteins and DNA) are often said to be “electrically charged.” The term can cause confusion. Doesn’t matter have to be neutral? The electrostatic self-energy Example on page 78 explains why people said that in first-year physics. But:

**Your Turn 6E**

a. Repeat the calculation for an object of radius \( R = 1 \mu m \) suspended in water. Recall that the static permittivity \( \varepsilon \) of water is about 80 times bigger than the value for air used in the Example.
b. Repeat for an \( R = 1 \) nm object in water.
c. Compare both answers to the thermal energy scale \( k_B T_f \), where \( T_f \) is room temperature, 298 K, and comment.

The energy of an object embedded in a medium is also called its **Born self-energy** after M. Born.\(^{13}\)

Thus, it is possible for thermal motion to separate a neutral molecule into charged fragments. For example, when you purchase DNA in bulk you actually get a salt; upon dissolving it, each DNA molecule liberates positive ions into solution and itself becomes a highly negatively charged macroion, surrounded by a layer of polarized solution.

### 6.8 POLAR FLUID MEDIA CAN BE HIGHLY POLARIZABLE

So far, we have considered molecules that have no intrinsic dipole moment ("nonpolar" molecules), but that can polarize by deforming slightly. We can also consider a medium consisting of polar molecules that initially are randomly oriented, or in any case oriented in such a way that their dipoles cancel, as in liquid water (Figure 6.4) or water vapor. In an applied electric field, the molecules can align to create net polarization.\(^{14}\)

\(^{13}\)The term “self” may be confusing. As in the preceding sections, we are discussing energy stored in the medium and field exterior to the charged body; the qualifier “self” only distinguishes this from a possible “interaction energy” with other charged objects.

\(^{14}\)Section 3.7.3 (page 47) already mentioned that although individual molecules in liquid water are randomly oriented on average, each has a definite instantaneous orientation and can respond to the others’ electric field.
Figure 6.4: Tetrahedral arrangement of water molecules in an ice crystal. In liquid water, the immediate neighborhood of any one molecule has a similar character, though with some randomness added by thermal motion. The sum of all electric dipole moments in a unit cell is 0, but in the liquid state an applied electric field can bias the distribution of orientations, generating a nonzero net dipole moment density. The gray outline of a tetrahedron is just to guide the eye. Dashed lines are hydrogen bonds that stabilize the structure.

So once again we face a puzzle: Won’t this system always cancel an applied \( E \), at least up until the molecules have reached perfect alignment? Section 6.5.2 escaped this paradox by acknowledging an elastic energy cost for polarizing individual molecules, but in liquid water they rotate freely. Nevertheless, they do pay a price: Aligning the molecules costs entropy, or equivalently raises the free energy of the system. In a weak field, the compromise between free energy cost and electrostatic energy reduction will be mathematically similar to what we previously worked out, again leading to an incomplete cancellation of the electric field.\(^\text{15}\) Although the net static permittivity is therefore not infinite, for water at room temperature it is quite high; indeed, \( \varepsilon \approx 80\varepsilon_0 \). Interestingly, solid water (ice) has a much smaller permittivity, because its molecules are not free to reorient. Like any other molecules, they may deform, but the effective spring constant for deformation is much stiffer than the one for alignment.

The reorientation of molecules in liquid water is accompanied by frictional loss as they rub against their neighbors. When \( E \) oscillates at microwave frequency, the associated heating can be considerable, and indeed, you know that a microwave oven heats liquid water, with its strong and mobile dipoles, faster than it does glass, plastic, or even ice.\(^\text{16}\)

At higher frequencies, however, the reorientation response is too slow to follow the field fluctuations. That is, the permittivity of water is strongly frequency-dependent and much smaller (closer to that of ice) in the optical range than at lower frequencies.\(^\text{15}\) Section 6.8\(^{\text{1}}\) (page 92) describes another consequence of the finite response time of polarization.

6.9 PARTITIONING OF IONS

6.9.1 Solubility of ionic solids follows a simple quantitative rule

It is hard to vaporize rock salt. You have never achieved this on your kitchen stove. Dissociating the individual ions requires enough thermal energy to overcome their enormous

\(^{15}\)You’ll work out a quantitative approach in Problem 6.11.

\(^{16}\)Food contains salty water, which is a conductor. The electric fields in the applied microwaves therefore also induce currents, which give rise to additional resistive (Joule) heating.
6.9 Partitioning of Ions


equation

\[ y_0 = 0.00100 \]

\[ y_1 = 0.01200 \]

\[ y_2 = 0.02300 \]

\[ y_3 = 0.03400 \]

\[ y_4 = 0.04500 \]

\[ y_5 = 0.05600 \]

\[ y_6 = 0.06700 \]

\[ y_7 = 0.07800 \]

\[ \frac{1}{\varepsilon} = \text{relative permittivity} \]

\[ 10^{0} \]

\[ 10^{-2} \]

\[ 10^{-4} \]

\[ 10^{-6} \]

\[ 0.00 \]

\[ 0.01 \]

\[ 0.02 \]

\[ 0.03 \]

\[ 0.04 \]

\[ 0.05 \]

\[ 0.06 \]

\[ 0.07 \]

Figure 6.5: [Experimental data.] Semilog plot of the solubility of sodium chloride in solvents of different static dielectric constants. Although these chemicals have many specific features, the solubility over a huge range is largely explained in terms of a single characteristic, the solvent permittivity \( \varepsilon \). The line goes to maximum solubility at large \( \varepsilon \). [Data from Israelachvili, 2011.]

Electrostatic attraction (Section 3.7.3, page 47). And yet, every day you dissociate salt by adding it to water. What accounts for these radically different behaviors?

Separating ions is easier with a highly polar solvent than in vacuum because of the \( 1/\varepsilon \) factor in the Born self-energy (Your Turn 6E) and the large permittivity of water. Figure 6.5 indeed shows a strong inverse correlation between the permittivity of the solvent and a salt’s willingness to dissolve (solubility). Quantitatively, the Boltzmann distribution leads us to expect that the fraction of Na\(^+\)Cl\(^-\) pairs that are separated should be a constant times \( e^{-\varepsilon/\varepsilon} \). You’ll follow up this observation in Problem 6.12.

6.9.2 Partitioning at a fluid interface or cell membrane; permeability

Next, imagine an oil-water interface. An ion, for example Na\(^+\), is dissolved in the water (\( \varepsilon \approx 80\varepsilon_0 \)). Suppose that this ion crosses the interface to the oil side (\( \varepsilon \approx 2\varepsilon_0 \)). The lower permittivity of oil means that the self-energy increases. Thus, even though there is no material barrier at the interface, ions will segregate to the water side, following the Boltzmann probability rule.

Living cells are surrounded by a bilayer membrane a few nanometers thick, a fluid layer with nonpolar hydrocarbon chains in its center. The water on either side of this membrane contains lots of ions, but they will not cross the membrane because of the high Born self-energy they would incur in the intermediate states while crossing.\(^{17}\)

Hence, cell membranes are electrically insulating, despite being so thin. Their thinness

\[^{17}\]See Problem 6.12. [Screening by salt reduces the self-energy still further on the water side (Section 10.3.4, page 156).]
also implies that such membranes also have very high capacitance per unit area (Equation 6.13). The passage of ionic current into or out of a cell can take place only via ion channels, protein complexes embedded in the membrane containing water-filled passages. Chapters 11–12 will show how the interplay of high capacitance, ionic imbalance, and controlled passage leads to the phenomenon of nerve impulses.

Note that we are not saying that membranes should be thought of as sealed bags. Indeed, they are far more permeable to water than they are to ions. Unlike ions, water molecules are neutral, and hence have lower Born self-energy in a membrane than do ions.

6.10 BOUNDARY CONDITIONS

The same discussion we gave at an interface between a conductor and vacuum continues to hold at interfaces between a conductor and a dielectric, a dielectric and vacuum, or between two different dielectrics (Section 2.4, page 31): $\vec{E}_\perp$ can jump at such an interface, because free charges (on a conductor) or bound charges (in one or both dielectrics) can be localized at the surface.

For example, suppose that a dielectric material $I$ faces vacuum or air, and let $\hat{n}$ be the perpendicular to a point on the surface that points away from the material. Then Equation 6.4 and the electric Gauss law (Equation 0.1, page 2) give that (Figure 6.6a)

$$\hat{n} \cdot (\vec{E}^{[\text{vac}]} - \vec{E}^{[I]}) = (\hat{n} \cdot \vec{P}^{[I]} + \sigma_f)/\varepsilon_0.$$ \hspace{1cm} (6.21)

If we know the polarization in terms of the electric field, for example via $\vec{P}^{[I]} = \varepsilon_0 \chi \varepsilon \vec{E}^{[I]}$, then we get a condition for how $\vec{E}_\perp$ jumps. Rephrasing in terms of the displacement (Equation 6.8, page 81) gives a generalization of Equation 6.9:

$$\Delta \vec{D}_\perp = \sigma_f. \hspace{0.5cm} \text{dielectric boundary} \hspace{1cm} (6.22)$$

Often it is reasonable to assume that there is no free surface charge; for example, the dielectric could have been neutral before an external field was applied.

Turning now to the components of $\vec{E}$ that are tangential to the surface, integrating both sides of $\vec{V} \times \vec{E} = \vec{0}$ over a small area that passes through the interface shows that $\vec{E}_\parallel$ may not jump as we cross the boundary (Figure 6.6b):

$$\Delta \vec{E}_\parallel = 0. \hspace{0.5cm} \text{dielectric boundary} \hspace{1cm} (6.23)$$

In particular, these two components must equal zero just inside a dielectric that adjoins a conductor.

---

18Chapter 9 will discuss how this prediction was confirmed experimentally.
19A water molecule does have a dipole field, whose self-energy is larger in the membrane’s nonpolar interior than in the surrounding bulk. But this energetic cost of entry is not as great as that due to the Born energy of a small ion.
Figure 6.6: [Sketches.] **Boundary conditions near a dielectric.** (a) The short *red cylinder* has one end cap just outside and the other just inside the dielectric. Integrating the electric Gauss law over it, and using the divergence theorem, shows that the component of $\vec{E}$ perpendicular to the surface can have different values just inside and outside the conductor, due to bound and free charges at the surface. (b) The *red rectangle* has one of its longer edges just outside and the other just inside the dielectric. Integrating the curl-free condition, and using Stokes’s theorem, shows that any component of $\vec{E}$ parallel to the surface must have the same values just inside and just outside the conductor.

**FURTHER READING**

*Semipopular:*
Adee, 2023.

*Intermediate:*
See also Pollack & Stump, 2002, chap. 4 and 6.
Bioelectricity, Coulter counter: Grodzinsky, 2011.

- Piezoelectricity: en.wikipedia.org/wiki/Piezoelectricity.

*Technical:*
Piezoelectric dressings to promote wound healing: Long et al., 2018.
6.5.1 Fermoelectricity, electrostriction and piezoelectricity

The main text introduced the common assumption that polarization density is zero at zero applied field. But nothing forbids a permanent electric polarization, analogous to the phenomenon of permanent magnetism, and indeed materials with this property, called ferroelectrics, are known. (Devices relying on ferroelectricity are sometimes called electrets.)

Regardless of whether ferroelectricity is present, the main text imagined a field-induced polarization resulting from deformation of individual molecular constituents. In fact, such deformations are real and can lead to a small but measurable bulk change in the overall size of a dielectric body, a phenomenon called electrostriction.

Conversely, some materials polarize when mechanically strained; they are called piezoelectric. The bound charge generated by this effect can be measurable and even useful (for example, in a microphone transducing mechanical pressure to electrical signals, or in a grill lighter that generates a spark when a crystal is struck).

6.6 Higher multipoles

Equation 6.18 described the charge density that arises from a nonuniform density of induced dipole moment. But individual molecules also have higher multipole moments, which could potentially change upon the imposition of an external field. A more complete analysis shows, however, that the density of induced quadrupole moment enters as a correction to Equation 6.18 via its second derivative, and so on for higher multipoles (Jackson, 1999a, §6.6). Dimensional analysis implies that induced higher moments will involve higher powers of the molecular size. That length scale, in turn, is divided by the inverse of a length scale characterizing spatial variation, which arises from the extra derivative(s).

Often we are interested in applied fields arising from an electromagnetic wave, and in particular visible light, whose wavelength is much larger than a molecule. Higher multipole contributions therefore will be suppressed by powers of the small ratio of scales.

6.8 Electrorotation

Chapter 3 discussed the torque on an electric dipole in an external field. The present chapter discussed how a polarizable object can acquire a dipole moment from an external field. Section 6.8 pointed out that this induction may not be instantaneous. Weaving these threads together, we see that polarization can lag behind a rapidly varying electric field; in particular, a rotating electric field can exert constant torque on a polarizable object. M. Washizu and coworkers applied this insight to living bacteria, polarizable micrometer-scale objects (Washizu et al., 1993). When an electric field rotating at megahertz frequency was applied to the cells, they experienced torques without suffering any damage.

The bacteria in question have rotary motors attached to flagella; anchoring one flagellum to a surface and applying electrorotation allowed H. Berg and coauthors to study details of the motor’s operation (Berg & Turner, 1993), including the unexpected discovery that it remodels itself (adds or removes some “pistons”) in response to changes in external load (Wadhwa et al., 2022).
6.1 Capacitor fun
A simple capacitor is a device formed by two insulated conductors separated by vacuum. If equal and opposite charges are placed on the conductors, there will be an electrostatic potential difference between them. The ratio of the magnitude of charge on one of them to the magnitude of is their capacitance (Equation 6.1). Using the electric Gauss law, calculate the capacitances of:

a. Two concentric conducting spheres with radii and .

b. Two concentric conducting cylinders of length that is large compared to their radii and . What is the diameter of the outer conductor in a vacuum-filled coaxial cable whose central conductor is a cylindrical wire of diameter 1 mm and whose capacitance per unit length is 0.5 \( \mu \text{F/cm} \)?

6.2 Twinlead cable
Two long, cylindrical conductors of radii and , are parallel, with centerlines separated by distance , which is much bigger than either and . Long ago, such “twinlead” cables were used as waveguides to bring television signals from an antenna to a tuner circuit.

a. Let \( c = \sqrt{ab} \) and show that the capacitance per length is approximately proportional to \( (\ln(w/c))^{-1} \). Find the constant of proportionality.

b. Now suppose that and . What diameter wire would be necessary to obtain 0.1 \( \text{pF/cm} \) if the separation is \( w = 5 \text{ mm} \)? (pF means \( 10^{-12} \text{ F} \).)

6.3 Can you take the pressure?
Let us understand why there is such a strong tendency for matter to be electrically neutral. Consider a spherical balloon filled with gas. At atmospheric pressure and room temperature, a balloon of radius \( R = 18 \text{ cm} \) contains about one mole of gas.

a. Now assume that we remove one electron from one out of every million gas atoms, while holding \( R \) fixed. The remaining uncompensated charges will repel each other, so they will distribute themselves on the surface of the sphere. Find the electrostatic self-energy of this assembly of charge.

b. Differentiate your result in (a) to find the pressure (change of energy per change of volume) exerted on the balloon. Express your answer as a multiple of atmospheric pressure, which is about \( 10^5 \text{ N/m}^2 \).

6.4 Fluid-filled capacitor
A parallel-plate capacitor, with vertical plates of width \( W \), height \( L \), and small, fixed separation \( a \), is partly immersed in dielectric fluid (for example, oil). The fluid has permittivity \( \epsilon \) and mass density \( \rho_m \) that we can look up.

Above the fluid there is vacuum. When a fixed potential difference \( \psi_0 \) is established between the plates, the fluid between the plates is observed to rise to a higher level \( \Delta h \)
above the surrounding fluid. We’d like to know why the fluid rises against the force of gravity, and what determines the equilibrium $\Delta h_\ast$.

a. Before you write any equations, explain physically why the fluid is pulled into the capacitor. There’s a difficult approach, which involves fringe fields. But there’s an easy approach, which involves energy considerations. A fixed-potential source, for example a battery, can be idealized as a black box whose internal energy (for example, chemical energy) rises or falls as electric charge passes through it (charging or discharging the battery), in such a way that the potential $\psi_0$ is always fixed (for example, to 1.5 V on a commercial battery). When it is connected to a capacitor, charge will initially flow, then stop. Neglecting gravity, consider the total system’s final energy if $\Delta h$ is fixed to zero (empty), versus if $\Delta h = L$ (full of oil). Now discuss how the situation changes when gravity is included, and so explain the phenomenon qualitatively.

b. Still not writing any equations, apply dimensional analysis. The equilibrium $\Delta h_\ast$ will depend on the given parameters of the system: the fixed potential $\psi_0$, geometry $L, W, a$, and fluid characteristics, as well as on relevant constants of Nature ($\varepsilon_0$ and the acceleration of gravity at Earth’s surface). Moreover, argue that $\Delta h_\ast$ will not depend on $W$ or on $L$. Then see how much you can predict about $\Delta h_\ast$ just from dimensions.

Now it is time to write some equations and make your answers to (a,b) more detailed. For any value of $\Delta h$, we may regard the system as two capacitors in parallel: One has area $W\Delta h$ and is filled with oil; the other has area $W(L - \Delta h)$ and has no dielectric.

c. Write an expression for the total energy of the system as a function of $q_{\text{oil}}, q_{\text{vac}}, \Delta h$, and fixed parameters. As described above, model the voltage supply as a subsystem whose internal energy is a constant minus $(\psi_0)q_{\text{tot}}$, where $q_{\text{tot}} = q_{\text{oil}} + q_{\text{vac}}$ is the net charge that it places on the plates.

d. Minimize the total energy, obtaining among other things a formula for the equilibrium $\Delta h_\ast$ in terms of fixed parameters.

e. Substitute some rough numbers appropriate to a classroom demonstration and see what value your formula predicts for $\Delta h_\ast$: $\varepsilon = 2.5\varepsilon_0$ for oil, $a = 5\text{ mm}$, $\rho_{\text{oil}} = 10^3 \text{ kg/m}^3$, $\psi_0 = 10^3 \text{ V}$, $W = L = 6\text{ cm}$.

6.5 Biocapacitor

a. Show that the electric field outside a line of charge in vacuum is $E = \hat{r} \frac{\rho_{\text{q}}^{(ID)}}{2\pi \varepsilon_0 r}$. Here $r$ is the distance from the observation point to the line and $\hat{r}$ is the unit vector pointing from the line, perpendicular to it, and passing through the observation point. $\rho_{\text{q}}^{(ID)}$ is the linear charge density (charge per unit length) on the line, which we assume to be uniform.

b. Suppose that instead, the charge is distributed on a cylinder of radius $R_1$, and that an equal and opposite charge is distributed on a larger cylinder, with radius $R_2$. The two cylinders are concentric (they have the same centerline). Use (a) to state the capacitance per unit length of this coaxial “cable.”
c. The neurons in your body each have a long, thin “output line” called the axon. It’s got a good conductor inside (axoplasm) and outside (mostly salt water), separated by a thin insulating layer (the cell’s outer bilayer membrane). The insulating layer has permittivity \( \epsilon \), which may be different from \( \epsilon_0 \), but with this modification we ought to be able to apply your result in (b) to find the capacitance of the membrane. And yet, people always use a formula that looks quite different from yours: the parallel-plate capacitor formula \( C = \epsilon \Sigma / w \). Here \( \Sigma \) is the total area of the membrane and \( w \) is its thickness. Resolve this apparent discrepancy. [Hint: An axon may typically have diameter 1 \( \mu \text{m} \). Its membrane may typically have thickness 3 nm.]

d. Suppose further that the membrane is made of an insulating material (lipid) with \( \epsilon \approx 3\epsilon_0 \). Get a numerical estimate for the capacitance per unit area, \( C \Sigma \), expressed in \( \mu \text{F}/\text{cm}^2 \).

6.6 Microwave heating
[[Not ready]]

6.7 Measure \( \epsilon_0 \)
If you know about fringe fields, neglect them in this problem (that is, pretend the plates are infinite).

a. A flat, circular plate, of radius \( r = 14 \text{ cm} \), in vacuum, carries total charge \( q \). Write an approximate expression for the electric field strength \( \vec{E} \) very near the plate as a function of distance \( w \) to the plate (so \( w \ll r \)).

b. A second such plate is held close to, and parallel to, the first one, and carries total charge \( -q \). Find the force \( df \) on each surface area element \( dA \) of the second plate due to the charge on the first plate.

c. Find the electric field strength and the electrostatic potential drop \( \Delta \psi \) between the plates as a function of their separation \( w \).

d. A mechanical force \( f \) is required to maintain the second plate at a fixed distance \( w \). Find this force as a function of \( r, w, \Delta \psi \) and physical constants.

e. One can readily measure the \( \Delta \psi \) needed to balance a force of \( 10^{-2} \text{ N} \) at separation \( w = 0.5 \text{ cm} \). One trial yielded \( \Delta \psi \approx 1055 \text{ V} \). Use this information to determine the approximate numerical value of \( \epsilon_0 \) and compare to the standard value. [Note: The plates were in air, so really you are finding the dielectric susceptibility \( \epsilon_{\text{air}} \). But air is similar to vacuum.]

f. Compute \( 1/\sqrt{\epsilon_0 \mu_0} \) based on your answer to (d) and the standard value of \( \mu_0 \). Does its value remind you of a well-known physical constant?

6.8 Mechanical analogy

a. Remind yourself of how two Hooke-law springs in series are equivalent to a single spring, and the formula for that equivalent spring constant. Rederive that formula by minimizing total elastic energy at fixed total extension.

Section 6.5.2 (page 82) introduced an analogy in which one “spring” is the total deforma-
tion energy of all the dipoles, and the other is the electric field energy. Thus:

\[
\text{extension of first spring} \leftrightarrow \text{bound charge on the dielectric surface} \\
\text{extension of second spring} \leftrightarrow \text{total (free plus bound) charge at each plate.}
\]

All vectors point along the x direction, so we will abbreviate by letting \( \vec{E} \) denote \( \vec{E}_1 \) and so on. The Gauss law implies that \( E = \sigma_{q,\text{tot}}/\varepsilon_0 = (\sigma_f + \sigma_b)/\varepsilon_0 \). We have a model in which each “molecule” contains charges \( \pm q_1 \) that can separate by an unknown distance \( a_1 \), with a spring constant \( k_1 \) resisting that deformation.

The polarization density is \( P = q_1 a_1 \rho_{\text{stuff}} \), so the areal density of bound charge is \( \pm q_1 a_1 \rho_{\text{stuff}} \), with sign chosen to partially cancel the free charge \( \sigma_f \). Thus, total energy per volume is

\[
\mathcal{E}_{\text{tot}}/(\text{volume}) = \frac{1}{2\varepsilon_0} (\sigma_f - q_1 a_1 \rho_{\text{stuff}})^2 + \frac{1}{2} \rho_{\text{stuff}} k_1 a_1^2.
\]

b. Rearrange this expression into the form

\[
\mathcal{E}_{\text{tot}} = \frac{1}{2} (K a_1^2 + K'(L - a_1)^2),
\]

for appropriate constants \( K \) and \( K' \), and a variable \( L \) related to free charge density.

c. Apply your result in (a) to find the minimal energy for fixed \( L \). Rewrite that result in terms of the original variables to get energy per volume as a quadratic function of \( \sigma_f \).

d. Rewrite some more to get the form \( \mathcal{E}_{\text{tot}}/(\text{volume}) = \frac{1}{2\varepsilon} \sigma_f^2 \), where \( \varepsilon \) is a combination of the original constants that you are to find.

e. Describe the limits of (i) unpolarizable medium, and of (ii) infinitely polarizable medium in the spring language, and comment on the reasonableness of your result in (d).

### 6.9 Permittivities of various polar liquids

Here are the permittivities of some common substances, all of which are liquids with similar mass density at room temperature:

- Methanol, \( \text{CH}_3-\text{OH} \), has \( \varepsilon/\varepsilon_0 = 33 \);
- Ethanol, \( \text{CH}_3-\text{CH}_2-\text{OH} \), has \( \varepsilon/\varepsilon_0 = 24 \);
- 1-propanol, \( \text{CH}_3-(\text{CH}_2)_2-\text{OH} \), has \( \varepsilon/\varepsilon_0 = 20 \).

Each of these also has about the same dipole moment per molecule, \( \approx 1.7 \text{ debye} \), where 1 debye \( \approx (0.021 \text{ nm})e \). Here \( e \) is the proton charge.

a. Explain qualitatively the differences in permittivities.

b. Attempt a quantitative comparison to your expectations.

### 6.10 States of matter

Two parallel, circular, conducting plates, each with diameter 20 cm, were arranged with variable distance between the plates and the material in the gap. Here are the resulting capacitances for several trials:
<table>
<thead>
<tr>
<th>Material</th>
<th>Separation, cm</th>
<th>Measured capacitance, nF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Air</td>
<td>6</td>
<td>0.040</td>
</tr>
<tr>
<td>Air</td>
<td>1</td>
<td>0.054</td>
</tr>
<tr>
<td>Air</td>
<td>0.1</td>
<td>0.34</td>
</tr>
<tr>
<td>Paper</td>
<td>3.3</td>
<td>0.058</td>
</tr>
<tr>
<td>Liquid water</td>
<td>3.8</td>
<td>2.45</td>
</tr>
<tr>
<td>Ice</td>
<td>3</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Make some appropriate quantitative and qualitative comments on these data.

6.11 *How to measure molecular dipole moments*

Section 6.8 (page 87) suggested that the polarizability of a fluid or gas of polar molecules could in part be interpreted in terms of the molecules’ dipole moments aligning in an external field, and that the degree of alignment would also depend on temperature. Before examining experimental data, let’s think about what we may expect.

a. A rigid electric dipole, in a uniform external field that points parallel to \( \hat{z} \), has a spatial orientation that fluctuates due to thermal motion. Use the Boltzmann distribution to write down an integral that expresses its mean, \( \langle \hat{D}_{E,z} \rangle \), in terms of the field strength \( E \), the fixed magnitude \( \hat{E} \), and temperature.

b. Do the integral.

c. Simplify your result by considering the limiting case of weak applied field, so that \( \langle \hat{D}_{E,z} \rangle \rightarrow 0 \), and get just the first term of its Taylor series expansion in \( E \). (State the condition needed for \( E \) to be “weak” in this sense.)

d. For a low-density gas, each molecule responds to the imposed field (the effect of other molecules is small). Thus, \( \epsilon \) differs only slightly from \( \epsilon_0 \) and the relation between dielectric constant and molecular dipole moment is simple: Combining Equations 6.10 and 6.12 gives

\[
\left( \frac{\epsilon}{\epsilon_0} - 1 \right) = \frac{\hat{P}}{(\epsilon_0 \hat{E}_z^2)}.
\]

This formula relates the experimentally measurable left side to the theoretically predicted right side. Make that relation more explicit, and hence predict the dependence of permittivity on temperature, molecular number density, and the dipole moment \( D_E \) of a single molecule.

Experimentally, temperature can be controlled; density and permittivity can be measured. So the relation you found gives us a way to infer \( D_E \) from data. C. Zahn did many such experiments. In Figure 6.7, the dimensionless quantity \( \nu \) denotes “specific volume relative to an ideal gas at standard temperature and pressure,” that is,

\[
\nu = \frac{\text{volume}}{\text{molecule}} / \left( \frac{k_B T_{\text{STP}}}{p_{\text{STP}}} \right).
\]

Here \( p_{\text{STP}} = 1.01 \cdot 10^5 \text{ Nm}^{-2} \) is standard pressure and \( T_{\text{STP}} = 273 \text{ K} \). In short, \( \nu \) is a constant divided by the number density of gas molecules.

e. Find that constant.

f. The vertical axis on the graph shows the dimensionless quantity \( Y = (\epsilon/\epsilon_0 - 1)(T/(1 \text{ K})\nu \) as a function of temperature. Interestingly, for each gas shown the re-
Figure 6.7: [Experimental data.] Specific polarizabilities of various gases as functions of temperature. The vertical axis gives permittivity relative to vacuum, multiplied by the temperature and a density factor $\nu$ described in Problem 6.11. [Data from Zahn, 1926.]

Polarizability is roughly linear. Work out the expected relation between $Y$ and temperature from your earlier result. Can you understand qualitatively the data for ammonia ($\text{NH}_3$) based on that relation?

g. Rather than reading the graph, use these tabulated values from Zahn's paper to deduce the dipole moment of an ammonia molecule:

<table>
<thead>
<tr>
<th></th>
<th>$T [K]$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>first</td>
<td>456.9</td>
<td>2.086</td>
</tr>
<tr>
<td>second</td>
<td>241.7</td>
<td>1.966</td>
</tr>
</tbody>
</table>

h. Re-express your answer in the customary unit debye $\approx (0.021 \text{ nm})e$, where $e$ is the proton charge.

i. Why aren’t the curves in the figure horizontal? Why indeed do some of them actually seem to pass through the origin?

[Note: It may seem inappropriate to treat individual molecules with classical physics. J. van Vleck carefully repeated the analysis quantum-mechanically and found that fortuitously, the answer is the same.]

6.12 Solubility
Examine Figure 6.5 (page 89), which shows the solubility of table salt in various liquids.

a. Describe the trend you see. [Hint: There are a lot of scary chemical words on this plot. Ignore them! Just think about what the curve is saying about the relation between two physical quantities.]

b. Qualitatively explain this trend using ideas discussed in the chapter.

c. Without doing any calculation: What in principle could we learn from the measured slope of the line?
CHAPTER 7

Vista: Electrohydrostatics

It is the discovering of the connection between physical phenomena and describing them by mathematical analysis, rather than the analysis itself, which is interesting.

— G. I. Taylor

7.1 FRAMING: AN IMPOSSIBLE SHAPE

Think about soap bubbles you have observed. The usual closed ones come to a hydrostatic equilibrium, where they stop wobbling and assume a spherical shape (Figure 7.1a). Authority figures have probably told you, “A sphere has the smallest surface area for a given volume, so surface tension dictates that shape.” Indeed, when we see videos of astronauts creating zero-gravity blobs of soup and then slurping them up, the equilibrium shapes are spherical, due to the air–liquid interfacial tension. Even with gravity and wind resistance, raindrops are also roughly spherical.

Think some more. A wire frame dipped in soap solution can lead to other kinds of equilibrium surface shapes. Dip a frame shaped like the edge of a potato chip, and you get a saddle-shaped film (Figure 7.1b). Dip two circular rings and if you’re careful, you can get a catenary-type surface spanning them (Figure 7.1c). (With even greater care, you may be able to get a cylinder with closed caps.)

But many other shapes never arise: You never get a cone, nor indeed any sort of isolated, sharp point (Figure 7.2e)—neither for open or closed soap films, nor for water

Figure 7.1: Some air–fluid–air interfaces in mechanical equilibrium. (a) Closed (distinct inside and outside regions). (b) Open. (c) Open. [(b) Photo by J. Jacobsen. (c) Photo by R. E. Goldstein, A. Pesci, and K. Moffatt.]
And yet, Figure 7.3 shows a conical surface of a fluid–fluid interface, in equilibrium, displaying a sharp point. We’d like to answer questions like:

- How can this seemingly impossible shape arise?
- Are there restrictions on the sort of conical shapes we can realize?
- Is there technological relevance? (Answer: Yes, lots.)

Another goal of this chapter is to foreshadow some ideas about tensors for future elaboration.

**Electromagnetic phenomenon:** The interface between a conducting and an insulating fluid can form a sharp conical point, despite surface tension.

**Physical idea:** Diverging electric fields at the cone’s apex modify the usual Young–Laplace law, allowing a matched divergence in equilibrium curvature.

## 7.2 SOME GEOMETRY OF CURVES AND SURFACES

The next sections introduce many symbols, which are summarized here for reference:

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s$</td>
<td>arc length coordinate along a curve $\vec{c}$</td>
</tr>
<tr>
<td>$\vec{c}(s)$</td>
<td>curve presented in parametric form</td>
</tr>
<tr>
<td>$\Delta(s)$</td>
<td>deviation from tangent line</td>
</tr>
<tr>
<td>$\pi$</td>
<td>curvature of a curve in a plane</td>
</tr>
<tr>
<td>$\xi$</td>
<td>small perpendicular displacement of a curve or surface</td>
</tr>
<tr>
<td>$T$</td>
<td>interfacial tension of a fluid–fluid interface or free film</td>
</tr>
<tr>
<td>$\Delta T = T_{in} - T_{out}$</td>
<td>jump in surface tension across a 1D barrier</td>
</tr>
<tr>
<td>$F$</td>
<td>line tension</td>
</tr>
<tr>
<td>$L$</td>
<td>total arc length of a curve in a plane</td>
</tr>
<tr>
<td>$u, v$</td>
<td>local coordinates centered on a point $P$</td>
</tr>
<tr>
<td>$F(u, v)$</td>
<td>surface presented in parametric form</td>
</tr>
<tr>
<td>$B_{ij}$</td>
<td>description of surface shape near a point; $k_i$, its eigenvalues when expressed in normal coordinates</td>
</tr>
<tr>
<td>$H = (k_1 + k_2)/2$</td>
<td>mean curvature of a surface in 3-space</td>
</tr>
<tr>
<td>$G = k_1 k_2$</td>
<td>Gauss curvature of a surface in 3-space</td>
</tr>
<tr>
<td>$\Delta p = p_{in} - p_{out}$</td>
<td>jump in fluid pressure across a 2D interface</td>
</tr>
<tr>
<td>$N(r), M(\theta)$</td>
<td>functions used in separation of variables</td>
</tr>
<tr>
<td>$\theta_0$</td>
<td>polar angle for a cone with opening angle $2(\pi - \theta_0)$</td>
</tr>
</tbody>
</table>

### 7.2.1 Curves in a plane can be characterized by a single curvature function

Before discussing surfaces in space with interfacial tension, let’s warm up by studying a curve in a plane, possibly with line tension, for example, a stretched rubber band. Consider
7.2 Some Geometry of Curves and Surfaces

Figure 7.2: Some illustrative 2-surfaces. In each panel, $G$ and $H$ refer to Gauss and mean curvatures, respectively (Section 7.2.3, page 103).

(a–b) A closed soap bubble can reach hydrostatic equilibrium as a surface of constant curvature, possibly constrained by a wire frame: (a), a free-standing sphere; (b), a cylinder with bulging caps. 
(c–d) An open soap film can reach equilibrium as a surface with mean curvature everywhere zero: (c), flat plane; (d), saddle. 
(e) A sharp conical point “should” never arise as an equilibrium shape—but see Figure 7.3.

Figure 7.3: [Photo.] Conical point of an oil–water interface (side view). The surface with polar angle $\theta_0$ is a cone with half-angle $\pi - \theta_0$. [Adapted from Taylor, 1964.]

the curve shown in Figure 7.4a. At the point $P$, construct the tangent line as shown. As we walk away from that point, the perpendicular distance $\Delta(s)$ from the curve to its tangent as a function of arc length $s$ becomes nonzero (unless the line is straight at $P$). The Taylor series of $\Delta(s)$ has no linear term (that’s what it means to be tangent). The quadratic term describes whether the curve is straight at $P$. Writing that term as $\frac{1}{2}\kappa(P)s^2$, the coefficient $\kappa(P)$ has dimensions $L^{-1}$ and is called the curvature at $P$.

The curvature as just defined also controls how a new curve, obtained by displacing the original by an amount $\xi(s)$ along the perpendicular, will have slightly different arc length from the original; see Figure 7.4b. Intuitively, a straight line is the shortest curve joining two given points, because if there’s a bend, “you could instead take a shortcut.”

\footnote{\textsuperscript{2}In flat euclidean space.}
Chapter 7  Vista: Electrohydrostatics

Figure 7.4: Measure of curvature for a curve in a plane. (a) \( \Delta(s) \) is distance from a curve to its tangent line at \( P \), after we travel arc length \( s \) away from \( P \) on the curve. The quadratic part of \( \Delta(s) \) is a measure of curvature. If we arbitrarily designate the lower region as “outside,” then \( \Delta \geq 0 \) when measured along the outward-pointing direction, and the curvature \( \kappa \) is positive at \( P \). (At other places along the curve, it may become zero or negative.)

(b) The curve has been shortened (dashed line) by displacing it a perpendicular distance \( \xi \). With the same choice of perpendicular as (a), \( \xi < 0 \) at \( P \). Because \( \kappa \xi \) is negative there, Idea 7.1a correctly predicts that the deformed curve will be shorter than the original. Idea 7.1b also correctly predicts how the area of the “outside” region grows at the expense of the “inside.”

To make that more precise, you’ll show in Problem 7.1 that:

\( a. \) To first order in \( \xi \), the curve’s total length change is the integral of arc length times \( \kappa \xi \) (a local formula).

\( b. \) In contrast, the area in the plane occupied by one side of the curve grows, and the other side shrinks, by the line integral of arc length times \( \xi \) (another local formula, but without any factor of curvature).

7.2.2 Mechanical equilibrium of an interface in a plane

Now imagine a floating skimmer designed to contain an oil slick. If you pin it between two fixed points and put it under line tension \( F \), and there’s no oil slick, then it will minimize length by assuming a curve of constant, zero curvature (a straight line). If one side confines an oil slick, however, then the skimmer will bulge out, due to the higher air–water interfacial tension on the oil-free side.\(^3\) It now assumes a shape that is a circular arc, that is, constant, but nonzero, curvature. Let’s see why.

To understand the situation, think in terms of energy. In mechanical equilibrium, the line tension \( F \) that we apply to the skimmer is constant along its length. The interfacial tension difference \( \Delta T = T_{\text{in}} - T_{\text{out}} \) is also constant, set by properties of water and oil. Mechanical equilibrium also requires that the curve’s shape must minimize total energy. We described a small shape disturbance by a function \( \xi(s) \), and Idea 7.1 says that the corresponding first-order change in energy has two parts: The interfacial tension difference \( \Delta T \) multiplies \( f \, ds \, \xi \), whereas the line tension \( F \) multiplies\(^4\) \( f \, ds \, \kappa \xi \). In mechanical equilibrium, the net first-order variation of energy must be zero:

\[
0 = \int ds (\Delta T + F \kappa) \xi.
\]

---

\(^3\)Try floating a closed loop of fine thread on water and adding a drop of detergent to the enclosed part of the water surface.

\(^4\)The sign reflects a particular choice of which direction of deviation from the tangent will be called positive (Figure 7.4a,b). Strictly speaking, interfacial tension involves the free energy cost.
This relation must hold for any displacement function \( \xi(s) \), so:

\[
\text{Mechanical equilibrium selects a shape that has constant curvature } \kappa = -(\Delta T)/F. \tag{7.2}
\]

Idea 7.2 confirms the intuitions at the start of this section:

- When we float an open thread on a surface, \( \Delta T = 0 \). If we pull the ends, then \( F > 0 \), and the thread stretches out straight (\( \kappa = 0 \)).
- When we float a closed loop of thread on a surface, then add some oil or detergent to the water it encloses, then \( \Delta T < 0 \). The thread jumps outward to form a circle (\( \kappa \) positive and constant).

### 7.2.3 Surfaces in space have two distinct curvature functions

Our real goal is to understand mechanical equilibrium of a 2D surface in 3D space. So we must make some substitutions in the preceding discussion:

- Line tension \( F \) along a skimmer \( \leftrightarrow \) surface tension \( T \) of a soap film, or the interfacial tension of a fluid–fluid interface.
- Interfacial tension difference \( \Delta T \) between two sides of skimmer \( \leftrightarrow \) pressure difference \( \Delta p \) between sides of our surface.
- Curvature of a curve in a plane \( \leftrightarrow \) … what?

To make progress, we must generalize Idea 7.1a to a formula for the change in area of a curved surface to first order in a small perpendicular displacement \( \xi \). Let’s proceed as before: At any chosen point \( P \), set up a tangent plane. Then measure the displacement \( \Delta(u, v) \) from the surface to the tangent, where \( u, v \) are two surface coordinates centered on \( P \). For example, starting from a sphere of radius \( r \) we could use \( u = (\theta - \theta_0)\alpha \) and \( v = (\varphi - \varphi_0)\alpha \), where \((\theta_0, \varphi_0)\) are the colatitude and the longitude of \( P \). If we measure \( \Delta \) with respect to the outward-pointing direction, then \( \Delta \geq 0 \).

- As before, the Taylor series expansion of \( \Delta \) has no linear terms: That’s what tangency means. Thus, to leading nontrivial order, \( \Delta \) is quadratic in \( u \) and \( v \) jointly. That function equals zero if the surface is flat, so we can use it to describe curvature.
- The quadratic part may be expressed as

\[
\Delta^{[2]}(u, v) = \frac{1}{2} \left( B_{11} u^2 + 2B_{12} uv + B_{22} v^2 \right). \tag{7.3}
\]

Unfortunately, the coefficients \( B_{11}, B_{12}, \) and \( B_{22} \) depend on our choice of coordinate system \( u, v \) for the surface. In one dimension, we removed this ambiguity by specifying arc length as the coordinate \( s \). But what’s the analog of that choice on a 2D surface?

Although there is no unique, standard coordinate system, we can at least restrict the choice by requiring that if we start at \( P \) and move a small distance along a straight line in coordinate space, then the arc length squared of the resulting curve on the surface must take the form

\[
ds^2 = du^2 + dv^2 + \cdots, \tag{7.4}
\]
where the ellipsis denotes terms of higher than quadratic order.\(^5\) If our coordinates don’t have that property, we can always find new coordinates that do have it just by applying a linear transformation to \(u\), \(v\). We’ll call any such choice **normal coordinates** for the surface near \(P\).

**Your Turn 7A**

a. Assume a spherical Earth. Look up the latitude \((\pi/2) - \theta_0\) and longitude \(-\varphi_0\) of, say, your hometown. You could choose the local coordinates \(u = \theta - \theta_0\) and \(v = \varphi - \varphi_0\), which are certainly centered, but do they satisfy Equation 7.4? If not, find a linear transformation that turns them into good coordinates.

b. Even with the choice you made in (a), does Equation 7.4 hold exactly, that is, without higher-order terms?

Once we find local coordinates that meet our criterion, they will still not be unique: Other choices will also obey Equation 7.4. However, all such choices are of the form

\[
\begin{bmatrix}
u' \\
u
\end{bmatrix} = S \begin{bmatrix} u \\ v \end{bmatrix} + \cdots,
\]

where \(S\) is a 2D rotation matrix, and the ellipsis again denotes possible higher-order terms. If we re-express Equation 7.3 in terms of \(u', v'\), then it will involve three new coefficients \(B_{11}', B_{12}', B_{22}'\). That is, none of these quantities *invariantly* characterizes the surface near \(P\), due to the residual coordinate freedom.

Luckily, there is a way out. Equation 7.3 expresses the quadratic function \(\Delta^{[2]}(u, v)\) in terms of a symmetric matrix \(\mathbf{B} = \begin{bmatrix} B_{11} & B_{12} \\ B_{12} & B_{22} \end{bmatrix}\). Upon rotation, its new form involves a transformed matrix \(\mathbf{B}' = (S^{-1})^T \mathbf{B} S^{-1}\). But 2D rotations have the special property that \(S' = S^{-1}\), so \(\mathbf{B}' = S \mathbf{B} S^{-1}\). The eigenvalues of a matrix are invariant under such “similarity” transformations, and hence do not care which local coordinates we chose (as long as they obey Equation 7.4). In the present context, the eigenvalues \(k_1\) and \(k_2\) are called the surface’s **principal curvatures** at \(P\). We may equivalently package them as the **Gauss curvature**, \(G = k_1 k_2 = \det \mathbf{B}\) and the **mean curvature**, \(H = (k_1 + k_2)/2 = \frac{1}{2} \text{Tr} \mathbf{B}\).

Now examine Figure 7.2. Panel (c) shows a case where both principal curvatures are zero. Panels (b,e) show cases where \(k_1 = 0\) while \(k_2\) is not zero but constant [lateral surfaces in panel (b)] or nonconstant [panel (e)]. Panel (a) shows both \(k_1\) and \(k_2\) nonzero with the same sign; both are positive. Finally, panel (d) shows opposite signs. Thus, the mean curvature is zero in panel (c), and potentially also in panel (d) (if \(k_1 = -k_2\) exactly). The Gauss curvature is zero in panels (b,c,e).

**Your Turn 7B**

Continuing Your Turn 7A, find Earth’s two principal curvatures at your hometown. (What about *my* hometown?)

---

\(^5\)The presence of the higher-order terms may be surprising—isn’t Equation 7.4, without any higher terms, just the pythagorean theorem? Indeed, on a flat plane, we may choose cartesian coordinates, in which the usual formula is exactly true. Certain curved surfaces may also admit such special coordinates; however, in general they don’t exist, and Equation 7.4 is the best we can do.
A cone has a sharp apex, so it shouldn’t surprise you that its mean curvature is infinite there, and hence nonconstant elsewhere. In fact, if we let \( r \) denote distance from the apex to \( P \), then axial symmetry implies that \( H = H(r) \), and you’ll show in Problem 7.2 that \( H \propto r^{-1} \).

7.2.4 The Young–Laplace formula describes a trade-off between surface tension and pressure

As in Section 7.2.1, we now imagine distorting a surface to a nearby one by moving each point \( P \) a variable distance \( \xi(P) \) perpendicular to the surface.\(^6\) We can now state the results we need, analogous to those in Idea 7.1:

a. To first order in perpendicular displacement \( \xi \), the surface’s total area change is the integral of its area element times \( 2H \xi \) (a local formula).

b. In contrast, the volume occupied by one side grows, and the other side shrinks, by the surface integral of its area element times \( \xi \) (another local formula, but without any factor of curvature). \(^7\)

We won’t prove Idea 7.5,\(^8\) but by now it should seem reasonable: Look at the example surfaces in Figure 7.2a,b,e. Flattening a patch of any of these surfaces will reduce the surface area. So Gauss curvature cannot be what controls this loss, because it’s zero in panels (b) and (e). Instead, all three of these surfaces have nonzero mean curvature. In contrast, panel (c) has extremal area and also zero mean curvature. So it’s reasonable to suppose that mean curvature controls the first-order change in area, as claimed in Idea 7.5a. To understand the factor of 2, consider a cylinder of radius \( a \). Its mean curvature is \( 1/(2a) \). If we expand its radius to \( a + \xi \), then a patch of area \( a \phi \, d\phi \, dz \) increases to \( (a + \xi) \phi \, d\phi \, dz \), a relative increase by a factor of \( 1 + (\xi/a) = 1 + 2H \xi \).

Now imitate the argument in Section 7.2.2, modified as at the start of Section 7.2.3. A soap bubble, or a fluid–fluid interface, costs some energy proportional to its surface area; the constant \( T \) is called surface or interfacial tension.\(^9\) In mechanical equilibrium it’s constant, because molecules can rearrange freely within the surface. A closed surface (closed soap bubble or liquid drop boundary) separates two sides that can have different hydrostatic pressures; in equilibrium, this pressure difference \( \Delta p = p_{\text{in}} - p_{\text{out}} \) is also constant throughout each region.\(^9\)

The equilibrium surface shape must minimize total free energy. Arguing as before

---

\(^6\)As in Section 7.2.1, \( \xi \) is to be measured along a particular perpendicular deemed outward-pointing.

\(^7\) See Section 7.2.4’ (page 111).

\(^8\) A soap film is an air–liquid–air interface. Again, interfacial tension is actually free energy cost per area.

\(^9\) Pressure can be nonconstant if a “body force” like gravity acts on the bulk of the fluid. In the experiments we are studying, the net effect of gravity involves the density difference of the two fluids and is negligibly small. Surface tension can also be nonconstant, for example, in the presence of temperature or chemical gradients (Marangoni effect), but those are nonequilibrium situations.
(Idea 7.2 but with Idea 7.5) now gives a result analogous to Equation 7.2:

\[
\text{Mechanical equilibrium selects a shape that has constant mean curvature. The value of mean curvature will be zero for an open soap film, or more generally } 2H = \Delta p / T \text{ for a closed bubble or fluid–fluid interface.}
\]

(7.6)

Pressure is measured in newtons per meter squared, whereas interfacial tension is in newtons per meter, so Idea 7.6 is dimensionally correct.

Section 7.2 has outlined some concepts and formulas needed to discuss curves and surfaces quantitatively. Although we didn’t prove the mathematical result Idea 7.5, it has led to Idea 7.6, which does accord with experience. Look at the examples in Figure 7.2, and note how the Young–Laplace formula applies to each one: Each is a possible equilibrium surface, except for the one in panel (e).

### 7.3 EFFECT OF ELECTRIC FIELD

#### 7.3.1 An electric field jump across an interface modifies the energy balance

Now you know why you have never seen a conical soap bubble or fluid–fluid interface. The only problem is that you have seen one in Figure 7.3. Contrary to Idea 7.6, this shape has mean curvature that is nonconstant and indeed diverges at the cone’s apex. What physics have we forgotten to include?

The new physics is that the lower fluid in the photo was electrically conducting, and the system was subjected to a strong electrostatic field. To see why this matters, recall that there is no static electric field inside a conducting body and hence no electric field energy there; any dielectric properties of the fluid are irrelevant. But there is field energy in empty space or an insulator, and unlike hydrostatic pressure, its density need not be uniform. Indeed, Chapter 6 argued provisionally that that density equals $\frac{1}{2} c E^2$. If we deform the interface, then the field energy cost changes proportional to the change of volume on the side with nonzero field.

Your Turn 7C

Translate the preceding words into a modified form of the Young–Laplace formula suitable for an interface between conducting and insulating fluids.

Benjamin Franklin told us (in modern language) that we may expect a nonconstant electric field in the region near a pointy conductor. Moreover, the field becomes huge near the point, so we can neglect any hydrostatic pressure difference (set $\Delta p = 0$) and attempt to balance the electric field energy against that associated with interfacial tension.

---

10See Problem 7.2.
11See Equation 6.15 (page 82). Chapter 53 will return to this point.
12See Chapter 5.
7.3.2 The modified mechanical equilibrium admits a conical point solution

Before asking about mechanical equilibrium, let’s first find what sort of static electric field could exist outside a cone-shaped conductor. It will be convenient to use spherical polar coordinates, because:

- The Laplace equation is separable in such coordinates; and
- They make axial symmetry easy to implement. For example, our boundary condition is simply that the cone with one particular value \( \theta_0 \) of polar angle must be an equipotential surface (Figure 7.3):

\[
\psi(r, \theta_0, \varphi) = 0 \quad \text{for all } r \text{ and } \varphi.
\] (7.7)

Following Chapter 5, let us therefore look for potentials of the form

\[
\psi(r, \theta, \varphi) = CN(r)M(\cos \theta), \quad \text{where} \quad M(\cos \theta_0) = 0.
\] (7.8)

Here \( C \) is an unknown overall constant. If such a solution exists, then our conducting cone will be the region \( \theta \geq \theta_0 \), and hence its half-opening angle will be \( \pi - \theta_0 \).

The unknown function \( N \) must obey\(^{13} \) the radial equation \( 2rN' + r^2N'' = \lambda N \) for some constant \( \lambda \). Moreover, we know how \( N \) must diverge at \( r \to 0 \). The electric field energy involves \( ||\vec{V}\psi||^2 \). Your result in Your Turn 7C says that its variation must balance the mean curvature, which as mentioned earlier diverges as \( r^{-1} \). So the electric field \( -\vec{V}\psi \) must diverge as \( r^{-1/2} \), which means that \( \psi \) itself, while not infinite, behaves like \( r^{1/2} \). Substituting that trial solution into the radial equation shows that the eigenvalue \( \lambda \) equals \( 3/4 \), and indeed, the solution is exactly \( N(r) = r^{1/2} \).

Meanwhile, the angular function obeys the Legendre equation (Equation 5.7, page 71):

\[
((1 - \mu^2)M')' = -\lambda M,
\]

where now prime indicates \( d/d\mu \), and \( \mu = \cos \theta \). For integer \( \lambda \), the solutions to this equation are the familiar Legendre polynomials. For other values, like our \( \lambda = 3/4 \), its solutions are called “Legendre functions.” Indeed, the standard form of the Legendre equation is

\[
(1 - \mu^2)M'' - 2\mu M' + [\ell(\ell + 1)] M = 0.
\]

Comparing shows that our \( M(\mu) \) is a Legendre function of order \( \ell = 1/2 \). It’s not a finite polynomial like the ones we’re used to, but it’s a perfectly well-defined function. You’ll evaluate it in Problem 7.3, but we can draw some simpler conclusions now.

We have found a unique solution, Equation 7.8, that satisfies the Laplace equation, is axisymmetric, and has the right kind of singularity at \( r \to 0 \). But we haven’t yet enforced the boundary condition Equation 7.7, which also requires \( M(\hat{\theta}_0) = 0 \). So remarkably, there is only one possible angle for an equilibrium cone singularity, regardless of the value of the interfacial tension. When you evaluate it in Problem 7.3, you’ll see that experimentally,
the angle shown in Figure 7.3 really is as predicted. This solution is often called the **Taylor cone**.

At last we have seen how a free, conical interface can be reconciled with surface tension. Figure 7.5a shows two electrodes shaped approximately as equipotentials of the solution to our equation, apart from a missing conical bit at the point labeled G. At the appropriate value of potential difference, a puddle of conducting fluid at G was observed to rise up and form the sharp point shown in Figure 7.3.

### 7.4 TECHNOLOGICAL APPLICATIONS

In 2002, J. Fenn shared a Nobel Prize, not for discovering the cone state, but in part for applying it. Fenn knew that at high applied potential, a molecular-scale jet of fluid can emerge from the apex of the cone (**electrospray**, Figure 7.5b). This proved to be a convenient way to gently isolate and ionize dissolved macromolecules without breaking them; it led to a big advance in mass spectrometry. When applied to a polymer solution, the result can instead be a fine fiber (**electrospinning**).

The Taylor cone is also important for colloid thrusters used in fine control of spacecraft.

### 7.5 PLUS ULTRA
7.5.1 A look ahead

Once again, a tensor quantity has popped out in the course of other business. Previously, this happened when we invented the quadrupole tensor;\(^{14}\) this time, the quadratic function \(\Delta^{(2)}(u, v)\) involved the coordinate-dependent, symmetric matrix \(B\). Later chapters will extend and systematize notions introduced informally in this chapter.

7.5.2 Other physical surfaces

We have barely scratched the surface of surfaces. Soap films and simple interfaces are characterized by a single parameter, the interfacial tension \(T\), but the description introduced in this chapter generalizes to more complicated situations. For example, fluid membranes, such as artificial lipid bilayers, resist bending with an energy cost per area of the form \((H - H_0)^2\), that is, different from the one that gave rise to the Young–Laplace formula. In this formula \(H_0\) is a constant, encoding a possibly asymmetry between the two sides of the membrane. Such membranes also resist local changes in areal density of molecules. Now when an external electrostatic field is applied it can actually create an instability where a hole opens in the membrane: \textit{electroporation} a phenomenon useful for the controlled introduction of foreign DNA into a living cell.

Generalizing still further, a cross-linked surface, such as the bacterial cell wall, will also resist shear deformation.

7.5.3 A glimpse of general relativity

Section 7.2.3 took some trouble to characterize a surface using intrinsically defined, local quantities (the scalar fields \(G\) and \(H\)). Only one of these, the mean curvature, was needed for the application in this chapter.

But the Gauss curvature has a remarkable property worth mentioning here. We defined curvature via a procedure involving points \textit{off} the surface (that is, via the deviation \(\Delta\) between the surface and its tangent plane). However, the Gauss curvature can be re-expressed solely in terms of distance properties \textit{within} the surface.\(^{15}\) We need not even imagine any surrounding 3D space. This realization set in motion B. Riemann’s study of intrinsic curvature for spaces of dimension greater than two. Much later, that framework was just what Einstein needed to understand gravitation.

Riemann found that in higher dimensions, Gauss’s simple scalar \(G\) becomes an entire tensor of intrinsic curvatures. Einstein proposed that Riemann’s curvature tensor plays a role roughly analogous to \(V^2\phi_N\) in the newtonian field equation,\(^{16}\) and that it also controls the separation of two nearby freely falling bodies.

\(^{14}\)Chapter 3.

\(^{15}\)Gauss called this fact his “theorema egregium” (outstanding theorem).

\(^{16}\)Equation 1.1 (page 17).
FURTHER READING

Semipopular:
Don’t miss the hilarious yet profound video: Lloyd Trefethen, *Surface tension in fluid mechanics* (National Committee for Fluid Mechanics films, 1963)
web.mit.edu/hml/ncfmf.html.

Intermediate:
Young–Laplace formula: Butt & Kappl, 2018; Safran, 2003; Nelson, 2020, §7.2.2.

Technical:
7.2.3' Metric and second fundamental form

We can rephrase the construction of Section 7.2.3 in a more elegant way, by using ideas to be developed in Chapter 34.

The quadratic part of the distance-squared function $ds^2$ defines a symmetric rank-$\left(\frac{\sqrt{2}}{2}\right)$ tensor field called the metric (also called "first fundamental form") of the surface.

The quadratic part of the deviation $\Delta^{\text{II}}$ defines a rank-$\left(\frac{\sqrt{2}}{2}\right)$ tensor field, called the "second fundamental form." We can use the metric to convert it to a rank-$\left(\frac{1}{2}\right)$ tensor ("raise an index"). The new tensor describes a linear transformation on tangent vectors, so its trace and determinant at any given point are invariants describing the surface at that point.

A similar situation arose in our discussion in Section 3.2. There we were working in flat 3D space, so we could just choose globally cartesian coordinates when defining the quadrupole tensor. We face the issue that there is some freedom to choose different cartesian systems, but again a different choice would amount to a similarity transformation acting on the components of $\mathbf{G}_n$. So again its three eigenvalues are intrinsic properties that characterize different kinds of quadrupole (uniaxial versus biaxial, Section 3.2'a).

7.2.4' Derivations of variational formulas

Here we establish the formulas in Idea 7.5.

Consider a 2-surface in 3-space with a point $P$ of interest to us. We can specify the surface near $P$ by a vector function $\mathbf{r}(u, v)$, where the parameters range over some region of the $uv$ plane. Let $\mathbf{n}(u, v)$ be a choice of perpendicular vector at each point of the surface, which we will deem "outward" even if the surface is not closed. Then the area of the surface can be written as

$$\Sigma = \int du dv \left| \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right|$$

(7.9)

Abbreviate $\partial / \partial u$ by $\nabla_u$ and so on, and let $J$ denote the square of the integrand above. Thus,

$$J = \left| \nabla_u \mathbf{r} \right|^2 \left| \nabla_v \mathbf{r} \right|^2 - (\nabla_u \mathbf{r} \cdot \nabla_v \mathbf{r})^2.$$

A new surface is specified by a perpendicular displacement function $\xi$ via $\mathbf{r}(u, v) = \mathbf{r}(u, v) + \xi(u, v)\mathbf{n}(u, v)$. Suppose that $\xi$ equals zero at the boundary of the chosen region in $u, v$. The first-order variation of the surface area is then

$$\delta \Sigma = \int du dv J^{-1/2} \left[ \nabla_u \mathbf{r} \cdot \nabla_u (\xi \mathbf{n}) \right] \left| \nabla_v \mathbf{r} \right|^2$$

$$\quad + \left( \nabla_u \mathbf{r} \cdot \nabla_v (\xi \mathbf{n}) \right) \left| \nabla_v \mathbf{r} \right|^2 - (\nabla_u \mathbf{r} \cdot \nabla_v \mathbf{r})(\nabla_u (\xi \mathbf{n}) \cdot \nabla_v \mathbf{r} + \nabla_v \mathbf{r} \cdot \nabla_u (\xi \mathbf{n})).$$

(7.10)

Now integrate by parts, using that $\xi = 0$ on the boundary, and write the result as the integral of $\xi$ times some function on the surface. We wish to find a convenient expression for that function at any point $P$ in terms of the surface shape near that point.

Equation 7.9 is unchanged under translations and rotations of $\mathbf{r}$, so we may suppose that our 3D coordinates are centered on $P$, and moreover, that the tangent to the surface is the $xy$ plane and $\mathbf{n}(P) = \hat{z}$. We can also shift the two parameters $u, v$ to center them on $P$ and make a linear transformation if needed to arrange that the leading terms are

$$\mathbf{r}(u, v) = u \hat{x} + v \hat{y} - \frac{1}{2} [u, v] \left[ \begin{array}{c} b_{11} \ b_{12} \\ b_{21} \ b_{22} \end{array} \right] \left[ \begin{array}{c} u \\ v \end{array} \right] \hat{z} + \mathcal{O}(3).$$
The last term represents contributions of order three or greater in \( u, v \). The constants \( B_{11}, B_{12} = B_{21}, B_{22} \) then have the same meaning as in Equation 7.3.

Next, note that
\[
\tilde{\nabla}_u \vec{r} = \hat{x} - (B_{11}u + B_{12}v)\hat{z} + O(2), \quad \tilde{\nabla}_v \vec{r} = \hat{y} - (B_{12}u + B_{22}v)\hat{z} + O(2).
\]
Hence, \( u \) and \( v \) are normal coordinates (Equation 7.4, page 103):
\[
||\tilde{\nabla}_u \vec{r}||^2 = 1 + O(2), \quad ||\tilde{\nabla}_v \vec{r}||^2 = 1 + O(2), \quad \tilde{\nabla}_u \vec{r} \cdot \tilde{\nabla}_v \vec{r} = O(2), \quad \text{and} \quad J = 1 + O(2).
\]
The unit vector perpendicular to the surface is then
\[
\hat{\eta}(u, v) = \frac{\tilde{\nabla}_u \vec{r} \times \tilde{\nabla}_v \vec{r}}{||\tilde{\nabla}_u \vec{r}|| ||\tilde{\nabla}_v \vec{r}||} = \hat{x} + \hat{y} + \hat{z} + O(2).
\]
Putting it all together, Equation 7.10 becomes
\[
\delta \Sigma = -\int dudv \xi \hat{\eta} \cdot \left[ \tilde{\nabla}_u \left( J^{-1/2} \tilde{\nabla}_u \vec{r} ||\tilde{\nabla}_u \vec{r}||^2 \right) + \tilde{\nabla}_u \left( J^{-1/2} \tilde{\nabla}_v \vec{r} ||\tilde{\nabla}_v \vec{r}||^2 \right) \right. \\
- \tilde{\nabla}_v \left( J^{-1/2} \tilde{\nabla}_u \vec{r} \cdot \tilde{\nabla}_u \vec{r} \right) - \left. \tilde{\nabla}_v \left( J^{-1/2} \tilde{\nabla}_v \vec{r} \cdot \tilde{\nabla}_v \vec{r} \right) \right].
\]
Evaluating the integrand at \( P \) gives
\[
-\xi \left[ \tilde{\nabla}_u (-B_{11}u + \cdots) + \tilde{\nabla}_v (-B_{22}v + \cdots) \right] = 2\xi H.
\]
Thus, Equation 7.10 is equivalent to the first statement in Idea 7.5.

The second statement concerns the volume of a thin shell of perpendicular thickness \( \xi(u, v) \).

Multiply the area element by the thickness to get the volume.
7.1 Variation of arc length and area
A curve in a plane is specified by a vector function \( \vec{c}(s) \), where \( s \) is arc length, \( 0 \leq s \leq L \). Let \( \vec{n}(s) \) be a field of perpendicular vectors all along the curve. A new curve is specified by a function \( \xi(s) \) via \( \vec{c}(s) = \vec{c}(s) + \vec{n}(s)\xi(s) \). The displacement function \( \xi \) equals zero at \( s = 0 \) and \( L \).

a. Establish the formula in Idea 7.1a (page 102). [Hints: Use integration by parts. Although the parameter \( s \) still runs from 0 to \( L \), it’s no longer arc length for the new curve, so the new total length will no longer be \( L \).]

b. Also establish the formula in Idea 7.1b.

7.2 Mean curvature of a cone
Show that the mean curvature of a cone with opening half-angle \( \alpha \) is \( H(r, \varphi) = (\cot \alpha)/(2r) \). Here \( r \) is distance from the cone’s apex to the point of interest, and \( \varphi \) is angular position on each “latitude” line. [Hint: If you have difficulty, first draw a very wide cone, with \( \alpha \) just slightly less than \( \pi/2 \). It’s nearly a plane, so its curvature must be smaller at a given \( r \) than that of a narrower cone. Make sure your derivation accounts for this.]

7.3 A pointed remark
Finish the derivation of the stability problem started in the main text. Set up spherical polar coordinates, and consider a cone of electrically conductive fluid occupying the region of space with \( \theta \geq \theta_0 \), as in Figure 7.3 (page 101). Thus, the half-opening angle of the cone is \( \pi - \theta_0 \). Take the electrostatic potential to have the form Equation 7.8, where \( N(r) = r^{1/2} \), \( M \) is the Legendre function of order 1/2, and \( C \) is an undetermined overall constant. You may assume that the pressure drop \( \Delta p \) is everywhere zero and the interfacial tension is fixed to some nonzero constant value \( T \).

a. Use a computer to find the only zero of the function \( M \) in the range \(-1 < \cos \theta_0 < 1\), and in that way predict \( \theta_0 \) and hence \( \pi - \theta_0 \).

b. Evaluate the electrostatic potential throughout the plane \( y = 0 \) (or just the half-plane with \( \varphi = 0 \)), display it as a contour plot, and comment.

c. Numerically evaluate the derivative \( dM(\mu)/d\mu \) at \( \theta_0 \), where \( \mu = \cos \theta \).

d. Using your results in (a–b), write a formula for the electric field squared just outside the surface (\( \theta \leq \theta_0 \)).

e. Generalize the Young–Laplace formula (Idea 7.6, page 106) appropriately by finding an expression for the electrostatic field energy density just outside the surface and setting it equal to \( 2TH \), where \( H \) is the mean curvature from Problem 7.2, and \( T \) is the interfacial tension of oil and water. Substitute the result that you found in (c).

f. Obtain a prediction for the constant \( C \) in terms of \( T \) and the relative permittivity \( \varepsilon /\varepsilon_0 \) of oil. After looking up those values, for example for an oil–water interface, then we can find \( \psi_{\text{cone}} \), the potential drop needed to create a conical singularity.

g. Evaluate \( \psi_{\text{cone}} \) at a distance \( r = 1 \) cm from the point of the cone, using estimated values...
appropriate for Figure 7.3 (page 101): $T \approx 3.7 \cdot 10^{-2} \text{ N/m}$ and $\varepsilon / \varepsilon_0 \approx 2.2$. Is such a potential achievable in the lab?
Chapter 8

Charge Flux, Continuity Equation, and Ohmic Conductors

Cavendish was his own galvanometer. In order to compare the intensity of currents he caused them to pass through his own body, and by comparing the intensity of the sensations he felt in his wrist and elbows, he estimated which of the two shocks was the more powerful.

— James Clerk Maxwell

8.1 Framing: Conservation

We now gradually start to look at non-static situations. First we must get precise about the meaning of charge flux, then find a useful identity about it that follows from the conservation of charge.

Electromagnetic phenomenon: A traveling nerve impulse or muscle contraction leads to a multipolar electrical disturbance that can be measured from far away.

Physical idea: In quasi-static conditions, a current source in a conducting medium leads to the same sort of multipole fields as in vacuum.

8.2 A Graphical Argument for the 1D Case

Imagine a long, thin pipe with some conserved “stuff” inside. Maybe it’s air, and the “stuff” is mass. Define a 1D density \( \rho^{(1D)}(t, z) \) (units kilograms per meter). At any \( z_0 \), also define the 1D mass flux \( j^{(1D)}_m \) as the net rate at which mass crosses the point \( z = z_0 \), moving from smaller to larger \( z \). Thus, a particle crossing in the opposite direction makes a negative contribution to the 1D flux of mass.

Mass therefore piles up in a small region near \( z_0 \) of width \( \Delta z \) at the rate \( j^{(1D)}_m(z_0) - j^{(1D)}_m(z_0 + \Delta z) \) (Figure 8.1). But the rate of pileup is also \( \frac{\delta}{\delta t}(\rho^{(1D)}_m \Delta z) \). Dividing through

Figure 8.1: Continuity equation warmup. In time \( \delta t \), two lumps of some conserved quantity enter a small zone of length \( \delta x \) from the left, but only one exits from the right. Thus, \( \frac{\delta j^{(1D)}}{\delta t} < 0 \). This imbalance leads to \( \frac{\delta \rho^{(1D)}}{\delta t} > 0 \).
Figure 8.2: Graphical understanding of the continuity equation. (a) Three trajectories in one spatial dimension. (b) A charge-conserving interaction in two spatial dimensions. The colored face of the cube shown is transverse to \( \vec{y} \), so any charge traversing it (in this case, the electron) contributes to \( J_z \).

by \( \Delta z \) yields

\[
\frac{\partial [\text{1D}] j_m}{\partial z} + \frac{\partial [\text{1D}] \rho_m}{\partial t} = 0. \quad \text{continuity, 1D} \quad (8.1)
\]

Here is a better pictorial way to understand Equation 8.1: Imagine a small range of space and time near \((t, z)\) (dashed box in Figure 8.2a). Then conservation of mass\(^1\) implies that the net mass entering this spacetime region is zero. For example, Figure 8.2a shows three particle trajectories: Point masses 1 and 2 contribute to the first term of Equation 8.2 at points 1 and 2', whereas points 2, 3, 1', and 3' contribute to the second term. Because every trajectory that enters the dashed box must also leave it, these terms must sum to zero:

\[
0 = \Delta t \left( \Delta z \frac{\partial [\text{1D}] j_m}{\partial z} \right) + \Delta z \left( \Delta t \frac{\partial [\text{1D}] \rho_m}{\partial t} \right). \quad (8.2)
\]

Dividing Equation 8.2 by \( \Delta t \Delta z \) recovers Equation 8.1.

8.3 TWO OR MORE DIMENSIONS

8.3.1 Any local conservation rule leads to a continuity equation

From now on, we will be more interested in electric charge than in mass, so unless otherwise stated the symbol \( \vec{j} \) will refer specifically to the flux of charge. Also, we will generalize to allow particles to exchange charge, merge, or even explode as shown in Figure 8.2b. In between such interactions, each particle's trajectory is a curve in spacetime carrying a fixed quantity (its “charge”). Even in an interaction, this number is conserved locally (at each vertex separately). For example, in the weak decay\(^2\) shown, the incoming line has charge zero, whereas the outgoing lines have charges 0, \( e \), and \( -e \).

\[^1\]Later, we will see that mass isn’t exactly conserved. Here we are just pursuing an illustration familiar from newtonian physics. Really we will be interested in charge, which is conserved, even in relativity.

\[^2\]The reaction shown is also an example of two other local conservation laws, those of lepton and nucleon numbers. Each has its own continuity equation analogous to the one for charge.
The overall charge entering any fixed region of spacetime, like the box shown in the figure, is therefore once again zero. In this example, we have:

- 0 (neutron trajectory enters via bottom face of the box);
- (e) (proton and antineutrino, total charge +e, exit via top face of the box);
- (e) (electron exits via right face of the box).

Those quantities do sum to zero: Everything that enters the box must also exit, carrying its charge.\(^3\)

In 2D, charge density \(\rho_q^{[1D]}(t, \vec{r})\) has units coul m\(^{-2}\). Charge flux \(j^{[2D]}_n(t_0, x_0, y_0)\) is defined as the net charge per time that crosses a short line segment of constant \(y = y_0\) near position \((x_0, y_0)\) and time \(t_0\), divided by the segment length. Here again, “net” means that a charge \(q\) passing from smaller to larger values of \(y\) contributes \(q\), while the same charge passing the opposite way contributes \(-q\).

What’s new compared to one dimension is that now we get a second component of flux, \(j^{[2D]}_1\), when we consider charge crossing a short line segment with constant \(x\). Overall, then, \(j^{[2D]}(t, \vec{r})\) in 2D is a vector field with units coul m\(^{-1}\) s\(^{-1}\), that is, A m\(^{-1}\).

Often, it’s reasonable to think of charge as a “river” of many particles, defining an essentially continuous flow. The overall charge entering the infinitesimal spacetime box in Figure 8.2(b) is then:

- \(+\rho_q^{[1D]}(0, 0, 0)\Delta x\Delta y\) from the \(t = 0\) (lower) face (plus terms of higher order in \(\Delta x\) and \(\Delta y\));
- \(-\rho_q^{[1D]}(\Delta t, 0, 0)\Delta x\Delta y\) from the \(t = \Delta t\) (upper) face;
- \(+j^{[2D]}_2(0, 0, 0)\Delta x\Delta t\) from the \(y = 0\) (left) face;
- \(-j^{[2D]}_2(0, 0, \Delta y)\Delta x\Delta t\) from the \(y = \Delta y\) (right) face;
- \(+j^{[2D]}_1(0, 0, 0)\Delta y\Delta t\) from the \(x = 0\) (rear) face;
- \(-j^{[2D]}_1(0, \Delta x, 0)\Delta y\Delta t\) from the \(x = \Delta x\) (front) face.

As in 1D, these contributions must sum to zero. Grouping them in pairs and using a Taylor expansion gives

\[0 = \left( -\frac{\partial}{\partial t}\rho_q^{[1D]} - \frac{\partial}{\partial y}j^{[2D]}_2 - \frac{\partial}{\partial x}j^{[2D]}_1 \right) \Delta x \Delta y \Delta t. \tag{8.3}\]

(Higher order terms vanish when we take the limit of a small box.) The spacetime box may be located anywhere, so analogously to Equation 8.1 we find

\[0 = \frac{\partial}{\partial t}\rho_q^{[1D]} + \nabla \cdot j^{[2D]}, \quad \text{continuity equation} \tag{8.4}\]

We can do the whole derivation again, with any number of spatial dimensions (for example, three). This time, the relevant definition says that

\(\tilde{j}_n\) is the function that, when integrated over \(\Delta y\Delta x\Delta t\) at fixed \(x\), yields the net charge crossing a small surface element from smaller to larger \(x\) during a small time interval. The other components are defined similarly.

\(^3\)And particles that never enter the box also never exit it.
The units of charge density and flux depend on dimensionality, but they always obey the same continuity equation.

### 8.3.2 The continuity equation bridges local and global conservation

Section 8.3.1 argued that local conservation of charge implies the continuity equation. A simple but important consequence comes when we integrate both sides of the continuity equation over a region of space containing all the charges at a particular time and use the divergence theorem:

\[
\frac{d}{dt} \rho_q + \nabla \cdot \mathbf{j} = \frac{d}{dt} \int d^3r \rho_q = \int \int d^3r \frac{\partial \rho_q}{\partial t} = \int \int \nabla \cdot \mathbf{j} = \int d^2S \cdot \mathbf{j} = 0. \tag{8.6}
\]

The last step follows because there are no charges, hence no charge flux, at the boundary of the region.

Not surprisingly, the local conservation of charge that led to the continuity equation implies global charge conservation.

### 8.4 REMARKS

- In this book, \textit{flux} will always mean “conserved stuff per transverse dimensions per time.”\footnote{As mentioned in Section 0.1.1 (page 2), This book reserves the word “density” to mean only “stuff per unit volume,” so we do not use the term “current density” to mean charge flux. Unfortunately, some books use “magnetic flux” to mean something quite different; we will not use that term.} In 1D, there are no transverse dimensions and \( j^{[1D]} \) was just stuff per time. In 2D, there is one dimension transverse to a given direction.\footnote{In Figure 8.2b, the colored cube face is transverse to \( \mathbf{y} \) and has one spatial dimension with extent \( \Delta x \). (In panel (a), the left and right edges have no spatial extent.)} In 3D, there are two.
- In \( D \) spatial dimensions we can integrate charge flux over a \((D - 1)\)-dimensional surface element to obtain the \textbf{current} \( I \) through that surface (coul/s).
- The continuity equation is a purely kinematic identity. It is valid regardless of whether the particle trajectories obey any equation of motion. So when we later discard newtonian dynamics in favor of relativistic dynamics, Equation 8.4 will reemerge \textit{unchanged}. It merely expresses local conservation of charge (or any other scalar quantity); beyond that physical assumption, it’s just bookkeeping.
- In a stationary situation, where charge density is unchanging in time (perhaps zero), the continuity equation guarantees that \( \mathbf{j} \) is divergence-free.

### 8.5 NONSTATIC SITUATIONS
8.5 Nonstatic Situations

8.5.1 Conductivity, resistivity, conductance, resistance

Many materials are insulators: \( \vec{j} = 0 \). Some others are approximately ohmic: they develop currents via a response function of the form

\[
\vec{j} = \kappa \vec{E}. \quad \text{ohmic material}
\]  

(8.7)

The constant \( \kappa \) is a material parameter called the conductivity of the material. Metals such as copper, at ordinary frequencies, are approximately ohmic, as is salt water.

Equation 8.7 may not look like “Ohm’s” “law” as it appeared in first-year physics. To make the connection, consider a thin wire of length \( h \) with cross-sectional area \( \Sigma \). Total current \( I \) flows, leading to a charge flux \( j = I/\Sigma \), or equivalently, \( I = \kappa E \Sigma \). The electric field within the wire leads to a potential drop as usual, \( \Delta \psi = hE \). Thus,

\[
\Delta \psi = IR \quad \text{where} \quad R = h/(\Sigma \kappa).
\]  

(8.8)

The resistance \( R \) depends on the geometry of the wire (via \( h \) and \( \Sigma \)) as well as on the material (via \( \kappa \)). The SI unit for resistance is called ohm and abbreviated \( \Omega \). The SI unit for conductivity is then \( \Omega^{-1} \) m\(^{-1} \). Another name for \( \Omega^{-1} \) is the siemens,\(^8\) abbreviated S.

Other quantities appearing in scientific literature include resistivity, a material parameter defined\(^9\) as \( 1/\kappa \), and conductance, defined as \( 1/R \).

Equation 8.7 is called “dissipative” because it relates \( \vec{j} \), a quantity that changes sign under time reversal, to \( E \), a quantity that doesn’t. Thus, this formula breaks time-reversal invariance: It entails the irreversible conversion of electric energy into heat. Let’s quantify that claim. Some external agency must expend energy \((dq)\Delta \psi\) to push a lump of charge through our wire. Multiplying \( \Delta \psi \) by the total rate of charge transport thus gives the power absorbed by the wire as

\[
P = (\Delta \psi)I = I^2R = (\Delta \psi)^2/R.
\]  

(8.9)

Indeed, that power ends up as heat, an effect called ohmic heating.\(^{10}\) If you plug an appliance with an internal short circuit \((R \lesssim 1 \Omega)\) into the wall \((\Delta \psi \text{ fixed})\), you get a lot more heat than when you plug in a normal light bulb \((R \gg 1 \Omega)\).

8.5.2 Salt water conducts electricity via the motions of ions

Passing direct current through a solution of table salt gives a vivid clue to the mechanism of conduction. Bubbles appear at the electrodes, and soon there is an unmistakable odor

---

\(^6\)More precisely, an insulator carries no free current. In nonstationary situations, the movement of bound charge leads to a “dielectric displacement charge flux” (Section 49.2.1, page 688).

\(^7\)Discovered by H. Cavendish, half a century before G. Ohm. Cavendish failed to publish this observation, and many others as well. So many exotic materials are not ohmic that it’s a bit silly to call it a “law.” It is, however, a useful approximation for certain materials, in certain conditions.

\(^8\)Don’t confuse siemens with the sievert (Sv), a unit of ionizing radiation dose, nor with the svedberg (also abbreviated S), used to describe sedimentation rate. An obsolete, whimsical synonym for siemens, still occasionally seen, is “mho,” abbreviated Ω\(^{-1}\).

\(^9\)The suffix “-ivity” generally denotes a material property independent of the size of a sample. The suffix “-ance” generally denotes a property of a specific object.

\(^{10}\)Sometimes called “Joule heating.”
of chlorine. That odor hints at what is happening: Chloride ions are attracted to the anode. When they arrive there, each Cl\(^{-}\) surrenders its excess electron, becoming a neutral atom of chlorine. Those chlorines bond in pairs and leave the solution as pungent Cl\(_2\) gas. The overall effect is thus that electrons leave the solution into the anode, even though free electrons did not literally pass through the solution.

A typical diffusion constant for an ion or small molecule in water at room temperature is \(D \approx 1 \mu m^2/\text{ms}\). What’s the mean velocity of a chloride ion in solution, in an applied electric field of 1 V/cm?

**Solution:** The viscous drag coefficient is given by Einstein’s relation, \(k_B T / D\). It has the same dimensions as force per velocity, so take the force \(qE\) on one ion and divide by the drag coefficient:

\[
\frac{qED}{k_B T} = (1.6 \times 10^{-19} \text{ coul})(100 \text{ V/m})(1 \mu \text{m}^2/\text{ms})(4 \text{ pN nm})^{-1}
\]

\[
= 1.6 \times 10^{-19} \times 100 \times \frac{10^{-12}}{10^{-3}} \frac{1}{4 \times 10^{-12} 10^{-9}} \frac{\text{J coul m}}{\text{m} \text{s} \text{N}}
\]

\[
\approx 10^{-5} \text{m/s}.
\]

Another name for the reciprocal of viscous drag coefficient is mobility.

The mean (drift) velocity just found may seem laughably small compared, say, to the thermal motion of each ion. But unlike thermal motion, the drift is *not random*, and there are lots of ions, so the resulting conductivity can be significant. **Section 8.5.2’ (page 128) gives more details about electrolysis of water.**

### 8.5.3 Conductivity depends on temperature

Even if a material may be adequately described as being ohmic, in general its electrical conductivity will vary with temperature.\(^1\) For example:

- Electrical conduction in salt water depends on thermal collisions that degrade the ordered motion of ions as they respond to an electric field, so \(\kappa\) increases slightly with increasing temperature.
- Conduction in metals is limited by imperfections in the crystal structure. Even the thermally excited flexing of a perfect crystal impedes conduction (by scattering electron waves), and so \(\kappa\) decreases with increasing temperature.\(^2\)
- Conduction in a semiconductor is limited by the sparse population of electrons thermally excited into the crystal’s conduction band, and so \(\kappa\) increases sharply with increasing temperature.

---

\(^1\)Other transport properties such as thermal conductivity are also variable.

\(^2\)See Media 3 and Section 10.2.2’ (page 156).
8.6 QUASI-STATIC APPROXIMATION

We will be interested in situations where everything is changing slowly in time, for example, on the millisecond time scale characteristic of nerve impulses. There is a useful simplification in this case.

In static (zero-frequency) situations, Section 2.4 argued that charge will rearrange to erase any electric field inside a conductor. Even at nonzero frequency, we get the same conclusion for a perfect conductor. What about a non-static situation with a non-perfect conductor? Charge takes time to move around, because moving fast incurs large frictional resistance. Combining the continuity equation, the ohmic hypothesis, and the Gauss law yields

\[ \frac{\partial \rho}{\partial t} = -\nabla \cdot (\kappa \mathbf{E}) = -\kappa \varphi / \varepsilon. \quad \text{spatially uniform, ohmic material} \]

We see that

In a spatially uniform ohmic material, any initial nonuniformity of net charge density gets exponentially suppressed over time scales longer than \( \varepsilon / \kappa \). \hfill (8.10)

**Your Turn 8A**

Check that \( \varepsilon / \kappa \) does have dimensions of time.

Here is another clue that an ohmic material breaks time-reversal invariance: Idea 8.10 says that an initial fluctuation in charge density will always shrink over time. Note that the restriction to uniform material is important: Net charge can still crowd up against an insulating layer, as it does in a charged capacitor.

For salt solution at concentration 100 mM, we can look up \( \kappa \approx 0.1 \, \Omega^{-1} \text{m}^{-1} \). We also know that pure water is highly polarizable: \( \varepsilon \approx 80 \varepsilon_0 \) at low frequency.\(^{13}\) So for frequencies below about 100 MHz, we can assume that salt water is everywhere locally neutral, and hence also that \( \nabla \cdot \mathbf{E} = 0 \), just as in electrostatics! This simplification will help us in Section 8.7 and in later chapters.

\(^{13}\)See Section 6.8 (page 87).

8.7 ELECTROENCEPHALOGRAM/ELECTROCARDIOGRAM

8.7.1 A steady current source in solution again leads to a Poisson equation

In your brain, vast numbers of nerve cells (neurons, Figure 8.3) are communicating with one another and with your muscles, sensory receptors, and even hormone-secreting cells throughout your body. The mechanism by which these signals travel long distances, without diminution, is the subject of Chapters 11–12. Right now we will instead study noninvasive ways to detect them.
Figure 8.3: **Two typical animal cells drawn to scale.** Upper: Human skeletal muscle cell. Dark blobs are cell nuclei. Lower: Human neuron. The unbranched tube on right is the “output line” (axon), which may extend for up to a meter to communicate with another neuron, a muscle cell, or an hormone-secreting cell. Other tubes represent “input lines” (dendrites), each of which communicate with other neurons (including sensory receptors). [Art by D. S. Goodsell.]

Figure 8.4: **A current dipole.** When immersed in an ohmic conductor such as salt water, this source sets up a distributed current, and hence an electric field. Both $\vec{E}$ and $\vec{j}$ assume a dipolar form far away.

To begin, think about the simple system in Figure 8.4: A battery is connected to two thin wires, each insulated except at their tips; everything is immersed in a salt solution. Each tip is a pointlike boundary between a good conductor and an ohmic medium, so Idea 8.10 does not apply and net charge can build up at the tips. Elsewhere, however, the electric field does obey $\nabla \cdot \vec{E} = 0$, so we have already done the math: We get the same electric field pattern as from a static charge dipole in vacuum! Unlike in vacuum, however, a continuous current flows: Charge emerges from one tip, passes through the solution, and returns to the other tip. To the outside world, each tip is a pointlike source or sink of charge. Very close to the + tip, charge emerges isotropically, following the electric field lines via Equation 8.7, and similarly at the other tip. The whole arrangement of source and sink is called a **current dipole**.

### 8.7.2 An isolated neuron creates an exterior potential

Next, imagine a single neuron in salt solution. The interior of the neuron is filled with a different solution of salts and various other molecules. More precisely, the interior and exterior fluids have well matched overall osmotic pressure, which is why delicate structures like cells and their axons, bounded by fragile membranes, can exist. But the concentrations of particular ions are quite different inside and outside of the cell. Figure 8.5a shows some of these concentrations for a well-studied axon in the squid *Loligo forbesi*:

- The exterior sodium ions have a big density gradient pushing them toward the interior, and an electric potential jump with the same sense, but they are frustrated by...
Figure 8.5: **Exterior effect of ion channel opening.** (a) Nonequilibrium ion concentrations inside and outside a “resting” nerve axon. The cell membrane separates inner and outer fluids with strikingly different ion concentrations, of which two key players are shown. The membrane prevents ion migration that would restore equilibrium. Ion pumps in the cell body (*not shown*) continually export sodium and import potassium, maintaining these resting concentrations at the expense of metabolic energy. (b) An imagined situation in which a narrow strip of the cell membrane opens sodium-specific ion channels (*circle*); elsewhere the channels remain closed. The influx of current locally discharges the capacitance of the cell membrane by releasing exterior cations that were initially attracted to the membrane (*lower right*) and by releasing interior anions localized to the membrane (*not shown*). The interior electric potential has its resting (negative) value $\psi^0$ at $x = \pm \infty$ but rises near the zone of open channels. (c) Equivalent circuit. The battery symbol represents the entropic tendency for sodium ions to enter the cell if permitted. (d) The linear density of current source seen by the outside world, $j$, has net current, and net current dipole, both zero.

the barrier membrane.\(^{14}\)

- Far from the cell the concentrations are uniform, but just outside the membrane there is a cloud of excess $+$ charge, attracted to the negative interior even though they cannot get there.\(^{15}\) Just inside, there is a corresponding depletion layer of $+$ charge, repelled by the exterior cloud. Similar but opposite remarks apply to the negative ions. In short, the resting membrane’s state amounts to a charged capacitor, with a static electric field created by those two layers of charge.\(^{16}\)

- The interior potassium ions are subject to conflicting forces: The negative interior potential tends to keep them in, but is overbalanced by the high interior concentration, leading to a net electrochemical force directed outward. Like Na\(^+\), however, the K\(^+\) ions are blocked by the membrane.

These nonequilibrium concentrations, enforced by the cell membrane, amount to a continuously distributed reservoir of excess free energy, constantly maintained by active transport of sodium out of, and potassium into, the cell.

Section 6.9 (page 88) mentioned that cell membranes are studded with doorways, ion channels that, while normally closed, can open upon command, permitting the transport

\(^{14}\)Recall Section 6.9 (page 88).

\(^{15}\)See Section 10.3.3 (page 146).

\(^{16}\)See Section 6.9 (page 88).
Exterior effect of a nerve impulse. An elaboration of Figure 8.5. (a) This time, the zone of open sodium-specific channels (circles) is moving to the right, and trailed by a zone of open potassium-specific channels (squares). The high interior concentration of potassium leads to their escape in that zone. Later still, all channels close and the membrane repolarizes (far left). So the interior electric potential has its resting (negative) value \( \psi^0 \) at \( x = \pm \infty \) but rises near the traveling wave. (b) Equivalent circuit. In between the two traveling fronts the membrane is discharged, but it slowly recharges after the double front has passed (left). (c) Imagined profile of the linear density of current source seen by the outside world, \( J \), at one instant of time. It is more complicated than in Figure 8.5, but again has net dipole moment zero (Equation 8.13).

of specific ion types across the membrane. A nerve impulse begins with the opening of ion channels specific for sodium in a small patch of membrane. As sodium ions rush into the long, narrow interior of the axon, a region of the axon becomes depolarized: Its electrostatic potential rises toward zero from its resting negative value. Some of the nearby ion cloud is free to depart, discharging the membrane capacitance locally (Figure 8.5b).

A real nerve impulse (action potential) is more complicated than the situation just outlined (Figure 8.6).

- The zone of open channels moves, traveling along the axon at constant speed. Chapter 12 will explore why this happens; for now, we treat this as a given empirical fact and explore measurable effects on the world outside the axon.
- Also, the zone of transiently open sodium channels is trailed by another limited zone of open channels, permeable only to potassium ions (Figure 8.6a).
- The overall effect is to create a traveling wave of momentary depolarization.

The world outside the axon sees the leading wavefront as a net source of positive charge from the released ions (right side of Figure 8.6abc), adjacent to a sink as sodium ions enter. Still farther to the left in the figure, there is a source (potassium ion outflow), and finally a sink as the membrane capacitance recharges to its pre-impulse value. Later still, all channels close and the whole system returns to its resting state after the impulse has passed.

In short, at any time \( t \) the exterior fluid sees a traveling array of apparent charge sources and sinks localized along the axon (Figure 8.6c). This current spreads into the surrounding fluid following the quasi-static rule, Idea 8.10. At any instant, it obeys \( \nabla \cdot \mathbf{J} = 0 \) with boundary conditions at the axon determined by the form of the nerve impulse. But this equation implies \( \nabla \cdot \mathbf{E} = 0 \), which is just the Laplace equation. We therefore know...
that far from an isolated axon, the electric field will have a multipole expansion of the usual form. Instead of a distribution of point charges, as we had in vacuum, we have a continuous distribution of current sources along the axon, but the math is the same as before.\footnote{Our picture strictly applies only to an isolated axon in solution. Corrections can be made for the inhomogeneity in the tissues of a complete animal.}

### 8.7.3 Electroencephalogram

Figure 8.6 looks complicated, but we can get its main qualitative feature by remembering charge neutrality. The axon’s cross-sectional area $\Sigma$ and its conductivity $\kappa$ determine the internal axial current $I_x$ created by the varying interior potential $\psi_{in}$ via the ohmic property of the interior salt solution:

$$I_x = -(\kappa \Sigma) \frac{d}{dx} \psi_{in}. \quad (8.11)$$

In an insulated wire, the continuity equation Equation 8.1 would require that nonuniformity of this current leads to charge buildup with rate proportional to the gradient of Equation 8.11. But for an axon, neutrality may be preserved if charge instead passes through the membrane. Charge can cross either literally, by ions passing through channels, or effectively, by charging/discharging the membrane capacitance (both mechanisms are shown in Figure 8.6). Either way, the axon maintains local neutrality by releasing charge to, or accepting it from, the exterior, forming the line of sources and sinks mentioned earlier. Each segment $dx$ releases charge at the rate $\mathcal{J}(x)dx$, where the linear density of current sources is

$$\mathcal{J} = -\frac{dI_x}{dx} = +\kappa \Sigma \frac{\partial \psi_{in}}{\partial x} \frac{\partial}{\partial x}. \quad (8.12)$$

**Your Turn 8B**

Confirm that this expression has appropriate dimensions.

The expression just found for $\mathcal{J}$ is a total derivative, and the potential approaches a constant at $x = \pm \infty$, so the net current source, $\int dx \mathcal{J}$, equals zero. Moreover, the quantity $x\mathcal{J}$ can be written as

$$(\kappa \Sigma) \frac{\partial}{\partial x} \left( x \frac{\partial \psi_{in}}{\partial x} - \psi_{in} \right),$$

which is also a total derivative. Because the potential approaches the same constant value at $x = \pm \infty$, we see that the dipole moment of the current source also equals zero:

$$\int dx x \mathcal{J} = (\kappa \Sigma) \left( x \frac{\partial \psi_{in}}{\partial x} - \psi_{in} \right)_{-\infty}^{\infty} = \kappa \Sigma (\psi_{in}(\infty) - \psi_{in}(-\infty)) = 0. \quad (8.13)$$

Hence, the leading-order electric field far from the axon is in general of quadrupole form (Figure 8.6c).\footnote{\textsuperscript{18} Other parts of a neuron, for example its dendrite, may also depolarize, potentially giving rise to a dipole contribution.} Any one nerve impulse will create extremely small distant currents
Figure 8.7: Exterior effect of a muscle cell activation. (a) Again there is a traveling wave of ion channel opening. However, repolarization is much slower than in a nerve axon, so the rise in interior potential persists all the way to the starting point of the activation (far left). (b) This time, there can be a nonzero dipole in the current released to the exterior region.

and fields. However, the concerted firing of impulses on many parallel axons in the brain can create a macroscopically measurable effect. Electric fields set up by the internal current can in turn penetrate even an intervening electrical insulator (such as the skull and surrounding skin). In this way, at least major brain activities can be measured noninvasively simply by attaching external electrodes to the skin and measuring the electric potential, a procedure called electroencephalography (EEG).

8.7.4 Electrocardiogram

Muscle cells also support traveling waves of membrane depolarization much like those in nerve cells, with the important differences that:

- Depolarization triggers the muscle cell to contract; and
- A single depolarization wave spreads over the entire cell for the duration of a contraction (Figure 8.7). Thus in this situation, $x = \pm \infty$ may have different potentials, Equation 8.13 does not apply, and the dipole moment of the current source need not equal zero.

Muscle tissue consists of huge numbers of parallel fibers that all contract in unison, leading to a big net dipole moment of the current distribution. Again, exterior electrodes on the skin can easily pick up this signal, determining not only the magnitude of the dipole (traditional electrocardiogram\textsuperscript{19}) but also its spatial direction (vector electrocardiogram, Figure 8.8). The time course of this net dipole vector is a more detailed diagnostic of heart disease than the more usual scalar time series.

FURTHER READING

Semipopular:

\textsuperscript{19}Abbreviated ECG. Nonmedical usage sometimes uses the older “EKG,” derived from a German word with the same meaning.
Figure 8.8: Electrocardiogram. The total current dipole vector moves periodically, rotating and stretching with each heartbeat.

EEG: www.youtube.com/watch?v=zG4tMchDTV8&list=PL8sA-GV63nB6AYDS1xLQA8bxom3ugwCMD.

Intermediate:

Technical:
Malmivuo & Plonsey, 1995; Gratiy et al., 2017.
8.5.2 More on electrolysis of water
The main text described one result of passing electricity through salt solution: the release of chlorine gas. Another option is for chloride to remain an ion and instead to assist in pulling an electron away from a water molecule, splitting it and ultimately creating $\text{H}^+$ ions and neutral oxygen gas.

At the cathode, sodium is too reactive with water to electroplate out. Instead, $\text{Na}^+$ ions assist in adding the excess electrons to water molecules, splitting them and ultimately creating $\text{OH}^-$ ions and neutral hydrogen gas.

The overall effect on the solution is that near the anode, some hydrochloric acid forms (H$^+$Cl$^-$), whereas near the cathode, some lye forms (Na$^+$OH$^-$). The effect is readily demonstrated with an acid/base indicator, especially if we use some gelatine to slow down diffusion in the solution.

8.6 Correction to the ohmic relation on microscopic scales
The ohmic relation, Equation 8.7 (page 119), attributes charge flux to driving electric force via a dissipative transport coefficient, the conductivity. However, in general a flux can also arise from a density gradient, leading to another contribution to this formula. Over macroscopic scales, we have seen that matter is effectively neutral; the density of charge carriers is therefore constant and we may neglect this contribution. However, Chapter 10 will show that acknowledging it leads to phenomena such as screening, which arises when net charge builds up in bulk over microscopic length scales (violating Idea 8.10).
8.1 Reactance
A real capacitor’s dielectric may not be a perfect insulator: Some current may “leak” across when a potential difference is applied. Here’s a way to measure both the capacitance $C$ and resistance $R$ at once, by applying a time-varying current $I(t)$ and observing the resulting membrane potential jump $\psi(t)$.

a. Write an expression for the total current into a membrane in terms of $\psi(t)$. The total current consists of the leakage plus the time change of the charge stored in the membrane’s capacitance.

b. Suppose that we impose a known current $I = I \cos(\omega t)$. Find the resulting $\psi(t)$, and show that it has both $\cos(\omega t)$ and $\sin(\omega t)$ terms; that is, it’s not in phase with the current. Show how to deduce $R$ and $C$ from this measurement.

8.2 Bulk conductor, $I$
Consider two macroscopic electrodes immersed in an infinite, 3D bath of a conducting fluid, such as salt water. The electrodes are insulated except for their ends, which are small metal spheres of radius $R_0$. The conductor obeys an ohmic relation, and the zero-frequency (DC) bulk conductivity of the medium is a constant, $\kappa$. The ends are separated by a distance $L \gg L_0$. Find the total DC resistance between the two electrodes as a function of $L$ and comment on the (possibly surprising) form of your answer.

[Hints: (a) Start by noticing that the units of conductivity are not the same as those of $1/(\text{resistance})$. Think about the possible forms (allowed by dimensional analysis) of the desired formula for resistance as a function of $x, L$, and $L_0$, in the stated limit. (b) Next, begin the problem by writing a trial solution for the electrostatic potential in the medium that solves the relevant equations and is approximately constant over each electrode in the stated limit. (c) The instruction “find the resistance” implies that no net current enters from, or exits to, infinity. From the potential, you can find the charge flux everywhere, as well as the total potential drop.]

8.3 Electrosurgery
Patients undergoing electrosurgery sometimes suffer burns. The damage is confined to the perimeter of the electrode, even though the entire electrode (a flat disk) is in contact with tissue. By reflection symmetry, we can equivalently consider a thin circular metal disk electrode of radius $a$ and potential $\phi_0$ completely surrounded by a bulk medium of conductivity $\kappa$. The circuit is completed by another electrode at some distant place; for example, you could imagine it as a spherical shell at potential 0, centered on the disk’s center, and of infinite radius.

The goal of this problem is to find the perpendicular component of charge flux at the surface of the electrode, $j_{\perp}$, and how it depends on position on the conductor.

We will model the electrode as an ellipsoid, that is, a solid with $xz$ and $yz$ cross-sections that are ellipses, and $xy$ cross-section that is a circle. Later, we’ll take the “pancake” limit where the axes of the ellipses along $z$ are much smaller than the ones along $x$ or $y$.
To define the ellipsoid, let $\sigma$ be some positive constant (the distance from the center to one focus of an ellipse). Set up cylindrical coordinates $\rho, \varphi, z$ centered on the center, with $z$ the axis of symmetry. Now define the distances

$$r_\pm = \sqrt{z^2 + (\rho \mp \sigma)^2}$$

and

$$\xi = (r_+ + r_-)/(2\sigma), \quad \eta = (r_+ - r_-)/(2\sigma).$$

The surfaces of constant $\xi$ are a family of nested ellipsoids. Our goal is to find the potential outside a conductor whose surface is one of these ellipsoids (at some $\xi_0$), given that the potential drop between the surface and infinity is $\psi_0$. Then the case with $\xi_0$ just slightly greater than 1 will correspond to a thin disk. Specifically, each elliptical cross-section has semimajor axis $\sigma \xi_0$ and semiminor axis $\sigma \sqrt{\xi_0^2 - 1}$.

The electric potential obeys

$$\nabla^2 \psi = 0 \text{ for } \xi > \xi_0, \quad \psi \to 0 \text{ at infinity, and } \psi(\xi_0, \eta, \varphi) = \psi_0.$$ 

Thus, the boundary conditions are simple when written in terms of oblate ellipsoidal coordinates. Let’s show that the laplacian is separable in these coordinates, following the approach in the chapter.

a. Invert the preceding formulas to solve for $\rho$ and $z$ in terms of $\xi$ and $\eta$. [Hint: Express $\xi \eta$ and $(\xi^2 - 1)(1 - \eta^2)$ in terms of $\rho$ and $z$, then think.] The intersection of a surface of constant $\xi$ with the $xz$ plane is a curve; use a computer to draw a few such curves to confirm that your formulas behave as you expect. Superimpose a few curves of constant $\eta$ to see the coordinate grid created by $\xi$ and $\eta$.

From this point on, all work will be analytic (not numerical). First, promote everything to 3D by expressing $x, y, z$ in terms of $\xi, \eta, \varphi$.

b. Differentiate to find the vector $\vec{e}_\xi \equiv \partial \vec{r} / \partial \xi$, and similarly $\vec{e}_\eta$ and $\vec{e}_\varphi$. These three vectors have a convenient property similar to the one found in Section 5.3.2 (page 68) for plane polar coordinates—what is it?

c. Use your answers to (b), and the nice property you observed, to express the volume element $d^3r$ as $d\xi \, d\eta \, d\varphi$ times a function of $\xi, \eta, \varphi$.

d. Use (b,c) to express the integral $\int f \, d^3r \, \hat{\nabla} \cdot \vec{V} \, g$ in the coordinates $\xi, \eta, \varphi$. Here $f$ and $g$ are any two functions, both independent of the azimuthal angle $\varphi$ and vanishing at infinity.

e. Use integration by parts to work out the laplacian $\nabla^2 g$ in these coordinates, for the relevant special case where $g$ is independent of $\varphi$.

f. What aspect of your answer to (e) suggests that we seek exact solutions to our physics problem of the form $\psi = A(\xi)B(\eta)$? Substitute this trial solution into your formula in (e), to obtain two ordinary differential equations linked by a common constant.

---

20Equation 8.14 is not quite the same as the corresponding quantities introduced in Problem 5.1, because now we need a squashed (“oblate”) ellipsoid, not one that is stretched (“prolate”).

---

November 14, 2023; Contents  Index  Notation  Glossary
g. Fix that unknown constant by imposing the boundary condition at the surface. Then find a simple solution for the function $B$.

h. Now solve the other ODE for $A$. It’s not quite as simple, but at least it takes the form $dA/d\xi = f(\xi)$, and hence can be done just by evaluating an integral. Assemble your results into the complete $\psi$. Ensure that your answer has the required behavior at infinity.

i. Find the charge flux perpendicular to the electrode at its surface. It could potentially be a function of $\eta$ and $\varphi$ at fixed $\xi_0$. Then integrate to find the total current, and hence also the net resistance between the electrode and infinity.

j. Find the rate of heat production in the medium close to the disk. [Hint: Let dimensional analysis guide you: You want an answer with the units such as watts per cubic meter.] Comment on how it depends upon position (that is, on $\eta$).

[Note: One might have worried that the sharp edge of the disk could generate a singularity that gives a pathological answer, such as infinite current and hence zero resistance. Indeed, you’ll find large charge flux at the rim of the disk. But you’ll also find that the overall resistance is finite.]

8.4 Current dipole
Imagine a small current source (hearing-aid battery) with narrow wires sticking out. Everything is insulated except for the tips of the wires, which are separated by 5 cm. The whole thing is immersed in an infinite bath of isotropic conductor, for example salt water, and the current source supplies a steady total current $I = 1$ mA (Figure 8.4, page 122).

a. What equation governs the steady electric potential throughout the salt water?

b. Write down a solution to that equation appropriate to the problem by superposing two simpler solutions.

c. The conductivity of seawater is $\kappa \approx 0.1 \, \text{S} \cdot \text{m}^{-1}$. Use that fact, and the form of your answer to (b) up close to one electrode tip, to get the overall constant that multiplies your solution, and hence finish explicitly evaluating the steady electric potential throughout the seawater.

8.5 Bulk conductor, revisited
A stationary (time-independent) current distribution is established in a medium that is isotropic but not necessarily homogeneous, for example, body tissue.

Specifically, the charge flux $\vec{j}$ is everywhere proportional to $-\vec{V}\psi$, but the constant of proportionality (the conductivity $\kappa$) may not be spatially uniform. However, you may assume that the dielectric constant $\epsilon/\epsilon_0$ is uniform and isotropic.

a. Show that the medium will in general acquire a nonzero free electric charge density $\rho_f(\vec{r})$. Show that this charge density may be written as the dot product of $\vec{V}\psi$ with a certain vector field, and find that vector field.

b. Repeat for the case where $\epsilon$ is also nonuniform, though still isotropic.

8.6
[[Not ready]]
CHAPTER 9

Vista: Cell Membrane Capacitance

Wonderful as are the laws and phenomena of electricity when made evident to us in inorganic or dead matter, their interest can bear scarcely any comparison with that which attaches to the same force when connected with the nervous system and with life.

— Michael Faraday

9.1 FRAMING: NONINVASIVE MEASUREMENT

Every living cell needs a wrapper to maintain a distinct interior environment. Section 6.9 mentioned that this bilayer membrane is just a few nanometers thick, which is why nobody could see it prior to the invention of the electron microscope. Nevertheless, H. Fricke “saw” it (that is, deduced its existence and thickness) in 1925.

Actually, a molecular-scale membrane had been hypothesized prior to this. There was some precedent. Benjamin Franklin had long ago done measurements on the spreading of oil on an air-water interface. Rayleigh made these more systematic: Oil could be spread to a layer just a few nanometers thick without holes appearing, but no further.\(^1\) Rayleigh was brave enough to propose the interpretation that this layer was exactly one molecule thick, at a time when the reality of molecules themselves was still controversial. Later, others realized that, even without an air-water interface, a double layer of such molecules could form, stably separating one aqueous compartment from another one. Could that be the physical object surrounding living cells? Fricke sought to confirm this hypothesis by characterizing the membranes of living cells, using an ingenious and noninvasive measurement technique.

Knowing that the lipid molecules constituting the cell’s bilayer membrane are similar to other oils let Fricke predict that the capacitance per unit area would be \(C = \varepsilon / \delta\), where \(\varepsilon\) is the permittivity of oil and \(\delta\) is the membrane thickness, at that time unknown.\(^2\) Thus, knowing the permittivity \(\varepsilon\) and measuring \(C\) would allow a determination of \(\delta\).

Although Fricke used blood cells, his result was especially significant in the context of neurons. Microscopy showed that they form a complex network. But debate raged about what happened at their junctions: Were they really separate cells, each enclosed in a distinct bag? Or was each junction a passageway, joining two cells’ interiors? By establishing the nanometer scale of membrane thickness, Fricke confirmed that the cell

\(^{1}\)Strutt, 1890.

\(^{2}\)See Equation 6.13 (page 81).
membrane was too thin to be seen via optical microscopy. So the fact that it had not been seen was not surprising, and certainly didn’t imply that it was absent.

Chapter 12 will build on these insights to make a fully quantitative theory of nerve impulses; the numerical value of membrane capacitance, established in this chapter, will play a big role in that theory.

Electromagnetic phenomenon: The capacitance of an object can be measured without placing electrodes on either side of it.

Physical idea: When the object is immersed in a current-carrying fluid, it will polarize, with measurable effects on its exterior.

9.2 FRICKE’S EXPERIMENT

9.2.1 Setup and solution

Naïvely, one could imagine stretching a bilayer membrane all the way across a chamber, imposing a potential drop across it, and measuring how much charge flowed while establishing that drop (“charging the capacitor”). Today such an approach may be possible via patch-clamp measurements, but not in 1925. To get there 60 years ahead of when the measurement “ought” to have been possible, Fricke found a more indirect approach.

Rather than having electrodes on either side of a membrane (Figure 9.1a), Fricke’s experiment involved suspending many cells in salt water and passing alternating current through the chamber. The frequency of the current was around 100 kHz, so we may use the quasi-static approximation for our analysis.³

We idealize the system as salt water on either side of an insulating spherical shell of radius $a$. (Later we will acknowledge that there are many cells, but they will be well separated in space.)

³Section 8.6 (page 121).
In a conducting medium, \( \mathbf{j} = \kappa \mathbf{E} \). Because we assume that no current may cross the membrane, we must have \( \mathbf{j} \perp = 0 \) at the inner and outer surfaces, and hence \( \mathbf{E} \perp = 0 \) also. The system arranges this by having thin layers of net charge pile up just outside the membrane as shown in Figure 9.1b. Elsewhere, there is no net charge, so \( \nabla \cdot \mathbf{E} = 0 \). Thus, we may write \( \mathbf{E} = -\nabla \psi \) as usual, but with a jump in \( \psi \) as we cross the membrane, due to the charge layers.

It may seem that we have another chicken/egg problem: We need the charge layers if we are to find the field, and vice versa. But we know by now that often such knots can be untangled by treating them as boundary-value problems. Indeed, we can regard ours as two decoupled quasistatic problems:

**Inside the cell**

\[ \nabla^2 \psi_{\text{in}} = 0, \text{ subject to } \mathbf{j} \perp = 0 \text{ on the boundary; that is,} \]

\[ \frac{\partial \psi_{\text{in}}}{\partial r} = 0 \text{ on the spherical surface } r = a. \]

That boundary condition is spherically symmetric, but there is only one spherically symmetric solution to the Laplace equation that is nonsingular at the origin: \( \psi_{\text{in}} = \text{const.} \) We will take the center to be our zero point of potential, so the constant is zero.

**Outside the cell**

\[ \nabla^2 \psi_{\text{out}} = 0, \text{ subject to } \mathbf{j} \perp = 0 \text{ on the boundary; that is,} \]

\[ \frac{\partial \psi_{\text{out}}}{\partial r} = 0 \text{ on the spherical surface } r = a, \text{ and} \]

\[ \psi_{\text{out}} \rightarrow -E_{\infty} z = -E_{\infty} r \cos \theta \text{ far away.} \]

The second condition is the requirement that \( -\nabla \psi \) should approach the imposed uniform electric field \( E_{\infty} \). We can easily guess one solution to the Laplace equation with the required behavior at infinity, that is,

\[ -E_{\infty} r \cos \theta \]

(9.2)

Itself. That solution doesn’t satisfy the boundary condition at the sphere, but we may add to it any other solution that vanishes at infinity, because such a modification won’t spoil the distant behavior. Indeed, we know many such solutions from the multipole expansion. Of these, however, only the dipole \( r^{-2} \cos \theta \) has the same angular dependence as Equation 9.2, and so is a candidate to help us satisfy the boundary condition at \( r = a \). (It doesn’t matter that this function is singular at \( r = 0 \), because we are only applying it in the exterior region.)

Imposing the boundary condition Equation 9.1 lets us find the unknown constant \( A \)

---

4See Equation 8.7 (page 119).
5Idea 8.10 (page 121).
9.2 Fricke’s Experiment

that multiplies the second solution:

\[
0 = \frac{\partial}{\partial r} \left[ -E_\infty r \cos \theta + Ar^{-2} \cos \theta \right] \\
0 = -E_\infty - 2Au^{-3} \\
\psi_{\text{out}} = -E_\infty (r + \frac{1}{2}a^2) \cos \theta + \text{const.} 
\]

(9.3)

Match the solutions

By symmetry, no charge piles up near the membrane at the equator, \( \theta = \pi/2 \). Hence, \( \psi \) must not jump as we cross the membrane there. We already found that \( \psi_{\text{in}} \) is zero throughout the interior, so

\[
\psi_{\text{out}}(r = a, \theta = \pi/2) = 0.
\]

Thus, the final constant in Equation 9.3 is zero.

9.2.2 The membrane stores electrostatic energy despite not being “in series” with the applied potential

We solved the electrostatic problem, but we still must connect to what was measured, and ultimately use the measurement to find the desired quantity: the capacitance per area \( C \) of cell membrane.

First, notice that the potential jump across the membrane is \( \Delta_{\text{memb}} \psi(\theta) = \psi_{\text{out}}(r = a, \theta) = -E_\infty \frac{3a}{2} \cos \theta \). Each surface area element is therefore a capacitor charged to that potential, and hence stores energy

\[
dE_{\text{memb}} = \frac{1}{2} (\Delta_{\text{memb}} \psi)^2 dC \text{ where } dC = C d\Sigma.
\]

We can now find the total stored energy by using Equation 9.3:

\[
E_{\text{memb}} = \int dE_{\text{memb}} = \int (a^2 d(\cos \theta) d\varphi) \frac{1}{2} (\Delta_{\text{memb}} \psi)^2 C \\
= \frac{1}{2} C \int (2\pi a^2 d(\cos \theta)) (E_\infty \cos \theta (a + \frac{1}{2}a))^2 = \frac{1}{2} C (E_\infty)^2 2\pi a^4 (3/2)^2 \int_{-1}^{1} d\mu \mu^2 \\
= \frac{3\pi}{2} a^4 E_\infty^2 C.
\]

(9.4)

For \( N \) well-separated cells in suspension, the total is \( N \) times this formula.

Fricke imposed a time-dependent potential drop \( \ddot{\psi} \cos(\omega t) \) across a chamber of length \( h \), so \( E_\infty(t) = -\ddot{\psi} \cos \omega t \). He then measured the resulting current. The current had the same angular frequency \( \omega \), so its form was \( I \cos(\omega t - \phi) \); Fricke therefore measured the dependence of peak current \( I \) and its phase shift \( \phi \) on \( \ddot{\psi} \) and \( \omega \) at fixed, known values of \( N \) and \( a \). We wish to see what our solution to the electrostatic problem predicts about this relationship, with the goal of extracting the numerical value of the only unknown parameter: the areal density of membrane capacitance, \( C \).

---

6Section 5.4 (page 70) found this solution in a different context (no charge flow).

7Problem 8.1 outlines Fricke’s strategy.

---

A suspension of insulating objects in conducting solution alters the phase relation between alternating current and potential.
9.2.3 The experimentally measured phase lag determines the capacitance

Each time an electron enters one end of the chamber, another exits the other end, with a net energy cost of $e\Phi \cos \omega t$. Thus, the net electric power entering the experimental chamber is

$$\mathcal{P} = \Phi \cos \omega t \mathcal{I} \cos(\omega t - \phi)$$

$$= \Phi \mathcal{I}(\cos^2 \omega t \cos \phi + \cos \omega t \sin \omega t \sin \phi). \quad (9.5)$$

The first term is always nonnegative. It represents ohmic (resistive) dissipation of energy into heat. The second term averages to zero. This indicates an “elastic” element, constantly storing energy and giving it back. The storage mechanism is the charging and discharging of the membrane capacitance, so this term must equal the time derivative of Equation 9.4:

$$-\frac{d}{dt}(N\epsilon_{\text{memb}}) = -N\frac{3\pi}{2}a^4 \epsilon_0 \left(\frac{\Phi}{L}\right)^2 \frac{d}{dt} \cos^2 \omega t,$$

where $h$ is the length of the chamber (distance between electrodes).

$$= N3h^{-2}\pi a^{4}\epsilon_{0}\Phi^{2} \cos \omega t \sin \omega t.$$

Compare that result to the second term of Equation 9.5 to find

$$\Phi \mathcal{I} \sin \phi = N3\pi a^{4} \epsilon_{0} \left(\frac{\Phi}{h}\right)^2.$$

Rearranging gives the desired result

$$\epsilon = \frac{I}{\Phi} \frac{h^2 \sin \phi}{N3\pi a^{4}\omega}. \quad (9.6)$$

The formula gives us membrane’s capacitance per area in terms of the known cell radius $a$ and count $N$, the chamber dimension $h$, the imposed $\Phi$ and $\omega$, and the resulting $I$ and $\phi$.

Fricke found $\epsilon \approx 1 \mu F/cm^2$. The permittivity of oil is around $3\epsilon_0$, so he inferred a membrane thickness value $\delta \approx 3\text{nm}$, within a factor of two of today’s accepted value. Remarkably, that value is also similar to the one implied by measurements made by Benjamin Franklin in 1773!

FURTHER READING

Semipopular:
On Franklin’s observations and more: Tanford, 1989.

Intermediate:
Sohn et al., 2000.

---

8The following derivation is easier with complex exponential notation (Section 18.7, page 278).
9See Problem 9.1.

Technical:

9.1 Measure cell membrane capacitance
In this problem you’ll find an experimentally practical way to measure the capacitance of a cell membrane.

Electrically speaking, a sea urchin egg is a thin spherical shell of insulator (the cell’s bounding membrane), surrounded by a medium-good conductor (sea water), and enclosing a medium-good conductor (the cell’s interior). The apparatus consists of a suspension of such eggs in a chamber, which is a rectangular prism. Plates at either end set up a potential drop from one end of the chamber to the other, that is, from \( z = 0 \) to \( z = h \).

The goal of the experiment was to measure \( C \), the membrane capacitance per unit area. But when designing the experiment, we often turn things around and use an estimate for \( C \) in order to predict whether the observed phase lag \( \phi \) between voltage and current will be large enough to measure (for example, with an oscilloscope). Thus, assume \( C \approx 1 \mu F \, cm^{-2} \).

Here are some other typical numbers:

- K. Cole followed up Fricke’s work, replacing the blood cells by egg cells, with radius \( a \approx 3 \times 10^{-4} \) cm.
- The applied current had a frequency of 87,000 Hz, or angular frequency \( \omega = 2\pi \times 87,000 \) rad/s.
- The overall resistance of the seawater in the chamber was \( \dot{\psi}/\dot{I} \approx 300 \) Ω.
- The number density of cells in the chamber was such that they occupied about 20% of the chamber volume.
- The chamber dimensions were: cross-section \( \Sigma \approx 15 \) cm², length \( h \approx 7 \) cm.

Use these numbers and the analysis in the chapter to find the predicted phase lag angle \( \phi \) in radians. (Make sure the units work out properly.) Does it seem likely to be measurable?

9.2 Fricke 2
Use a computer to visualize the electrostatic potential outside a cell, modeled as a spherical shell of insulator in conducting solution, with an applied \( \vec{E} \) field at infinity that is uniform along \( \hat{z} \):

a. Make a contour plot of \( \psi(x, 0, z) \). Describe in words the relevant physical aspects of the solution.

b. Then show the same function as a surface plot showing \( \psi(x, 0, z) \) as height above or below the xz plane.

c. Finally, make a vector-field plot of the corresponding \( \vec{E} \) field.
CHAPTER 10

Statistical Electrostatics of Solutions

10.1 FRAMING: ION CLOUDS

Section 2.1 mentioned that it is often important to find condensed (implicit) descriptions of some of the actors in a complex system. Thus, we would like to follow mobile charges explicitly but not have to think about everything else. One example of this approach was our introduction of a modified permittivity to account for a dielectric medium. This chapter introduces another example, where we account for the incessant thermal bumping of uncharged actors, for example, water molecules, against the charges of interest via a Boltzmann distribution.

Electromagnetic phenomenon: DNA falls apart into separate strands in pure water.

Physical idea: Excess salt in solution screens electrostatic interactions via the formation of neutralizing ion clouds.

10.2 SOLUTION IS DIFFERENT FROM VACUUM

10.2.1 The Nernst relation sets the scale of membrane potentials

Many of the molecules floating in water carry a net electric charge, unlike the water molecules themselves. When table salt dissolves, for example, the individual sodium and chlorine atoms separate, but the chlorine atom grabs one extra electron from sodium, thereby becoming a negatively charged chloride ion, \( \text{Cl}^- \), and leaving the sodium as a positive ion, \( \text{Na}^+ \). Any electric field present in the solution will then exert forces on the individual ions, dragging them just as gravity drags colloidal particles toward the bottom of a test tube. But colloidal particles do not fall all the way to the bottom of a chamber. Let’s recall why not, in an electrical context.

Suppose that we begin with a uniform-density solution of mobile, charged particles, each of charge \( q \), in a region with electric field \( \vec{E} \). For example, we could place two parallel, flat plates just outside the solution’s container, a distance \( h \) apart, and connect them to a battery that maintains a fixed electric potential difference across them, \( \Delta \psi = \psi_{\text{bot}} - \psi_{\text{top}} < 0 \). Even in solution, Equation 2.2 (page 28) still implies that \( E = -\Delta \psi / h \), and each charged particle still feels an electric force \( q\vec{E} \). Initially, then, each charged particle drifts with net speed \( v_{\text{drift}} = qE/\eta \), where \( \eta \) is a constant describing viscous friction. In salt solution

---

1Negative ions are also called anions, because they’d be attracted to an anode; similarly, positive ions are generically called cations. The terms “cathode,” “anode,” “ion,” “cation,” “anion,” “electrode,” and “electrolyte” were all coined by Michael Faraday.
Figure 10.1: [Sketch.] Origin of the Nernst relation (Equation 10.3). An electric field drives positively charged ions downward. Initially after connecting the battery, the number flux $j_{\text{ion}}$ for the ion species shown points downward with magnitude equal to the number density $c_{\text{ion}}$ times $v_{\text{drift}}$. The corresponding contribution to charge flux is $qj_{\text{ion}}$. Eventually the system comes to equilibrium with a downward density gradient of positive ions (and an upward gradient of negative ions, not shown.)

there are two ionic species with opposite charge, and hence opposite drift velocities, but for now we only consider one of the species.

Imagine observing a small surface element of area $d\Sigma$ stretched out perpendicular to the electric field (that is, parallel to the plates; see Figure 10.1). To find the flux of ions induced by the field, we ask how many ions pass this surface each second. The average ion drifts a distance $v_{\text{drift}}dt$ in time $dt$, so, in this time, all the ions contained in a slab of volume $v_{\text{drift}}dt d\Sigma$ pass the surface. The number we seek equals this volume times the ion density $c_{\text{ion}}$. The number flux in the $x$ direction is then the number crossing per area per time, or $c_{\text{ion}}v_{\text{drift}}$. (Check to confirm that this formula has the proper units.) Substituting the drift velocity gives

$\nabla \cdot j_{\text{ion}} = qE_{\text{ion}}/\eta$, the electrophoretic flux of the ion species we are considering.

Now suppose that the density of ions is not uniform. For this case, we add the driven (electrophoretic) flux just found to the diffusive flux (from “Fick’s law”), obtaining

$\nabla \cdot j_{\text{ion},x} = \frac{qE_{x}(x)c_{\text{ion}}(x)}{\eta} - D \frac{dc_{\text{ion}}}{dx},$

where $D$ is the diffusion constant for the ion species in question. The minus sign in this formula is important: It says that if concentration is increasing at larger $x$, then to erase that gradient diffusion will send a net flux of ions in the opposite direction.

We next rewrite the viscous friction coefficient in terms of $D$, using the Einstein relation$^2 \eta D = k_B T$ to get$^3$

$\nabla \cdot j_{\text{ion},x} = D(-\frac{dc_{\text{ion}}}{dx} + \frac{q}{k_B T}E_{x}c_{\text{ion}}).$  \textbf{Nernst–Planck formula} (10.1)

The Nernst–Planck formula helps us to answer a fundamental question: What electric field would be needed to get zero net flux, that is, to cancel the diffusive tendency to erase

---

$^2$We encountered this relation in the Example on page 120.

$^3$Recall Section 8.5.2. More generally, in non-planar geometry the Nernst–Planck formula becomes $\nabla \cdot j_{\text{ion}} = D(-\nabla c_{\text{ion}} + (q/k_B T)E c_{\text{ion}}).$
nonuniformity? To find out, set $j_{\text{ion}} = 0$ in Equation 10.1. In a planar geometry, where everything is constant in the $y, z$ directions, we get the condition

$$\frac{1}{c_{\text{ion}}} \frac{dc_{\text{ion}}}{dx} = \frac{q}{k_B T} \vec{E} \cdot \hat{x}. \quad \text{(thermal equilibrium)} \quad (10.2)$$

The left side of this formula can be rewritten as $\frac{d}{dx} \ln c_{\text{ion}}$.

To use Equation 10.2, integrate both sides from the top plate to the bottom one. The left side is $\int_0^h dx \frac{d}{dx} \ln c_{\text{ion}} = \ln(c_{\text{bot}}/c_{\text{top}})$, that is, the difference in $\ln c_{\text{ion}}$ from one plate to the other. On the right side, recall that $\Delta \psi = -\vec{E} \cdot \hat{x} h$. Thus, the condition for thermal equilibrium is

$$\Delta (\ln c_{\text{ion}}) = -\frac{q (\Delta \psi_{\text{eq}})}{k_B T}. \quad \text{Nernst relation} \quad (10.3)$$

The jump in electric potential across a cell membrane is usually abbreviated membrane potential (or “transmembrane potential”). The subscript “eq” reminds us that this is the potential change needed to maintain a concentration gradient in equilibrium.

The minus sign in Equation 10.3 says that positive ions will migrate toward larger $x$ (downward in Figure 10.1). It makes sense: They’re attracted to the negative plate. We have so far been ignoring the corresponding negative charges (for example, the chloride ions in table salt), but the same formula applies to them as well. Because they carry negative charge, Equation 10.3 says they migrate toward the positive plate.

Substituting some numbers into Equation 10.3 yields a suggestive result. Consider a singly charged ion species like $\text{Na}^+$, for which $q = e$. Suppose that we have a moderately big concentration jump, $c_{\text{bot}}/c_{\text{top}} = 10$. At room temperature, $T_r = 298 K$, we have $(k_B T_r/e) \approx 1.40 \cdot 10^{-3}$. Hence we find $\Delta \psi \approx +58 \text{ mV}$. What’s suggestive about this result is that many living cells, particularly nerve and muscle cells, really do maintain a potential difference across their membranes of a few tens of millivolts! We haven’t proven that these are equilibrium (Nernst) potentials, and indeed they’re not. But the observation does show that dimensional arguments successfully predict the scale of membrane potentials with almost no hard work at all.

Something interesting happened on the way from Equation 10.1 to Equation 10.3: When we consider equilibrium only, the value of the diffusion constant drops out. That’s reasonable: $D$ controls the rate of response to a field; its units involve time. But equilibrium is an eternal state; it can’t depend on time. In fact, exponentiating the Nernst relation gives that $c_{\text{ion}}(x)$ is a constant times $e^{-q \psi(x)/k_B T}$. This result is an old friend: It says that the equilibrium distribution of ions follows the Boltzmann distribution. A charge $q$ in an electric field has electrostatic potential energy $q \psi(x)$ at $x$; its probability to be there is proportional to the exponential of minus its energy, measured in units of the thermal energy $k_B T$. A positive charge doesn’t like to be in a region of large positive potential, and vice versa for negative charges. Our formulas are mutually consistent.

---

*Normally it is meaningless to speak of a nonlinear function like log applied to a quantity with units. However, a difference of two such logs can be written as the log of the dimensionless ratio, so we always get the same result regardless of what units we choose.*
10.2.2 The electrical conductivity of a solution reflects frictional dissipation

Suppose that we place the metal plates in Figure 10.1 inside the container of salt water, so that they become electrodes. Then the ions in solution migrate, but they don’t accumulate: The positive ones get electrons from the − electrode, whereas the negative ones hand their excess electrons over to the + electrode. The resulting neutral atoms leave the solution; for example, they can electroplate onto the attracting electrode or bubble away as gas. Then, instead of establishing equilibrium, our system continuously conducts electricity, at a rate controlled by the steady-state ion fluxes.

According to the Nernst–Planck formula (Equation 10.1), this time with uniform $c_{\text{ion}}$, the electric field is $E = (k_B T/(D q c_{\text{ion}})) j_{\text{ion}}$. Recall that number flux and charge flux are related by $j_{\text{ion}} = j/q$. Comparing with Equation 8.7 (page 119) then shows that our solution is ohmic with conductivity

$$\kappa = \frac{D q^2 c_{\text{ion}}}{k_B T}.$$ (10.4)

Indeed, saltier water conducts better. To use Equation 10.4, remember that each ionic species contributes to the total current; for table salt, we need to add separately the contributions from Na\(^+\) with $q = e$ and Cl\(^-\) with $q = -e$. Because all small ions have similar diffusion constants, the effect is to approximately double the right-hand side of the formula.

The resistance of the solution depends not only on its chemical makeup but also on the geometry of the chamber, via Equation 8.8 (page 119):

$$\Delta \psi = IR \quad \text{where} \quad R = h/\langle \Sigma \kappa \rangle.$$ [8.8, page 119]

Section 10.2.2 (page 156) mentions other points about electrical conduction.

10.3 A REPULSIVE INTERLUDE

10.3.1 Electrostatic interactions are crucial for proper functioning of living cells

Section 6.7 (page 87) pointed out that when we put a macromolecule such as DNA in water, some of its loosely attached cations wander away, leaving some of their electrons behind. The remaining macromolecule then has a net negative charge: DNA becomes a negative macroion. The lost ions are called counterions, because their net charge counters (neutralizes) the macroion.

The counterions diffuse away because they were not bound by chemical (covalent) bonds in the first place and because by diffusing away, they increase their entropy. But having left the macroion, the counterions now face a dilemma. If they stay too close to home, they won’t gain much entropy. But to travel far from home requires lots of energy, to pull away from the opposite charges left behind on the macroion. The counterions thus need to make a compromise between the competing imperatives to minimize energy and

---

\(^5\)See Section 8.5.2 (page 119).
maximize entropy. The following sections will show that for a large flat macroion, the compromise chosen by the counterions is to remain hanging in a **counterion cloud** near the macroion’s surface. After working Your Turn 6E (page 87), you won’t be surprised to find that the cloud can be a couple of nanometers thick. Viewed from beyond the counterion cloud, the macroion appears neutral. Thus, a second approaching macroion won’t feel any attraction or repulsion until it gets closer than about twice the cloud’s thickness. This behavior is quite different from the behavior of charges in a vacuum: In that case, the electric field outside a flat, charged object doesn’t fall off with distance at all! In short,

*Electrostatic interactions are of long range in vacuum. But in solution, a screening effect reduces this interaction’s effective range, typically to a nanometer or less.* (10.5)

The counterion cloud is sometimes called the “diffuse charge layer.” Together with the charges left behind in the surface, it forms an **electric double layer** near a charged macroion. The forces on charged macroions have a mixed character: They are partly electrostatic and partly entropic. Certainly, if we could turn off thermal motion, the counterion cloud would collapse back onto the macroion, thereby leaving it neutral; we’ll see this in the formulas we ultimately obtain.

Before calculating properties of the counterion cloud in Section 10.3.3, this section will close with a few comments on broader biophysical implications.

**Electrostatic repulsion opposes macromolecular aggregation**

The cells in your body contain a variety of macromolecules. A number of attractive forces are constantly trying to stick the macromolecules together, for example, van der Waals forces. It wouldn’t be nice if they just acquiesced, clumping into a ball of sludge at the bottom of the cell, with the water on top. The same problem bedevils many industrial colloidal suspensions, for example, latex paint. One way Nature, and we its imitators, avoid this “clumping catastrophe” is to arrange for the colloidal particles to have the same sign of net charge. Indeed, most of the macromolecules in a cell are negatively charged and hence generically repel one another.

**Specific binding**

Idea 10.5 says that electrostatic forces are effectively of short range in solution, and moreover that this range is smaller than a typical macromolecule. That observation matters crucially for cells, because it means that two macromolecules will not feel one another until they’re nearby. Even when they are nearby, only immediately juxtaposed elements of their surfaces will “feel” each other. Thus, the detailed shape and surface pattern of positive and negative residues on a protein can be felt by its neighbor, not just the overall charge. This observation goes to the heart of how cells organize their myriad internal biochemical reactions (Figure 10.2). Although thousands of macromolecules may be wandering

---

6Similar methods can be applied to a long, thin line of charge, such as a DNA molecule; see Problem 10.2.

7See Section 3.7.3.
**Figure 10.2:** [Cartoon.] **One source of binding specificity.** The bottom pair of macromolecules have complementary shapes and charge distributions, creating multiple attractive contributions to their mutual electrostatic energy. The molecules at the top are overall attractive, but due to their shape mismatch and the short-range character of electrostatic interactions in solution, they are not as strongly bound. The molecules in the middle panel match in shape but not in the detailed distribution of charges.

around any particular location in the cell, typically only those with precisely matching shapes and charge distributions will bind together. One reason for this amazing specificity is that

> Even though each individual electrostatic interaction between matching charges is rather weak (relative to $k_B T$), still the combined effect of many such interactions can lead to strong binding of two molecules—**if** their shapes and patterns of charged groups match precisely. (10.6)

Nor is it enough for two matching surfaces to come together; they must also be properly oriented before they can bind. We say that macromolecular binding is **stereospecific**.

Thus, understanding molecular recognition, which is crucial for the operation of every cell process, requires that we first understand the counterion cloud around a charged surface, and hence establish Ideas 10.5–10.6.

**Energy of ATP**

It is sometimes said that the molecule ATP is suitable as an energy carrier because it contains “high energy bonds” that when broken “release their energy.” But that seems paradoxical: The formation of a bond always lowers energy (that’s what makes it a bond), so breaking a bond always **costs** energy.

We get some insight when we recall that the Born self-energy in pure water is proportional to charge squared (the Example on page 78); a similar result holds in salt solution.8

---

8See Problem 10.5.
positive layer

negative surface

Figure 10.3: [Schematic.] A **planar distribution of charges.** A thin sheet of negative charge (*hatched, bottom*) lies next to a neutralizing positive layer of free counterions (*shaded, top*). The individual counterions are not shown; the shading represents their average density. The lower box encloses a piece of the surface; so it contains total charge $-\sigma_q d \Sigma$, where $d \Sigma$ is its cross-sectional area and $-\sigma_q$ is the areal charge density. The upper box encloses charge $\rho_q(x) dx d \Sigma$, where $\rho_q(x)$ is the charge density of counterions. The electric field $E(x)$ at any point equals the electric force on any ion at that point, divided by the ion’s charge. For all positive $x$, the field points along the $-\hat{x}$ direction. The field at $x_1$ is weaker than that at $x_2$, because the repelling layer of positive charge between $x_1$ and $x=0$ is thicker than that between $x_2$ and $x=0$. Moreover, there is less positive charge between $x_1$ and infinity pushing a positive test charge downward than between $x_2$ and infinity.

So a small molecule with charge $-4e$ reduces its electrostatic energy when it splits into fragments with charges $-e$ and $-3e$, because $(-1)^2 + (-3)^2 < (-4)^2$. If that energy gain outweighs the net energy cost of rearranging chemical bonds, then there can indeed be a net release of energy upon hydrolysis.\(^9\)

The situation may remind you the energy release in nuclear fission: Here again, a short-range attractive interaction (the nuclear force) competes against the long-range electrostatic repulsion. If a uranium nucleus deforms enough to get past the resulting activation barrier, then it can greatly reduce its overall energy by separating completely into two well separated fragments, each with about half the original charge.\(^10\)

### Counterion cloud near a polarized membrane

Section 6.9 pointed out that a cell’s bilayer membrane acts as an insulator, preventing the free passage of ions into or out of the cell and hence allowing a sharp change in the electric potential from one side to the other. Positive ions then form a cloud just outside the cell, whereas negative ions are depleted there, and vice versa just inside, as claimed in Figures 8.5 (page 123) and 8.6.

### 10.3.2 The Gauss law

Let’s express the foregoing qualitative ideas mathematically. Figure 10.3 shows a thin, negatively charged sheet with uniform areal charge density $-\sigma_q$, next to a spread-out layer

---

\(^9\)A quantum-mechanical effect (resonance) also reduces the bond energies of the fragments more than it does the original ATP.

\(^10\)Recall Your Turn 6A (page 78).
of positive charge with volume charge density $\rho_q(x)$. Thus, $\sigma_q$ is a positive constant with units $\text{coul m}^{-2}$, whereas $\rho_q(x)$ is a positive function with units $\text{coul m}^{-3}$. Everything is constant in the $\hat{y}$ and $\hat{z}$ directions. We’ll simply write $E$ for the component of the electric field in the $\hat{x}$ direction.

The electric field above the negative sheet is a vector pointing along the $-\hat{x}$ direction, so the function $E(x)$ is everywhere negative. Just above the sheet, the electric field is proportional to the areal charge density: Applying the Gauss law for a flat, charged surface gives

$$E|_{\text{surface}} = -\sigma_q/\epsilon.$$  \hfill (10.7)

Away from the surface, the Gauss law gives (see Figure 10.3)

$$\frac{dE}{dx} = \frac{\rho_q}{\epsilon}.$$  \hfill (10.8)

The following section will use this relation to find the electric field everywhere outside the surface.

10.3.3 Detailed form of the counterion cloud outside a charged surface

The mean field

Now we can return to the problem of ions in solution. A typical problem might be to consider a thin, flat, negatively charged surface with areal charge density $-2\sigma_q$ and pure water on both sides. For example, cell membranes are negatively charged. You might want to coax DNA to enter a cell (say, for gene therapy). Because both DNA and cell membranes are negatively charged, you’d need to know how much they repel.

An equivalent, and slightly simpler, problem is that of a solid surface carrying charge density $-\sigma_q$, with water on just one side (Figure 10.4a). Also for simplicity, suppose that the loose positive counterions are monovalent (for example, sodium, $\text{Na}^+$). That is, each carries a single charge: $q_+ = e \approx 1.6 \cdot 10^{-19} \text{coul}$. A real cell has additional ions of both charges from the surrounding salt solution. The negatively charged ones are called coions, because they have the same charge as the surface. We will neglect the coions for now (see Section 10.3.4, page 156).

As soon as we try to find the electric field in the presence of mobile ions, an obstacle arises: We are not given the distribution of the ions, but instead must find it. Moreover, electric forces are of long range. The unknown distribution of ions will thus depend on each ion’s interactions not only with its nearest neighbors but also with many other ions! How can we hope to model such a complex system?

Let’s try to turn adversity to our advantage. Perhaps we can approach the problem by thinking of each ion as moving under the influence of an electric potential created by the average charge density of the others, or $\langle \rho_q \rangle$. We call this approximate electric potential $\psi(x)$ the mean field and this approach the mean-field approximation. The approach is reasonable if each ion feels many others; then the relative fluctuations in $\psi(x)$ about its
average will be small. To make the notation less cumbersome, we will drop the averaging signs; from now on, $\rho_q$ refers to the average density.

The Poisson–Boltzmann equation

We wish to find $c_+(x)$, the number density of counterions. We are supposing that our surface is immersed in pure water; hence, far away from the surface, $c_+ \to 0$. The electrostatic potential energy of a counterion at $x$ is $e\psi(x)$. We are treating the ions as moving independently of each other in a fixed potential $\psi(x)$, so the density of counterions, $c_+(x)$, is given by the Boltzmann distribution. Thus, $c_+(x) = c_0 e^{-e\psi(x)/k_B T}$, where the normalization $c_0$ is a unknown. We can add any constant we like to the potential because that change doesn’t affect the electric field $E = -d\psi/dx$. It’s convenient to choose the constant so that $\psi(0) = 0$. This choice gives $c_+(0) = c_0$; that is, $c_0$ is just the concentration of counterions at the surface.

Unfortunately, we don’t yet know $\psi(x)$. To find it, apply the Gauss law (Equation 10.8), taking $\rho_q$ equal to the number density of counterions times $e$. The potential obeys the Poisson equation: $d^2\psi/dx^2 = -\rho_q/e$. Given the charge density, we can solve the Poisson equation for the electric potential. The charge density, in turn, is given by the Boltzmann distribution as $e c_+(x) = e c_0 e^{-e\psi(x)/k_B T}$.

Despite the simplification of mean field approximation, we still seem to be facing a chicken-and-egg problem (Figure 10.5): We need the average charge density $\rho_q$ to solve the Poisson equation for the potential $\psi$. But we need $\psi$ to find $\rho_q$ from the Boltzmann distribution! Luckily, each of the arrows in Figure 10.5 represents an equation in two unknowns, namely, $\rho_q$ and $\psi$. We just need to solve these two equations simultaneously to find the two unknowns.

Before proceeding, let’s take a moment to tidy up our formulas. First, define the
Figure 10.5: [Diagram.] **Strategy to find the mean-field solution.** Neither the Poisson equation nor the Boltzmann distribution alone can determine the charge distribution, but solving these two equations in two unknowns simultaneously does the job.

Dimensionless rescaled potential $\tilde{\psi}$:

$$\tilde{\psi}(x) \equiv e\psi(x)/k_B T.$$  \hfill (10.9)

That change simplifies the exponential:

$$\frac{d^2 \tilde{\psi}}{dx^2} = -\frac{e^2 c_0}{k_B T \varepsilon} e^{-\psi}.$$  \hfill (10.10)

We can simplify still further by changing variables from $x$ to a dimensionless rescaled variable:

**Your Turn 10A**

Let $\tilde{x} = x/A$, where $A$ is a constant.

a. Confirm that the choice $A = \sqrt{\varepsilon k_B T/(e^2 c_0)}$ has dimensions of length.

b. Confirm that this choice simplifies our equation to the dimensionless form

$$\frac{d^2 \tilde{\psi}}{d\tilde{x}^2} = -e^{-\tilde{\psi}}.$$  \hfill (10.10)

The payoff for introducing the abbreviations $\tilde{\psi}$ and $\tilde{x}$ is that now Equation 10.10 is less cluttered, and we can verify at a glance that its dimensions work: Both sides are dimensionless.

**Solution of the Poisson–Boltzmann equation**

We could just ask a computer to solve our problem, but in this case we are lucky and can do it analytically. We need a function whose second derivative equals minus its exponential. We recall that the logarithm of a power of $\tilde{x}$ has the property that both its derivative and its

---

12 This formula seems to contradict our discussion of the quasistatic approximation (Idea 8.10, page 121), which was that the net charge density in bulk rapidly relaxes to zero. Section 8.6 (page 128) explained that that result is valid only on macroscopic length scales, because it neglects the diffusive current contribution in Equation 10.1 (page 140).
exponential are powers of $x$. We don’t want $\psi(\tilde{x}) = \ln \tilde{x}$, because that’s divergent (equal to minus infinity) at the surface. Nevertheless, a slight modification gives something promising:

$$\psi(\tilde{x}) = \beta \ln(1 + (\gamma \tilde{x})).$$

Let’s see whether we can find values for the constants $\beta$ and $\gamma$ for which Equation 10.11 solves the Poisson–Boltzmann equation.

### Boundary conditions

Like any differential equation, (10.10) doesn’t specify the solution completely. Instead, the equation has a family of solutions; we must choose the one that satisfies appropriate boundary conditions. We require:

(i) Our convention that $\psi(0) = 0$. The trial solution Equation 10.11 always has that feature, regardless of what values we choose for $\beta$ and $\gamma$.

(ii) Our expectation that there will be no electric field at infinity because no charge is located there: $d\psi/dx \to 0$. Our trial solution also automatically satisfies this condition.

We now check whether we can choose values for the constants $\beta$ and $\gamma$ in such a way that the trial solution also solves the Poisson–Boltzmann equation. Substituting $\psi(\tilde{x}) = \beta \ln(1 + (\gamma \tilde{x}))$ into Equation 10.10, we indeed find that it works if we take $\beta = 2$ and $\gamma = 1/\sqrt{2}$.

We have not yet introduced the areal charge density, so we are not yet done. The surface form of the Gauss law (Equation 10.7) gives $-d\psi/dx|_{\text{surface}} = -\sigma_q/\varepsilon$, or

$$\left.\frac{d\psi}{dx}\right|_{\text{surface}} = \frac{eA\sigma_q}{k_B T \varepsilon}.$$  \hspace{1cm} (10.12)

When using this formula, remember that $\sigma_q$ is a positive number; the surface has charge density $-\sigma_q$. The constant $A$ is the combination that you found in Your Turn 10A.

**Ex.** Check that the sign is correct in this formula.

**Solution:** The electrostatic potential $\psi$ gets more negative as we approach a negatively charged object. Thus, approaching counterions feel their potential energy $e\psi$ decrease as they approach the surface, so they’re attracted. If $x$ is the distance from a negatively charged surface, then $\psi$ will be decreasing as we approach it, or increasing as we leave: $d\psi/dx > 0$, so the sign is correct in Equation 10.12.

It may now seem as though we are in trouble: We have used up all the freedom in our family of trial solutions, and yet we still must impose Equation 10.12! To make progress, note that one of the constants entering $A$ was not given to us, namely $e_0$. We are given the areal charge density, but the system chooses the counterion concentration in a way fixed by Equation 10.12. Substituting the trial solution and the definition of $A$ yields

$$\frac{k_B T}{e} \left( \frac{e k_B T}{e^2 e_0} \right)^{-1/2} 2^{1/2} = \frac{\sigma_q}{\varepsilon}.$$
which we can solve for the unknown $c_0$.

**Your Turn 10B**

a. Show that $c_0 = \frac{\sigma_q^2}{(2\varepsilon k_BT)}$.
b. Hence show that in the original variables the electrostatic potential is

$$\psi(x) = \frac{k_BT}{\varepsilon} 2 \ln(1 + x/x_0),$$

(10.13)

where $x_0 = 2\varepsilon k_BT/(\varepsilon \sigma_q)$. Check the units.

Notice that *increasing the areal charge density makes the counterion cloud thinner* (reduces $x_0$), and raises the concentration at the surface.

**Your Turn 10C**

Find the equilibrium concentration profile $c_+(x)$ away from the surface. Check your answer by calculating the total areal density of counterions, $\int_0^\infty dx c_+(x)$, and verifying that the whole system is electrically neutral.

The solution you just found is sometimes called the **Gouy–Chapman layer**; $x_0$ is called the Gouy–Chapman length. This solution is appropriate in the neighborhood of a flat, charged surface in pure water.\(^{13}\) Let’s extract some physical conclusions from the math.

First, your answer to Your Turn 10C shows that indeed, a counterion cloud forms, with thickness roughly $x_0$. As argued physically in Section 10.3.1, the counterions are willing to pay some electrostatic potential energy (separating from their macroion) in order to gain entropy. More precisely, the counterions pull some thermal energy from their environment to make this payment. They can do this because doing so lowers the entropic part of their free energy more than it raises the electrostatic part. If we could turn off thermal motion (that is, send $T \rightarrow 0$), the energy term would dominate and the layer would collapse. We see this mathematically from the observation that then the layer thickness $x_0 \rightarrow 0$.

How much electrostatic energy must the counterions pay to dissociate from the surface? We can get an estimate by thinking of the layer as a planar sheet of charge hovering at a distance $x_0$ from the surface. When two sheets of charge are separated, we have a parallel-plate capacitor. Such a capacitor, with area $\Sigma$, stores electrostatic energy $\mathcal{E} = q_{\text{tot}}^2/(2C)$. Here $q_{\text{tot}}$ is the total charge separated; for our case, it’s $\sigma_q \Sigma$. The capacitance of a parallel-plate capacitor is given by $C = \varepsilon \Sigma / x_0$ (Equation 6.13, page 81). Combining the preceding formulas gives an estimate for the density of stored electrostatic energy per unit area for an isolated surface in pure water:

$$\mathcal{E}/(\text{area}) \approx k_BT(\sigma_q/\varepsilon). \quad \text{(electrostatic self-energy, no added salt)}$$

(10.14)

\(^{13}\) Or more realistically, a highly charged surface in a salt solution whose concentration is low enough; see Section 10.3.4\(^{\dagger}\) (page 156).
This estimate makes sense even for a realistic counterion cloud: The environment is willing to invest about $k_B T$ per counterion in electric field energy. This energy gets stored in forming the counterion cloud.

**Ex.** Is it a lot of energy?

*Solution:* A fully dissociating bilayer membrane can have one unit of charge per lipid head group, or roughly $|\sigma_q/e| = 0.7 \text{ nm}^{-2}$. A vesicle (spherical balloon of membrane) with radius $10 \mu m$ then carries stored free energy $\approx 4\pi (10 \mu m)^2 \times (0.7/\text{nm}^2) k_B T \approx 10^9 k_B T$. It’s a lot!

We’ll see how cells harness this stored energy in Section 10.4.

### 10.3.4 Excess salt shrinks the electric double layer

For simplicity, the preceding calculations assumed that a dissociating surface was immersed in pure water: All counterions come from the surface and there are no coions. In real cells, however, the cytosol is an electrolyte (salt solution). In this case, the density of counterions at infinity is not zero, so the counterions originally on the surface have less to gain entropically by moving away from it. We may then expect that the counterion cloud will hug the surface more tightly than it does in Equation 10.13. That is,

\[ \text{Increasing salt in the solution shrinks the counterion cloud.} \quad (10.15) \]

You’ll make this expectation quantitative in Problem 10.4.

Section 10.3.4′ (page 156) solves the Poisson–Boltzmann equation for a charged surface in a salt solution, arriving at the concept of the Debye screening length and making Equation 10.15 quantitative. It then considers more complex chemical reactions than dissociation, arriving at a model for how voltaic cells work.

### 10.3.5 The repulsion of like-charged surfaces arises from compression of their ion clouds

Returning to the case with pure water,\textsuperscript{14} we’re ready to find the force between two charged surfaces. Figure 10.4b shows the geometry. One might be tempted to say, “Obviously, two negatively charged surfaces will repel.” But wait: By symmetry, everything to the left of the central plane $x = 0$ (that is, the surface, together with its counterion cloud) is net electrically neutral, as is everything to the right. Thus, the electrostatic force that one side exerts on the other must equal zero! But electrostatic force is not the only kind of force in the problem. As the surfaces get closer than about twice their Gouy–Chapman length $x_0$, their diffuse counterion clouds begin to overlap, then get squeezed; they resist that confinement just as an ideal gas resists compression. Here are the details.

If we could turn off thermal motion, the mobile ions would collapse down to the surfaces, and there would be no net charge anywhere. That observation motivates us to

\textsuperscript{14} Even in the presence of salt, our result will be accurate if the surfaces are highly charged because in this case, the Gouy–Chapman length is less than the Debye screening length (again see Section 10.3.4′, page 156).
look at entropic forces. Examining Figure 10.4b, we see that charged particles are required to be in the gap, by charge neutrality. That is, the concentration of a dissolved ion species is higher in the gap than in the bulk. In such a situation, we expect an osmotic pressure in the gap, proportional to the concentration difference times the absolute temperature. This hydrostatic pressure is what physically pushes the two surfaces apart, not a literal electrostatic repulsion.

Unlike the case of a single surface, this time it’s convenient to measure distance from the midplane between the two surfaces, which are therefore located at $x = \pm h$ (Figure 10.4b). Each surface has areal charge density $-\sigma_q$. We again choose the constant in $\psi$ so that $\psi(0) = 0$, but now the parameter $c_0 = c_+(0)$ is the unknown concentration of counterions at the midplane. $\psi(x)$ will then be symmetrical about the midplane, so our previous trial solution (Equation 10.13) won’t work. Keeping the logarithm idea, though, this time we try $\hat{\psi}(x) = \beta \ln \cos(\gamma x)$, where $\beta$ and $\gamma$ are again unknown constants. Certainly this trial solution is symmetrical and equals zero at the midplane, where $\hat{x} = 0$.

The rest of the procedure is familiar. Substituting the trial solution into the Poisson–Boltzmann equation (Equation 10.10) again shows that it works with $\beta = 2$ and $\gamma = 1/\sqrt{2}$. The boundary condition at $x = -h$ is again Equation 10.12. Imposing the boundary conditions again fixes $c_0$: Making the convenient abbreviation $\xi = (c_0 e^2/(2\varepsilon k_B T))^{1/2}$ gives

$$\tan(h\xi) = \frac{1}{\xi} \frac{\sigma_q e}{2\varepsilon k_B T}. \tag{10.16}$$

Given the areal charge density $-\sigma_q$, we solve Equation 10.16 for $\xi$ as a function of the spacing $2h$; then the desired solution is

$$\hat{\psi}(x) = 2 \ln \cos(\xi x), \quad c_+(x) = c_0 (\cos \xi x)^{-2}. \tag{10.17}$$

As expected, the charge density is greatest near the plates, whereas the potential is maximum in the center.

By symmetry, the electric field at the midplane is zero, so an ion feels zero external force there. However, an ion that tries to diffuse out of the gap gets pulled back in. The osmotic pressure difference equals $k_B T$ times the difference between $c_0$ and the concentration outside the gap (which is zero), so the repulsive force per unit area on the surfaces is given approximately by the ideal gas law:

$$f/(\text{area}) = c_0 k_B T \quad \text{repulsion of like-charged surfaces, no added salt} \tag{10.18}$$

In this formula, $c_0 = 2\xi^2 e^2 k_B T/e^2$ and $(h, \sigma_q)$ is the solution of Equation 10.16. You can solve Equation 10.18 numerically (see Problem 10.1), but a graphical solution shows qualitatively that $\xi$ increases as the plate separation decreases (Figure 10.6). Thus, the repulsive pressure increases, too, as expected.

Your Turn 10D

Make a similar graphical argument to find qualitatively what happens to $\xi$ if we change the areal charge density, holding $h$ fixed.
Figure 10.6: [Mathematical functions.] **Graphical solution of Equation 10.16.** The sketch shows the dimensionless function $\sigma_q e/(2ek_BT\xi)$, as well as $\tan h\xi$ for two values of the plate separation $2h$. The value of $\xi$ at the intersection of the rising and falling curves gives the desired solution. The figure shows that smaller plate separation gives a larger solution $\xi_2$ than does large separation (yielding $\xi_1$). Larger $\xi$ in turn implies a larger ion concentration at the midplane and larger repulsive pressure.

Figure 10.7: [Experimental data with fits.] **The repulsive pressure between two positively charged surfaces in water.** The surfaces were egg lecithin bilayers containing 5 mole% or 10 mole% phosphatidylglycerol (*circles* and *stars*, respectively). The curves show one-parameter fits of these data to the numerical solution of Equations 10.16 and 10.18. The fit parameter is the areal charge density $\sigma_q$. The *dashed line* shows the solution with one proton charge per $24 \text{ nm}^2$; the *solid line* corresponds to a higher charge density (see Problem 10.1). At separations below 2 nm, the surfaces begin to touch and other forces besides the electrostatic one appear. The membrane with a larger density of charged lipids is found to have a larger effective charge density, as expected. [Data from Cowley et al., 1978; see Dataset 1.]

Note that the force in Equation 10.18 is not simply proportional to the absolute temperature, because $\xi$ has a complicated temperature dependence. This means that our pressure is not a purely entropic effect, but a mixed effect: The counterion layer reflects a balance between entropic and energetic imperatives. As remarked in Section 10.3.4, the qualitative effect of adding salt to the solution is to tip this balance away from entropy, thereby shrinking the counterion clouds on the surfaces and shortening the range of the interaction.

This theory works (see Figure 10.7). You'll make a detailed comparison with experiment in Problem 10.1.

Section 10.3.5′ (page 163) derives the electrostatic force directly as a derivative of the free energy.
Chapter 10  Statistical Electrostatics of Solutions

10.3.6 DNA denatures in pure water

We may summarize qualitatively by saying

*The distribution of co- and counterions outside a charged object adjusts to partially screen that object’s far fields.*

Indeed, far enough outside of a charged plane we found complete cancellation of electric field; you’ll explore a cylindrical object in Problem 10.2. This reduction of far fields implies a corresponding reduction of the Born self-energy of the object due to those fields.

DNA consists of two highly charged strands that hold together in a precarious balance, in which their mutual electrostatic repulsion is overridden by hydrogen bonds between their bases. Changing the concentration of surrounding excess salt alters that balance. In fact, when DNA is placed in pure water, repulsion gains the upper hand and the two strand separate (the DNA “denatures”). In normal physiological salt levels, the double helix is stable.

10.4 OPPOSITELY CHARGED SURFACES ATTRACT BY COUNTERION RELEASE

Now consider an encounter between surfaces of opposite charge (Figure 10.4c, page 147). Without working through the details, we can understand the attraction of such surfaces in solution qualitatively by using the ideas developed earlier. Again, as the surfaces approach each other from infinity, each presents a net charge density of zero to the other; there is no long-range electrostatic force, unlike the constant attractive force between two such planar surfaces in vacuum. Now, however, as the surfaces approach, they can shed counterion pairs without sacrificing the system’s neutrality. The released counterions leave the gap altogether and hence gain entropy, thereby lowering the free energy and driving the surfaces together. If the charge densities are equal and opposite, the process proceeds until the surfaces are in tight contact, with no counterions left at all. In this case, there is no separation of charge, and no counterions remain in the gap. Thus, all the self-energy estimated in Equation 10.14 gets released. The Example on page 151 showed that this energy is substantial: Electrostatic binding between macromolecular surfaces of matching shape can be very strong.

10.5 PLUS ULTRA

An “n-type semiconductor” is a material with a small concentration of mobile electrons (not involved in covalent bonding), for example, a crystal of silicon doped with a small impurity of antimony. A “p-type semiconductor” has a similarly small concentration of atoms with an electron bonding state that is normally unfilled, for example, silicon doped with indium. When a slab of n-type is placed in contact with a slab of p-type, some of the mobile electrons cross over to the other side and occupy the extra bonding orbitals, leaving
the n-side positively charged and the p-side negatively charged. This charge separation costs electrostatic energy, but nevertheless it happens at nonzero temperature; the details are mathematically similar to the counterion cloud studied here.

**FURTHER READING**

*Semipopular:*
Electroosmotic pump: www.youtube.com/watch?v=zzVa_tX1OiI.

*Intermediate:*
Smith et al., 2020; Safran, 2003
   Electrostatic model of protein stability: Bahar et al., 2017, §§3A and 9C.

*Technical:*
Voltaic cells: Saslow, 2021; Schmidt-Rohr, 2018.

Surface forces including electrostatic: Leckband &Israelachvili, 2001.
Bagotskii, 2006.
10.2.2 Electric currents in metals

The conduction of electricity through a copper wire is also a diffusive transport process and also obeys an ohmic relation. But the charge carriers are electrons, not ions; and the nature of the collisions is quite different from that in salt solution. In fact, the electrons could pass perfectly freely through a perfect single crystal of copper; they only bounce off imperfections in the crystal lattice. Figuring out this story required the invention of quantum theory, but one prediction from this idea is straightforward: Thermally induced crystal distortions also count as imperfections, so as mentioned in Section 8.5.3 (page 120) the conductivity of copper is much larger in cryogenic conditions than it is at room temperature.

10.3.4a Solutions with added salt or acid

The solution Equation 10.13 has a disturbing feature: The potential goes to infinity far from the surface! It’s true that physical quantities like the electric field and concentration profile are well behaved (see Your Turn 10C, page 150), but still, this pathology hints that we have missed something. For one thing, no macromolecule is really an infinite plane. But a more important and interesting omission from our analysis is the fact that any real solution has at least some coions; the concentration $c_\sigma$ of salt in the surrounding water is never exactly zero. Rather than introducing the unknown parameter $c_\sigma$ and then going back to set it, this time we’ll choose the constant in $\psi(x)$ so that $\psi \to 0$ far from the surface; then the Boltzmann distribution says

$$c_+(x) = c_\sigma e^{-\psi(x)/k_BT} \quad \text{and} \quad c_-(x) = c_\sigma e^{-(\psi+x)/k_BT}$$

(10.19)

for the monovalent counterions and coions, respectively. The corresponding Poisson–Boltzmann equation is

$$\frac{d^2\psi}{dx^2} = -\frac{1}{2} \lambda_0^{-2} [e^{-\psi} - e^\psi],$$

(10.20)

where again $\psi = \epsilon \psi / k_BT$ and where $\lambda_0$ is defined as

$$\lambda_0 \equiv (\epsilon k_BT/(2e^2c_\sigma))^{1/2}. \quad \text{Debye screening length}$$

(10.21)

In a solution of table salt, with $c = 0.1$ m, the screening length is about 1 nm.

Even in 1D, the general solutions to Equation 10.20 are not elementary (they’re called elliptic functions), but once again, we get lucky for the case of an isolated surface.

Your Turn 10E

Check that

$$\tilde{\psi}(x) = -2\ln \left( \frac{1 + e^{-(x+x_\sigma)/\lambda_0}}{1 - e^{-(x+x_\sigma)/\lambda_0}} \right)$$

(10.22)

solves the equation. In this formula, $x_\sigma$ is any constant. [Hint: It saves some writing to define a new variable, $\xi \equiv e^{-(x+x_\sigma)/\lambda_0}$, and rephrase the Poisson–Boltzmann equation in terms of $\xi$, not $x$.]

Before we can use Equation 10.22, we must fix the value of $x_\sigma$, by imposing the surface boundary condition. Equation 10.7 (page 146) gives

$$e^{x_\sigma/\lambda_0} = \frac{2e^\epsilon}{\epsilon \lambda_0} \left( 1 + \sqrt{1 + \left( \epsilon \lambda_0 \sigma_\sigma / (2e^\epsilon k_BT) \right)^2} \right).$$

(10.23)
10.3.4'b Low-salt limit
Let’s examine the low-salt limit ($\lambda_0 \to \infty$ for fixed $\sigma_q$ and $x$).

**Your Turn 10F**
Show that in this limit, the solution Equation 10.22 becomes a constant plus our pure-water result (Equation 10.13, page 150).

10.3.4'c Far field limit
We can now look at a more relevant limit for biology: This time, hold the salt concentration fixed but consider large distances, where the pure-water result (Equation 10.13, page 150) displays its odd behavior. For $x \gg \lambda_D$, Equation 10.22 reduces to

$$\hat{\psi} \to -(4e^{-x/\lambda_D} e^{-x/\lambda_D}).$$

That is,

*The electric fields far outside a charged surface in salt solution are exponentially screened at distances greater than the Debye screening length $\lambda_D$. (10.24)*

Idea 10.25 and Equation 10.21 confirm an earlier expectation: Increasing $c_m$ decreases the screening length, shrinking the counterion cloud and hence shortening the effective range of the electrostatic interaction (Idea 10.15).

The behavior just found contrasts with a dielectric, whose charges could move slightly but were not fully mobile: In that case, Section 6.5 (page 79) found no exponential screening, just a changed prefactor in the electric field.

10.3.4'd Weakly charged limit; linearized Poisson-Boltzmann equation
In the special case of a weakly charged surface ($\sigma_q$ is small), Equation 10.23 gives $e^{-x/\lambda_D} \approx e\lambda_D \sigma_q/(4\varepsilon k_BT)$, and so the potential simplifies to

$$\psi(x) \approx -\frac{\sigma_q \lambda_D}{\varepsilon} e^{-x/\lambda_D}, \quad \text{potential outside a weakly charged surface}$$

(10.26)

There is a shortcut to this result. If a surface is weakly charged, then $\hat{\psi}$ will never deviate much from zero, and we may work with a simplified form of Equation 10.20:

$$\frac{d^2\hat{\psi}}{dx^2} \approx \frac{\lambda_D^2}{\varepsilon} \hat{\psi}.$$ (10.27)

That linearized equation certainly has solutions of exponential form (like Equation 10.26).

Indeed, we have seen that even a highly charged surface will have weak fields if we stand far enough away from it, and hence a far-field solution of the general form Equation 10.26. However, the prefactor in that solution will not accurately reflect the true surface charge, because the approximate solution breaks down as we approach the surface. Other corrections, such as a breakdown of mean field theory near the surface, can also modify the apparent far-field charge (charge renormalization).

\[15\text{See Problem 10.4.}\]
10.3.4’e Stored energy

In the presence of added salt, the layer thickness no longer grows without limit as the layer charge gets smaller (as it did in the no-salt case, Equation 10.13); rather, it stops growing when it hits the Debye screening length. For weakly charged surfaces, then, the stored electrostatic energy is roughly that of a capacitor with gap spacing \( \lambda_D \), not \( x_0 \). Repeating the argument at the end of Section 10.3.3, we now find the stored energy per unit area to be

\[
\mathcal{E}/(\text{area}) \approx \frac{\sigma_q^2 \lambda_D}{2\varepsilon}.
\]

(10.28)

10.3.4’f How voltaic cells push electrons

We are now in a position to understand the action of a simple voltaic cell. The cell supplies electrical energy until it is discharged (“dead”); some can later be “recharged.” We must carefully distinguish these everyday uses of “charge” from electric charge as used elsewhere in this book; actually, the cell always remains electrically neutral. To avoid confusion, this section will substitute “depleted/restored” for the everyday senses.

First-year physics texts offer a phenomenological model of a voltaic cell as a black box that maintains a fixed electric potential difference between its terminals, regardless of whether, or how fast, it is pushing charge out of its high-potential end. (Some texts modify this description to acknowledge an “internal resistance.”) The energy imparted to charge carriers as they pass through the cell is attributed to “chemistry,” but we are left with several questions:

- What agency pushes electrons down a gradient of \( \psi \), overcoming the electrostatic force whose net effect is to push them the other way?
- The free energy change of a chemical reaction is a scalar, so how does it convert to a vector (the directed force just mentioned)?
- How does the chemical reaction “know” whether the cell is part of a complete circuit, and proceed only if it is? How does the reaction “know” to proceed in reverse when we are restoring the cell?
- Why is the potential drop nearly independent of the total degree of depletion? To what extent is this idealized statement actually true?

We can address all of these questions using ideas from the preceding sections and main text.

Setup

Gasoline-powered auto engines, and certain other devices such as backup power supplies for computers, contain lead–acid batteries, a technology pioneered by G. Planté in 1859. In its modern form, each of six cells in series consists of three elements, which we will simplify as:

- One electrode, a flat plate of metallic lead (Pb), drawn on the left in Figures 10.8–10.9.
- Another electrode, a flat plate of lead oxide (PbO\(_2\)), which also conducts electricity, drawn on the right in the figures.
- A solution of sulfuric acid (H\(_2\)SO\(_4\)) between the plates. We will assume that this solution is dilute, and hence fully dissociated into H\(^+\) and SO\(_4^{2-}\) ions.

\[16\) A “battery” consists of multiple voltaic cells joined in series (to increase their voltage) and/or parallel (to increase capacity). In everyday speech, we often also use “battery” to refer to a single cell, speaking of both a “D battery” and a “9 V battery.” The former is actually a single cell, whereas the latter is six cells in series.
Figure 10.8: [Schematics; sketches of concentration profiles.] **Half-reactions of a lead–acid storage cell.** (a) “Left” reaction. In the lower panel, square brackets around an ion species denote its concentration. The acid solution is neutral in bulk, deviating from neutrality in a thin layer of width roughly the Debye screening length. (b) “Right” reaction.

[In a real automobile battery the plates are actually porous, and initially the solution is concentrated (hence not fully dissociated).]

The main text, and earlier track-2 sections, considered only the simplest chemical reaction: We assumed that when a neutral surface was placed in pure or salt water, certain species would dissociate completely, with coions entering solution and leaving behind immobile charges in the surface. Now we admit more complex reactions at each electrode (“half reactions”):

- On the left, lead atoms can join with SO\textsuperscript{4\textsuperscript{-}} ions to form neutral lead sulfate PbSO\textsubscript{4}, liberating two excess electrons into the conducting lead electrode. PbSO\textsubscript{4} is insoluble, so it remains confined to the left electrode.
- On the right, molecules of lead oxide can join with SO\textsuperscript{4\textsuperscript{-}} and four H\textsuperscript{+} ions to form neutral species: two H\textsubscript{2}O and one PbSO\textsubscript{4}. Two additional electrons must be supplied by the electrode for this reaction to occur.

Like any chemical reaction, each of these is reversible. Like any chemical reaction, each will proceed in a direction that depends on the availability of the various species (their “chemical potentials”), and also on the energy change when the molecules and electrons rearrange. For illustration, we will assume in the following that each is initially favorable in the direction described above; however, all that is really needed is for the overall combined reaction to be favorable. In the case of the lead-acid battery, formation of H\textsubscript{2}O with its very strong chemical bonds makes the overall reaction strongly favored in this direction.

**Left half-reaction**

The key to the first puzzle listed earlier lies in the Nernst–Planck formula (Equation 10.1, page 140), which implies that *entropy can overrule electrophoretic motion*. That is, positively charged ions will climb a potential hill (and negative ions will descend a valley) if a steep enough concentration gradient is pushing them. Suppose that some agency (a “demon”) immobilizes any ion that arrives
Chapter 10  Statistical Electrostatics of Solutions

at a certain plane, removing it from solution. Then the resulting ion sink sets concentration to zero at that boundary, and so diffusion will bring in more ions from the bulk, despite their repulsion from the charge immobilized there.

More realistically, suppose that we plunge an electrically neutral plate of lead into an infinite bath of acid (Figure 10.8a). Because we are assuming that the reaction is chemically favorable under these conditions (temperature, acid concentration), initially it proceeds, depositing some neutral PbSO₄, as well as liberating some electrons (confined to the electrode) and sequestering some SO₄²⁻ ions (previously in solution).

Unlike the case with the imagined demon, however, eventually the reaction will slow, then stop. Although diffusion brings in more ions from the bulk, the repulsion leads to a concentration profile reminiscent of the ones we got in Section 10.3.4d (page 157). Eventually the ion concentration at the plate is so low that the reaction is no longer favorable—it comes to equilibrium, with concentration profiles sketched in Figure 10.8a. The final charge density on the electrode depends on the chemical details of how favorable the reaction was initially. (After all, if acid concentration is zero, then sulfate is unavailable and the reaction cannot occur.)

The final equilibrium state has a layer of excess electrons confined to the electrode, next to a diffuse cloud of net positive charge in the nearby solution (and the original neutral solution beyond that). That is, the reaction has led to charge separation over a length scale set by the Debye screening length of the solution. Although that length scale is typically nanometers, this charge separation still leads to an electrostatic potential gradient near the electrode: The electric potential $\psi$ increases as we move away from the electrode (to the right), then levels off in the neutral bulk solution. The direction of that gradient is set by the spatial asymmetry (electrode on left, solution to its right), answering another of our puzzle points:

- As an analogy, suppose that we pull air out of the top of an otherwise sealed chamber. Molecules from below will rush in to replace those removed, even though they must rise against gravity to do so.
- Similarly, at the electrode ions are being removed, and they can only be replaced from the bulk, even though that requires moving against the electrostatic force. The energy needed to accomplish this motion ultimately comes from the chemical reaction, which liberates slightly less of its free energy change as heat than it did initially.

In each case, a wall (top of the chamber or electrode, respectively) breaks spatial symmetry, choosing a direction.

Now that we understand what is pushing the ions, we can also state the result in energetic terms: Work must be done to move sulfate ions to the charged surface against the potential gradient that their predecessors have built up; eventually this extra effort to take a reaction step cancels the free energy gain of the reaction and the system comes to equilibrium.

Similar reasoning would also apply if the left half-reaction were unfavorable (that is, if its reverse were favorable), leading to a potential gradient with the opposite sign.

Right half-reaction

Next, suppose instead that we plunge an electrically neutral plate of lead oxide into an infinite bath of acid (Figure 10.8b). Initially, the reaction proceeds,
depositing of some neutral PbSO₄, as well as removing two electrons from the electrode, and eliminating some ions from the solution.

Thermal motion (diffusion) again brings in more ions from the bulk.

This reaction also slows, then stops, once repulsion of H⁺ ions from the positive electrode reduces their concentration there to an equilibrium value. This time, we end with a positively charged electrode next to a counterion cloud of negatively charged solution (and beyond that, neutral bulk solution). That spatial separation of charge implies that the electric potential ψ decreases as we move away from the electrode (to the left). Once we get beyond the thin layer of disturbed solution, then ψ levels off in the neutral bulk solution.

We can again restate the conclusion in energetic terms: Work must be done to move H⁺ ions to the charged surface against the electrostatic gradient that their predecessors have built up; eventually this extra effort cancels the free energy gain of the reaction and the system comes to equilibrium, with concentration profiles sketched in Figure 10.8b.

Combine reactions: Open circuit

When we plunge both electrodes into solution, separated by many Debye lengths, then both of the preceding stories play out. Combining the panels of Figure 10.8 and moving from left to right, in equilibrium we find an uptick in ψ, a long region of constant ψ, and another uptick, leading to a net increase, consistent with net negative charge on the left plate and net positive charge on the right plate. However, unlike in a capacitor, the regions of potential change are the thin layers close to each electrode.¹⁹

In contrast, suppose that we plunge two identical lead electrodes into acid. In this symmetrical situation, electric potential rises as we move rightward from the left electrode, holds steady through the bulk, then falls as we approach the right electrode, for no net change.

Even if the left reaction were slightly unfavorable initially, leading to a downtick of ψ, the strongly favorable right reaction could override it, leading to an overall increase as we cross the entire cell from left to right.

Closed circuit

Now imagine joining the two electrodes by wires connected to a very large resistor, for example, an incandescent light bulb (Figure 10.9a). Excess electrons can now migrate slowly through the load, from low potential (left) to high potential (right). As the left plate partially discharges, sulfate ions are slightly less repelled and the reaction becomes once again favorable, pulling more sulfate out of solution. As the right plate gets partially neutralized by incoming electrons, its reaction, too, becomes once again favorable, pulling more sulfate and H⁺ ions out of solution.

Replacement ions now move through the bulk according to the Nernst–Planck formula:²⁰ The bulk solution now sees a net positive object on its left (partially discharged electrode plus its ion cloud) and a net negative object on the right (electrode with extra electrons plus its ion cloud), which create an electric field that pulls sulfate ions to replace those lost on the left, and H⁺ to replace those lost on the right. (Some sulfate ions also move to the right under their concentration gradient.) The net motion of ions incurs some energy dissipation (“internal resistance” of the cell), but this will be negligible if the external resistance is large.

¹⁹Faraday correctly identified the drive as being chemical reactions localized at the electrodes.
²⁰Equation 10.1 (page 140).
Chapter 10  Statistical Electrostatics of Solutions

Figure 10.9: [Schematic; sketch graph.] Closed circuit with load. (a) Both half-reactions, plus continuous depletion of the bulk. (b) Electric potential profile.

The net effect of all these processes is to set up a profile for $\psi$ like that sketched in Figure 10.9b. The electric potential is still single-valued, as it must be, but chemical energy is now being continuously converted to electric form, then converted to heat and light by the load. At the same time, acid in the bulk of the cell is continuously being replaced by water (as well as Pb and PbO$_2$ being replaced by PbSO$_4$).

Eventually the acid becomes so dilute that the free energy cost of extracting another sulfate ion (proportional to the logarithm of concentration) becomes significant. Then even if we disconnect the cell from its load, its potential jump will be less than it was initially: The cell is “depleted.”

On the other hand, if by some means we force even more excess electrons onto the left electrode than are present in the open circuit, then the overall reaction can become favorable in the direction opposite to that shown in the figure, and we restore the acid concentration, for example converting the lead sulfate back to neutral lead and sulfate ions.

Some familiar phenomena

The preceding outline is highly simplified, but it already explains many things about voltaic cells:

- For a given initial concentration, the total charge that can be pushed through a cell before it becomes depleted is proportional to the volume of solution. This quantity is often expressed in units of mA hour, and indeed the rating of a big “D” cell is larger than that of a small “AAA” cell.
- However, the potential jump across the cell depends only on the chemical reaction (and temperature and acid concentration), not on the physical size of the cell. Indeed, both the “D” and “AAA” cells supply the same 1.5 V.
- Moreover, the dependence of $\Delta\psi$ on acid concentration is initially slight, because $\ln c$ is a slowly varying function of $c$ when $c$ is large. Indeed, commercial voltaic cells maintain nearly...
constant \( \Delta \psi \) throughout most of their life, justifying the textbook simplification that a battery simply sets up a fixed potential jump, possibly degraded by internal resistance, then passes whatever current is implied by that jump and the rest of its circuit. This observation answers another of our starting puzzles.

- For a given initial concentration, the maximum rate of depletion is limited by the chemical reactions, and hence by the surface areas of the electrodes. We have considered only the high-load limit, where these rates are not limiting, but real batteries use porous electrodes to maximize surface area and hence peak current.
- After a cell is depleted, we can connect it to a fresh battery (a "jumpstart battery box" for autos) or other source of electricity (the "alternator" in an auto). Then it actually becomes favorable for the reactions to run in reverse, restoring the acid concentration and removing \( \text{PbSO}_4 \) from the electrodes. Indeed, each time you use the battery to start a gasoline-powered engine, the alternator later restores the original state, so that you can start again next time.
- If you repeatedly try to start an auto engine unsuccessfully (for example, because there is no gasoline), eventually the battery becomes depleted and even the starter won’t operate. But if you wait an hour or two (and add gasoline), then miraculously the engine may start even though you took no action to restore the battery. The explanation is that in order to generate a potential jump, the relevant ions must be present right at the electrodes, and their replenishment via diffusion from the bulk takes time. So after rapid partial depletion, the battery may only seem to be dead.\(^{21}\)

10.3.5' Alternative derivation of force

The crucial last step leading to Equation 10.18 may seem too slick. We can instead work out the force in the same direct way we calculate any entropic force, by taking a derivative of the free energy. To that end, we now compute the Helmholtz free energy \( \mathcal{F} \) of the system of counterions+surfaces, holding fixed the charge density \( -\sigma_0 \) on each surface but varying the separation \( 2h \) between the surfaces (see Figure 10.4b, page 147). Then the force between the surfaces will be \( p\Sigma = -d\mathcal{F} / dh(2h) \), where \( \Sigma \) is the surface area.

As in the main text, suppose singly-charged counterions (charge \( +e \)). Define a convenient length scale, the Bjerrum length:

\[
\ell_B \equiv \frac{e^2}{4\pi\varepsilon k_B T}.
\]  

\( (10.29) \)

First we notice an important property of the Poisson–Boltzmann equation (Equation 10.10, page 148). Multiplying both sides by \( d\Psi / dx \), we can rewrite the equation as

\[
\frac{d}{dx} \left( \frac{d\Psi}{dx} \right) = 8\pi \ell_B \frac{dc_+}{dx}.
\]

Integrating this equation gives a first-order equation:

\[
\left( \frac{d\Psi}{dx} \right)^2 = 8\pi \ell_B (c_+ - c_0).
\]  

\( (10.30) \)

To fix the constant of integration, we noted that the electric field is zero at the midplane, and \( c_+(0) = c_0 \) there.

\(^{21}\)Our simplified discussion considered only the limit of very slow depletion, so that diffusion could keep pace with it.
The free energy density of an inhomogeneous ideal gas (or dilute solution) is
\[ c_s(\vec{r}) \left( q\vec{\phi}(\vec{r}) + k_B T \ln(c_s(\vec{r})/c_\infty) \right). \]

Here \( c_s \) is a constant whose value will drop out of our final answer because the integral \( \int c_s \, dx = 2\sigma_0/e \) is a constant, by charge neutrality. The free energy for our problem is the integral of this quantity, plus the electrostatic energy of the two negatively charged plates at \( x = \pm h \):
\[ \mathcal{F}/(k_B T \times \text{area}) = -\frac{1}{2} \frac{\sigma_0}{e} \left( \frac{\psi(h) + \psi(-h)}{2} \right) + \int_{-h}^{h} dx \left[ c_s \ln \frac{c_s}{c_\infty} + \frac{1}{2} q_s \psi \right]. \]

We simplify our expression by first noting that \( \ln(c_s/c_\infty) = \ln(c_0/c_\infty) - \psi \), so the terms in square brackets are \( c_s \ln(c_0/c_\infty) - (1/2)c_s \psi \). The first of these terms is a constant times \( c_s \), so its integral is \( 2(\sigma_0/e) \ln(c_0/c_\infty) \). To simplify the second term, use the Poisson–Boltzmann equation to write \( c_s = -(4\pi e)^{-1} (d^2\psi/dx^2) \). Next integrate by parts, obtaining
\[ \mathcal{F}/(k_B T \times \text{area}) = \frac{\sigma_0}{e} \left[ \ln \frac{c_0}{c_\infty} - \frac{1}{2} \psi(h) \right]_{-h}^{h} + \frac{1}{8\pi e B} \frac{d\psi}{dx} \bigg|_{-h}^{h} - \frac{1}{8\pi e B} \int_{-h}^{h} dx \left( \frac{d\psi}{dx} \right)^2. \]

We evaluate the boundary terms by using Equation 10.12 (page 149) at \( x = -h \) and its analog on the other surface; they equal \(-\sigma_0/e \psi(h)\).

To do the remaining integral, recall Equation 10.30: it's \( -\int_{-h}^{h} dx (c_s - c_0) + 2(\alpha c_0 - \sigma_0/e) \). Combining these results gives
\[ \mathcal{F}/(k_B T \times \text{area}) = 2hc_0 + \frac{\sigma_0}{e} \left( \ln \frac{c_0}{c_\infty} - \psi(h) - 1 \right) = \text{const} + 2hc_0 + \frac{\sigma_0}{e} \ln \frac{c_s(h)}{c_\infty}. \]

The concentration at the wall can again be found from Equations 10.30 and 10.12: \( c_s(h) = c_0 + (8\pi e B)^{-1} (d\psi/dx)^2 = c_0 + 2\pi \sigma_0 (\sigma_0/e)^2 \).

A few abbreviations will make for shorter formulas. Let \( b = 2\pi e B \sigma_0/e \) and \( u = \xi h \), where \( \xi = \sqrt{2\pi e B c_0} \) as in Section 10.3.5 (page 151). Then \( u \) and \( \xi \) depend on the gap spacing, whereas \( b \) does not. With these abbreviations,
\[ \mathcal{F}/(k_B T \times \text{area}) = 2hc_0 + \frac{b}{\pi e B} \ln \frac{c_0 + b^2/(2\pi e B)}{c_\infty}. \]

We want to compute the derivative of this expression with respect to the gap spacing, holding \( \sigma_0 \) (and hence \( b \)) fixed. We find
\[ \frac{p}{k_B T} = \frac{1}{k_B T} \left( \frac{d(\mathcal{F}/(k_B T \times \text{area}))}{d(2h)} \right) = -c_0 - \left( h + \frac{b}{2\pi e B c_0 + b^2} \right) \frac{dc_0}{dh}. \]

In the last term, we need
\[ \frac{dc_0}{dh} = \frac{d}{dh} \left( \frac{u^2}{h^2 2\pi e B} \right) = \frac{u}{\pi e B h} \left( h \frac{du}{dh} - u \right). \]

To find \( du/dh \), we write the boundary condition (Equation 10.16 (page 152)) as \( bh = u \tan u \) and differentiate to find
\[ \frac{du}{dh} = \frac{b}{\tan u + u \sec^2 u} = \frac{bu}{hb + u^2 + (hb)^2}. \]

This has gone far enough. In Problem 10.8, you'll finish the calculation to get a direct derivation of Equation 10.18. For a deeper derivation from thermodynamics, see Israelachvili, 2011, §12.7.

---

Notice that adding any constant to \( \psi \) leaves this formula unchanged. To understand the reason for the factor \((1/2)\) in the first and last terms, think about two point charges \( q_1 \) and \( q_2 \). Their potential energy at separation \( r \) is \( q_1 q_2 / (4\pi \varepsilon r) \) (plus a constant). This is one half of the sum \( q_1 \psi_1(r) + q_2 \psi_2(r) \). (The same factor of \((1/2)\) also appeared in the electrostatic self-energy Example on page 78.)

November 14, 2023; Contents Index Notation Glossary
10.1 **Charged surfaces**

a. Use numerical software to solve Equation 10.16 for $\xi$ as a function of plate separation $2h$ for fixed charge density $\sigma_q$. For concreteness, take $\sigma_q$ to equal $e/(20 \text{ nm}^2)$. Now convert your answer into a force by using Equation 10.18 and compare your answer qualitatively with Figure 10.7.

b. Obtain Dataset 1. Repeat (a) with other values of $\sigma_q$ to find the one that best fits the upper set of points in the figure at separation greater than 2 nm. If this surface were fully dissociated, it would have one electron charge per 7 nm$^2$. Is it fully dissociated?

10.2 **Counterions in cylindrical geometry**

Section 10.3.3 discussed the counterion distribution for a planar, charged surface. The text concluded that the counterions do not run away to infinity; that is, there is a nonzero concentration of ions near the surface.

One way to understand this result is to consider a single ion (of charge $e > 0$) near a surface with areal charge density $\sigma_q < 0$. Suppose that the ion is initially confined to a distance $a$ from the surface. If the ion is now allowed to explore a larger distance $R$ from the surface, then the increase in its entropy is $k_B \ln(R/a)$. However, the electrostatic energy cost for the ion to travel out to a distance $R$ is $e(R - a)\sigma_q/\varepsilon$. The change in free energy is thus approximately $\Delta F \approx e(R - a)\sigma_q/\varepsilon - k_B T\ln(R/a)$, which increases as $R$ gets very large. Therefore, to minimize the free energy, the ion does not run away to infinity but remains near the surface.

a. Using a similar argument, determine whether or not the counterions will run away to infinity for an infinite-length charged cylinder of radius $b$ and charge per unit length $\sigma_q^{(1D)}$.

b. Apply your result to the case of DNA, which has two ionized phosphate groups (charge $-2e$) for every basepair.

10.3 [Not ready]

10.4 **Weak-charge limit**

Section 10.3.3 considered an ionizable surface immersed in pure water. Thus, the surface dissociated into a negative plane and a cloud of positive counterions. Real cells, however, are bathed in a solution of salt, among other things; there is an external reservoir of both counterions and negative coions. Section 10.3.4 (page 156) gave a solution for this case, but the math was complicated; here is a simpler, approximate treatment.

Instead of solving Equation 10.20 exactly, consider the case where the surface’s charge density is small. Then the potential $\psi(0)$ at the surface will not be very different from the value at infinity, which we took to be zero. (More precisely, the dimensionless combination $\tilde{\psi}$ is everywhere much smaller than 1.) Approximate the right-hand side of Equation 10.20 by the first two terms of its Taylor series expansion in powers of $\tilde{\psi}$. The resulting approx-
imate equation is easy to solve. Solve it, and give an interpretation to the quantity \( \lambda_D \) defined in Equation 10.21.

10.5 \( \text{T}_2 \) Counterion cloud

If you haven’t done Problem 10.4, look at it before attempting this problem, then use a similar approach here.

Consider a spherical macromolecule of charge \( q = ze \) and radius \( a \) in a solution containing a monovalent salt, such as sodium chloride. As discussed in Problem 10.4, in the limit that the potential satisfies \( |\psi(r)| \ll k_B T / e \), you may approximate the Poisson–Boltzmann equation in its linearized form. In spherical coordinates, the resulting equation is

\[
\frac{1}{r^2} \frac{d}{dr} \left( r^2 \frac{d\psi(r)}{dr} \right) = \frac{1}{\lambda_D^2} \psi(r),
\]

where \( \lambda_D \) is the Debye length.

a. Justify the following boundary conditions:

\( \psi(r) \to 0 \) as \( r \to \infty \), \hspace{1cm} \( - \frac{d\psi}{dr} \bigg|_{r=a} = E_i \) (surface) = \( \frac{q}{4\pi \varepsilon a^2} \).

b. Find \( \psi(r) \) in terms of \( \lambda_D \), \( a \), and \( q \).

c. The net charge density from salt ions is given by

\[
\rho_q(r) = e \rho_\infty \left( e^{-\psi} - e^{+\psi} \right) \approx - \frac{e k_B T}{\varepsilon \lambda_D^2} \psi.
\]

Using your result from (b), show explicitly that the integral of this charge density is equal to \(-q\).

d. Imagine placing the charge \( q \) on the surface of the spherical macromolecule by successive increments \( dq \). By integrating the work required to bring the charge from zero up to \( q \), find the total potential energy of the charged macromolecule and its neutralizing cloud.

e. The solubility of proteins in dilute salt solution generally increases with increasing ionic strength of the solution. Use your result from (d) to explain this effect qualitatively.

10.6 \( \text{T}_2 \) Salt I

a. Calculate the Debye screening length for a 100 mM solution of sodium chloride. That is, the concentration of \( \text{Na}^+ \) ions is 0.1 mole per liter.

b. But magnesium chloride, for example, dissociates into \( \text{Mg}^{2+} \) ions (and twice as many \( \text{Cl}^- \) ions). So recalculate the Debye screening length for a salt solution whose ions are not necessarily monovalent (singly charged). Do this by writing the appropriate Poisson–Boltzmann equation, linearizing it, and collecting terms.

c. Evaluate your answer for a 100 mM solution of magnesium chloride. That is, the concentration of \( \text{Mg}^{2+} \) ions is 0.1 mole per liter.
10.7  \( T_2 \)  Salt II

*Context:* The main text claimed that electrostatic interactions in solution have a number of features that make them well suited to implement the remarkable specificity of interactions between biomacromolecules. In this problem, you’ll explore the ranges of both the overall attraction due to total net charge, and also of the pattern-dependent part of the attraction.

*Setup:* Consider a surface that is the infinite \( xy \) plane. Suppose that the electric field inside the surface is everywhere zero, so that the potential gradient at the surface reflects the areal charge density. But unlike the discussion the main text, suppose that the fixed charge distribution on the surface is a constant plus a “chessboard” component, that is, that

\[
\frac{\partial \psi}{\partial z} \bigg|_{z=0} = A + B \sin(kx) \sin(ky).
\]

Suppose that the surface is immersed in a salt solution with Debye screening length \( \lambda \). Suppose that \( A \) and \( B \) are both small enough to justify linearizing the Poisson–Boltzmann equation (Problem 10.4).

*Do:* Find \( \psi(x, y, z) \). Comment on the \( z \) dependence of your solution in light of the above remarks.

10.8  \( T_2 \)  Direct calculation of a surface force

Finish the derivation of Section 10.3.5’ (page 163). The goal is to establish Equation 10.18.

10.9  \( T_2 \)  

[[Not ready]]
CHAPTER 11

The Cable Equation

11.1 FRAMING: THE ILL-FATED TRANSATLANTIC CABLE

By 1854, the first industrial revolution (steam power) had already transformed the world, and the second one (electric generation, motors, lights and related technology) was underway. But in at least one sense, the world remained unimaginably primitive: It still took weeks for any information to pass between Europe and America. The telegraph, by then a decade old, had eliminated communication barriers within continents, but between them, the only method of communication was by ship. In that year, a retired industrialist named Cyrus West Field decided to rectify this unsatisfactory situation. How hard could it be? One could simply string a cable across the narrowest part of the Atlantic ocean. With the growing economic significance of the United States, the first corporation to accomplish this simple task could reap enormous profits.

Field was ready to supply some of the needed capital investment, and he had the connections to bring in others like himself. But he also had the foresight to engage William Thomson, the future Lord Kelvin and already a noted expert on electricity. Thomson took the assignment, but he saw some clouds on the horizon: Existing, but shorter, undersea cables in the Mediterranean were not behaving as expected. When electric current was poured in one end of such cables, a lot of it... disappeared. Worse, when crisp on/off telegraph signals were sent in one end they arrived blurry at the other end (to the extent that they arrived at all).

Undersea cables had a “coaxial” structure. The one eventually laid across the Atlantic contained seven strands of a good conductor (copper) down the middle, surrounded by insulators, and then a layer of iron strands, similar to those used in suspension-bridge cables. The iron was a poor conductor of return current; its main job was to supply strength, so that the entire cable could withstand undersea currents, as well as the stress from its own weight as it was reeled out from a giant spool on the ship initially laying it. The overall diameter was 1.8 cm. Here is a small chunk of the original cable:
Developing older ideas from Michael Faraday, Thomson realized that part of the transmission problem must be the capacitance of existing cables: Instead of passing all the way through the cable and out the other end, some charge could simply stop in the middle, paying a finite energy cost to create an electric field across the insulating layer. Charge could also leak across the finite resistance of the insulating layer, again never arriving at the other end at all. Both loss mechanisms were unexpected because for overland transmission cables they were negligible: There the standard design was a pair of wires separated by a meter of air, with negligible capacitance per unit length (and enormous leak resistance per length).

Thomson therefore recommended reengineering the cables with a much thicker insulation layer than had originally been planned. Unfortunately, the thin cable had already been ordered and paid for. Field took the time-honored approach of finding another engineer willing to reassure him that everything would be fine. The new chief engineer in turn pulled the elderly Faraday out of retirement for a public meeting to reassure the investors, after first misleading Faraday about some recent experimental results. Cable-laying began in 1857.

The first attempt ended in failure with the cable snapping in water too deep to retrieve the lost end. Another attempt the following year involved two ships. They planned to meet in the middle of the Atlantic, splice their respective cables together, then head for Ireland and Newfoundland respectively, paying out cable as they went.

The operation immediately encountered one of the worst storms recorded in the North Atlantic. The ships were damaged; the cable snapped more than once and had to be spliced; one ship was attacked by an angry whale. Nevertheless, ultimately an intact cable at last stretched across the ocean. Wild celebrations ensued before the device had even been tested, including a torchlight procession that set fire to New York’s City Hall.

Most of the initial telegraph traffic on the cable consisted of “Send more slowly,” “Repeat,” or simply “What?” It took sixteen hours to transmit the Queen’s 99-word congratulations to the US President, and thirty hours for the equally brief reply. Desperate to get a stronger signal, the lead engineer increased the voltage supplied to the cable, until the insulation broke down somewhere in the middle of the ocean, turning the entire cable into worthless undersea trash. The investors lost their money. Eventually a rumor spread that the entire project had been a massive hoax. After another such disaster in the Red Sea, a parliamentary inquiry was mounted to see who should be blamed. Not until 1866 did a successful cable, following Thomson’s original advice, come into operation.

This chapter will explore Thomson’s analysis, then develop a remarkable parallel between undersea cables and living nerve cells. We’ll see that:

*Electromagnetic phenomenon:* Small electrical disturbances on a nerve axon spread diffusively.

*Physical idea:* The cable equation is closely related to the diffusion equation.
Table 11.1: **Symbols used in this chapter.** See also Appendix B.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>distance along cable axis</td>
</tr>
<tr>
<td>$a$</td>
<td>cable radius (Figure 11.1)</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>conductivity of central conductor</td>
</tr>
<tr>
<td>$g$ or $g_{\text{tot}}$</td>
<td>conductance per area of insulating sheath; $g_{\ell}$, membrane conductance per area for ion species $\ell$</td>
</tr>
<tr>
<td>$g_{\ell}^0$, $g_{\ell}'$</td>
<td>specifically the resting and excited conductances per area, respectively.</td>
</tr>
<tr>
<td>$\Delta \Sigma$</td>
<td>area of a segment of insulating sheath</td>
</tr>
<tr>
<td>$C$</td>
<td>capacitance of a segment of insulating sheath; $C_{\ell}$, per area</td>
</tr>
<tr>
<td>$R_a$, $R_r$</td>
<td>axial and radial (“leak”) resistances, respectively, for a segment of length $\Delta x$ and surface area $\Delta \Sigma$</td>
</tr>
<tr>
<td>$R_{\ell r} = (g_{\ell} \Sigma)^{-1}$</td>
<td>membrane resistances for individual ion species $\ell$</td>
</tr>
<tr>
<td>$\psi_{\text{in}}(x, t)$</td>
<td>interior electric potential (Figure 11.1)</td>
</tr>
<tr>
<td>$\psi_{\text{out}}$</td>
<td>exterior electric potential, $= 0$ in our simplified model so $\Delta \psi = \psi_{\text{in}} - \psi_{\text{out}} = \psi_{\text{in}}$ (Figure 11.1)</td>
</tr>
<tr>
<td>$\psi_{\text{Nernst}}^\ell$</td>
<td>Nernst potential for ion species $\ell$ (Section 10.2.1, page 139)</td>
</tr>
<tr>
<td>$\psi^0$</td>
<td>combination of Nernst potentials giving the resting potential (Your Turn 11F, page 178)</td>
</tr>
<tr>
<td>$w$</td>
<td>modified potential (Your Turn 11B, page 173)</td>
</tr>
<tr>
<td>$v(x, t)$</td>
<td>depolarization ($\Delta \psi$ shifted by $\psi^0$) (Equation 11.10, page 178)</td>
</tr>
<tr>
<td>$I_a$</td>
<td>axial (rightward) electric current inside cable (Figure 11.1)</td>
</tr>
<tr>
<td>$I_r$</td>
<td>radial (outward) electric current through a segment (leak plus capacitive, Figure 11.1)</td>
</tr>
<tr>
<td>$j_r$</td>
<td>radial charge flux (current per area) actually passing through axon membrane; $j_{\ell r}$, contribution from ion species $\ell$ (Equation 11.9, page 176)</td>
</tr>
<tr>
<td>$\lambda_{\text{cable}}$, $\tau_{\text{cable}}$</td>
<td>space constant and time constant of cable (Equation 11.6, page 172)</td>
</tr>
<tr>
<td>$\theta$</td>
<td>speed of a traveling wave (Your Turn 11E, page 174)</td>
</tr>
<tr>
<td>$c_{\ell, \text{in}}$, $c_{\ell, \text{out}}$</td>
<td>concentration of ion species $\ell$ inside (respectively outside) a cell</td>
</tr>
</tbody>
</table>

### 11.2 COAXIAL CABLE

This chapter introduces a lot of notation. For reference, Table 11.1 lists some symbols introduced below.

#### 11.2.1 A mathematical hyperlink to heat conduction

Thomson had understood both the loss and the spread of signals before the transatlantic cable was even attempted. He found his way through the physical problem by an approach that is routine today but astonishing in the mid-19th century: He set up the problem mathematically, then noticed that it involved the same equation as a problem that seemed physically to be completely different. The same equation must have the same solutions, so Thomson benefited at once from extensive work that had already been done on the other problem. Let’s see how that worked.

We’ll make some idealizations. Imagine a cable consisting of a solid cylindrical core of ohmic conductor (such as copper), surrounded by a sheath of partially insulating material, which in turn is surrounded by a perfect conductor. That last assumption is purely for mathematical convenience; if we relax it, the equations just get a bit longer.¹

¹Actually, an undersea cable is surrounded by an infinite bath of salt water, so it’s not so unreasonable to neglect...
Finally, we continue to work in the quasi-static regime, where we may neglect the back-reaction of any magnetic fields on electric fields and currents.²

Let \( a \) be the radius of the central core and \( \kappa \) its conductivity. Let \( g \) be the leak conductance per unit area of the insulating sheath. It’s positive and has units \( \Omega^{-1} \text{m}^{-2} \). Also let \( C \) denote the capacitance per area.

If the system is isolated, it will eventually come to the boring state with potential everywhere uniform. We are interested in transient solutions that have not yet arrived at that state, so we need to find and solve some equation.

11.2.2 Discrete-element models as stepping-stones to distributed elements

Both capacitance and resistance are continuously distributed along our cable. However, things will look more familiar if we imagine dividing the cable into segments of length \( \Delta x \) and surface area \( \Delta \Sigma = 2\pi a \Delta x \), treating them as discrete elements (see Figure 11.1). This is not an approximation, because later we’ll take the limit \( \Delta x \to 0 \).

What is an approximation is that we’ll assume that the potential is uniform throughout every cross-section of the central conductor.⁴ The potential may jump across the insu-

---

²See Section 8.6. This approximation breaks down at high frequency; Chapter 18 gives a more general discussion.

³See Section 8.5.1 (page 119). Note that conductance per area has units different from those of the conductivity, \( \kappa \), of a bulk material: The latter has units \( \Omega^{-1} \text{m}^{-1} \).

⁴At ultra-high frequencies, a “skin effect” confines current to just the outermost part of a wire, invalidating this assumption.
lating sheath, however, and it may also vary along the length of the (very long) conductor. Each segment has axial resistance \( R_x = \Delta x / (\pi a^2 \kappa) \) for the inner conductor. Because we are imagining that the corresponding axial resistance \( R_x^s \) for the outer material is zero, right away we learn that the exterior potential is a constant, which we may take to be \( \psi_{\text{out}} = 0 \), and so the potential drop across the sheath is just \( \psi_{\text{in}} \).

Another resistance, \( R_t = (g \Delta \Sigma)^{-1} \), impedes radial current passage through the insulating sheath (“leakage”). However, charge can instead approach the sheath and pile up against it, as long as an equal charge leaves the other side. The capacitance \( C = \epsilon \Delta \Sigma \) accounts for the electrostatic cost of this local separation. The combined effect of charge passage and charge pileup is symbolized in the figure by a resistor \( R_t \) and a capacitor \( C \) in parallel for each segment.

Currents must balance in the bulk of the interior and exterior compartments, because in the quasi-static approximation, no net charge can build up in a uniform medium. Thus, for example, the three-way junctions at the top must each have zero net current flowing into them:

\[
I_x(t, x) - I_x(t, x + \Delta x) = I_x(x) = \psi_{\text{in}}(t, x)/R_t + C \frac{\partial \psi_{\text{in}}}{\partial t}.
\]  
(11.1)

(We used the fact that charge entering each resistor on the top must all leave it: \( I_x \) is the same on both sides of a resistor.) Finally, the hypothesis of ohmic behavior in the core says

\[
\psi_{\text{in}}(x - \Delta x) - \psi_{\text{in}}(x) = I_x(x) R_x.
\]  
(11.2)

To summarize, we have expressed the discrete element properties in terms of material characteristics and geometry parameters:

\[
C = \epsilon \Delta \Sigma \quad \text{and} \quad R_t = 1/(g \Delta \Sigma), \quad \text{where} \quad \Delta \Sigma = 2\pi a \Delta x.
\]  
(11.3)

Also we have (Section 8.5.1, page 119) that

\[
R_x = \Delta x / (\kappa \pi a^2).
\]  
(11.4)

11.2.3 The linear cable equation explains the observed dispersion of signals

**Your Turn 11A**

Combine the preceding formulas and take the continuum limit, obtaining

\[
\kappa \pi a^2 \frac{\partial^2 \psi_{\text{in}}}{\partial x^2} = 2\pi a \left( g \psi_{\text{in}} + C \frac{\partial \psi_{\text{in}}}{\partial t} \right).
\]  
(11.5)

Define the **space constant** and **time constant** as

\[
\lambda_{\text{cable}} \equiv \sqrt{a \kappa/(2g)}; \quad \tau_{\text{cable}} \equiv C/g.
\]  
(11.6)

(Check that these expressions have the units of length and of time, respectively.) These abbreviations yield

\[
(\lambda_{\text{cable}})^2 \frac{\partial^2 \psi_{\text{in}}}{\partial x^2} - \tau_{\text{cable}} \frac{\partial \psi_{\text{in}}}{\partial t} = \psi_{\text{in}}.
\]  
**linear cable equation**  
(11.7)
Your Turn 11B

Change variables from $\psi_{\text{lin}}$ to $w(x, t) \equiv e^{t/\tau_{\text{cable}}} \psi_{\text{lin}}(x, t)$ and show that the linear cable equation becomes

$$\frac{(\lambda_{\text{cable}})^2}{\tau_{\text{cable}}} \frac{\partial^2 w}{\partial x^2} = \frac{\partial w}{\partial t}. $$

Thomson’s great insight was to recognize this equation as mathematically identical to the diffusion equation, at that time famous from Fourier’s recent study of heat conduction. The analog of the diffusion constant is $(\lambda_{\text{cable}})^2/\tau_{\text{cable}} = x a/(2c)$, so we see that a cable with small capacitance will transmit signals without much spreading.\(^5\)

We already know some solutions to the diffusion equation.

Your Turn 11C

Show that the following function solves Equation 11.7 (Figure 11.2a):

$$\psi_{\text{lin}}(t, x) = \text{const} \times e^{-t/\tau_{\text{cable}}}t^{-1/2}e^{-x^2/(4Dt)}, \quad \text{passive-spread solution \quad (11.8)}$$

where $D$ is the combination of cable parameters just mentioned.

This particular solution gives the response of our cable to a localized injection of current.

\(^5\)This is the result that led Thomson to propose redesigning the cable with thicker insulation (smaller $c$) and thicker central conductor (bigger $a$). But the Suits declared it was too late and too expensive to change the design.

Figure 11.2: [Mathematical functions.] Functions of two variables. (a) A function $\phi(x, t)$, describing diffusion as a concentrated lump of solute begins to spread. Notice that time is drawn as increasing as we move diagonally downward in the page (arrow). The heavy line is the concentration profile at one particular time, $t = 1.6$. (b) This surface plot represents a function $\phi(t, x)$, describing a traveling wave. The heavy line shows the time course as seen by an observer fixed at $x = 0.7$. An initial disturbance dies out due both to leakage and capacitive charging.
Chapter 11 The Cable Equation

It’s a gaussian profile at any instant of time, which initially widens out fast, then slows down, all the while dying off exponentially in time.

**Your Turn 11D**

Imagine sitting at a fixed location $x_s$ and observing the time course of the potential disturbance. At what time does the disturbance reach its peak? How does peak strength vary as a function of $x_s$? Use a computer to draw $\psi_{in}(t, x_s)$ for various $x_s$.

Even if there is no leak conductance ($g \to 0$), our passive cable still suffers from dispersion. (Indeed, $g$ dropped out altogether in the expression for the diffusion constant.) Moreover, the linear cable equation has no traveling wave solutions:

**Your Turn 11E**

Substitute a trial solution of the form $\psi_{in}(t, x) = f(x - \delta t)$, into Equation 11.7, where $\delta$ is a constant, the speed of the proposed traveling wave (Figure 11.2b). Is there any value of $\delta$ that yields a physical solution?

11.3 NEURONS

11.3.1 Action potentials propagate without dispersion or attenuation

People speak casually about the brain as a “computer” and its neurons as “wires,” but a little thought shows they must be very different from ordinary wires. A coaxial cable brings Internet into your apartment via signals that move at around $2 \cdot 10^8$ m/s. Your nerves carry signals that move at around 10–20 m/s—ten million times slower than the coaxial cable! They also manage to do this despite being surrounded by a conductive medium.

A neuron has a long projection, its axon, that is a “cable” of the sort we are considering: it is a tube of conductor (largely salt water) surrounded by a partially insulating layer (cell membrane), which is surrounded by another conductor (salt water). So we may expect that electrochemical disturbances will also spread diffusively along an axon.

For some nerve cells, that’s good enough (for example, photoreceptors in the eye). They are short, and diffusive spreading is not a problem over a few micrometers. Longer nerve axons also exhibit passive-spread behavior when stimulated with very small disturbances. But that wouldn’t be very useful for, say, the axons that start in your spinal cord and end a meter away in your foot! In fact, above a threshold of stimulation, axons transmit a traveling nerve impulse, called the action potential, that moves unchanged in form, at constant speed. The top curve in Figure 11.3 shows a typical waveform. Your result in in Your Turn 11E seemed to show that such behavior is impossible, so we have work to do.

---

6Hermann von Helmholtz measured this speed in 1850.
11.3 Neurons

Figure 11.3: Experimental data. The role of sodium in the conduction of an action potential. One of the top traces was taken on a squid axon in normal seawater before exposure to low sodium. In the middle trace, external sodium was reduced to one-half of its concentration in seawater, and in the bottom trace, to one-third. (The other top trace was taken after normal seawater was restored to the exterior bath.) The data show that the peak of the action potential tracks the sodium Nernst potential across the membrane (Equation 10.3, page 141), an observation supporting the idea that the action potential is a sudden increase in the axon membrane’s sodium conductance. [Data from Hodgkin & Katz, 1949.]

It is true that axons are filled with lots of other machinery, including microtubules. P. Baker and coauthors did experiments in which all those contents were emptied out of the axon and it was refilled with just a salt solution with concentrations of sodium and potassium similar to the interior of a living cell (and hence different from the exterior, Figure 8.5, page 123). Amazingly, all the phenomena we will discuss (passive spread and the action potential) behaved identically with these gutted axons as they do in living cells. That is, action potentials depend on just two key elements:

- Ion concentration imbalance. Specifically, excess exterior sodium ions are required (Figure 11.3). In the gutted axon experiment, there is not even any ATP nor other “energy molecule” present whose hydrolysis could sustain an action potential, countering dissipative (ohmic) loss.
- There must also be some specific property of the cell membrane that we have not yet accounted for. Certainly an ordinary glass capillary containing the same ion solution won’t support action potentials.

In the rest of this chapter and the next, these clues will lead us to the mechanism of the action potential.

11.3.2 Some ion species are far out of equilibrium

Let’s begin by considering ionic concentrations. We are studying a quasi-static situation, so the net charge density in bulk must be everywhere zero. For electrons in a metal, the neutralizing atomic nuclei are fixed in space. Charge neutrality then implies that, although the electrons are mobile, their density cannot vary. Salt water conducts electric current by the movement of ions, not electrons, but we studied this already in Section 10.2.2. There we saw that one key difference with ordinary conduction in metals is that there are several types of ions, in contrast to just one charge carrier (electrons or holes) in a metal. Each ion species $\ell$ has its own concentration $c_\ell$.
Thus, in aqueous solution charge neutrality does not prohibit a change in one ion’s concentration, as long as the other species make compensating changes.

The membrane leakage conductances per area for each ion species, \( g_\ell \), can all have different values, because the membrane itself is insulating (Section 6.9); ions are passed only through ion channels embedded in the cell membrane.\(^7\) Far from being featureless tubes, each class of channels is sculpted in a way that selects for a particular ion (or type of ions).\(^8\)

Thus, the net charge flow (current) through a channel due to ion species \( \ell \) is the conductivity for that species times the sum of two driving forces:

- There is an electrostatic force proportional to the difference of electric potentials on either side of the membrane times the charge on species \( \ell \).
- There is also a thermodynamic force, involving the difference of concentrations. Just like the air in a balloon, ions will “want” to escape from the side where their concentration is greater.

Indeed, Chapter 10 showed that equilibrium with given inner and outer concentrations requires a potential drop called the Nernst potential for species \( \ell \):

\[
\psi_\ell^{\text{Nernst}} = -\frac{k_B T}{q_\ell} \ln(c_{\ell, \text{in}}/c_{\ell, \text{out}}).
\]

But beware: The Nernst potential may not be equal to the actual potential drop. If they disagree, that just means that species \( \ell \) is out of equilibrium, and hence will flow if given the opportunity. So we expect that, at least for small deviations from equilibrium, the resulting ion flow will give rise to a charge flux via a linear relation:

\[
J_{r, \ell} = (\Delta \psi - \psi_\ell^{\text{Nernst}})g_\ell. \quad \text{ohmic conductance hypothesis} \quad (11.9)
\]

This formula gives the radial charge flux contribution from species \( \ell \), with the sign convention that positive means net charge leaving the axon (radially outward). The potential drop is defined as \( \Delta \psi = \psi_{\text{in}} - \psi_{\text{out}} \), and in our simplified model \( \psi_{\text{out}} = 0 \). The conductance per area \( g_\ell \) involves the permeability of a channel, the density of channels in the membrane, and the square of the charge carried by species \( \ell \);\(^9\) it is always a positive quantity.

Equation 11.9 makes precise a claim introduced in Section 8.7.3: The two terms mean that there can be net flow of ions against the electrostatic gradient, if the “pressure” term outweighs the “field” term.\(^10\)

Here are some typical values for three ion species that are relevant in the squid “giant” axon (so called because it can be up to a millimeter in diameter—not because it comes from a giant squid):\(^11\)

---

\(^7\)See Section 6.9 (page 88) and Section 8.7.2 (page 122).

\(^8\)However, we will make the approximation that ions of each species all have the same mobility in bulk solution (the Example on page 120), leading to an overall conductivity \( \kappa \) that doesn’t care which species is moving.

\(^9\)The reasoning is similar to Section 8.5.2.

\(^10\)Also see Section 10.3.4f (page 158).

\(^11\)Nor from a superconducting quantum interference device!
11.3 Neurons

The key feature appears in the last column of this table: There is no value of $\Delta \psi$ that even approximately satisfies all three of these ion species. In fact, resting neurons are polarized with $\Delta \psi$ negative. Sodium is therefore far out of equilibrium under those conditions.

In its resting state, the neuron creates and maintains these nonequilibrium concentrations by continuously pumping ions across its membrane, but we don’t need to worry about that. Even when we shut down a living cell’s metabolism (and hence its ion pumps), it still preserves the preceding values of ion concentrations for several minutes, because the interior and exterior are large reservoirs and membrane conductances are small. During that time, the neuron’s axon can conduct action potentials, and it otherwise behaves electrically like a normal cell’s axon. The pumps just set up and maintain the conditions given in the table.

We can summarize the preceding discussion with a little circuit diagram representing the contribution of one species to the current through a patch of membrane:

$$I_{r, \ell} = j_{r, \ell} \Delta \Sigma$$

$$R_{r, \ell} = 1/(g_{r, \ell} \Delta \Sigma)$$

$$\Delta \psi = \psi^{\text{in}} - \psi^{\text{out}}$$

Placing the resistor and battery symbols in series, as shown, encodes the fact that current is driven by the difference between actual $\Delta \psi$ and the Nernst potential for this species (Equation 11.9).

11.3.3 Linear cable equation for an axon

Let’s see how the preceding considerations affect signal propagation along a “resting” axon, that is, one in steady state. Each ion species makes its own contribution to the electric current, so we can simply represent the driving forces and conductances by three modules in parallel.\(^{12}\)

\(^{12}\)The circuit diagram also correctly represents the fact that all ion species share the same exterior and interior values of the electrostatic potential at any position $x$. 
Because we assume zero external resistance, \( \psi_{\text{out}} \equiv 0 \) and \( \Delta \psi = \psi_{\text{in}} \). The dashed arrow reminds us that, although the resting membrane transmits no net current, still individual ion species are flowing.

Section 11.3.1 suggested that the distributed free energy source, symbolized by the battery symbols in the diagram, could regenerate a disturbance as it travels along the axon.

**Your Turn 11F**

To investigate, first show that the entire preceding diagram can be equivalently replaced by a single resistor/battery unit, and find formulas for the effective overall battery potential \( \psi^0 \) and radial resistance \( R_{r, \text{tot}} \). Explain the sense in which “the ion species with the biggest conductance gets the biggest vote determining the membrane potential.”

Your answer involves the overall conductance per area of a resting axon membrane. For squid giant axon, a typical magnitude is \( g_{\text{tot}} = \sum g_\ell \approx 5 \text{ } \Omega^{-1} \text{ m}^{-2} \).

With this insight, we see that the axon’s overall diagram is almost exactly the same as the one in Figure 11.1, just with the addition of a battery in each module. Thus, the needed modification to the linear cable equation amounts to introducing \( \psi^0 \):

\[
\kappa \pi a^2 \frac{\partial^2 \psi_{\text{in}}}{\partial x^2} = 2\pi a \left( g_{\text{tot}}(\psi_{\text{in}} - \psi^0) + c \frac{\partial \psi_{\text{in}}}{\partial t} \right).
\]

We can then eliminate the battery term altogether by changing variables to \( v = \psi_{\text{in}} - \psi^0 \):

\[
(\lambda_{\text{cable}})^2 \frac{\partial^2 v}{\partial x^2} - \tau_{\text{cable}} \frac{\partial v}{\partial t} = 0.
\]  

(11.10)

Here the space constant and time constant are defined as before. We conclude that small disturbances from resting behavior are governed by exactly the same equation as the one we found for a cable (Equation 11.7, page 172).

Some illustrative numerical values are revealing: Taking \( a = 0.5 \text{ mm} \), \( g_{\text{tot}} \approx 5 \text{ } \Omega^{-1} \text{ m}^{-2} \), \( c \approx 1 \text{ } \mu \text{F} \text{ cm}^{-2} \), and \( \kappa \approx 3 \text{ } \Omega^{-1} \text{ m}^{-1} \) yields

\[
\lambda_{\text{cable}} \approx 12 \text{ mm}, \quad \tau_{\text{cable}} \approx 2 \text{ ms}.
\]

A signal won’t get from your spinal cord to your big toe if it dies out in twelve millimeters!
We seem to have hit an impasse. All that stored electrochemical energy seems unable to affect the propagation of a disturbance—it dropped out of the equation, which has the same disappointing solutions as before! Indeed, experimentally that’s the observed behavior for weak disturbances. For example, when we inject a subthreshold charge into the axon, we do find passive spread, which in this context is also called “electrotonus.” For the more spectacular action potential, we must look for another physical idea. And Section 11.3.1 suggested where to look: at the membrane.

11.3.4 Threshold behavior foreshadows a role for nonlinearity

That key word threshold in the preceding paragraph is a big clue. Linear equations, such as the linear cable equation, don’t exhibit threshold behaviors. We need to look for something nonlinear.

Experimentally, the resting membrane potential $\psi^0$ in squid axon is $\approx -50\text{mV}$. This is not far from the Nernst potential of potassium ions given in the earlier table. That coincidence suggests one possible interpretation: In the resting state, the conductance for potassium ions is much bigger than that for sodium ions.

When an action potential travels along the membrane, the membrane potential locally and temporarily shoots up to something more like $+40\text{mV}$. This is not so different from the Nernst potential of sodium, again suggesting an interpretation:

*The conductance for sodium ions briefly overtakes that for potassium, and a resulting ion flow tries to establish the sodium Nernst potential as the new steady membrane potential.*

In fact, Hodgkin and B. Katz had previously found that during an action potential, the conductances do change momentarily from their resting values, which are

\[
g_{K^+}^0 \approx 2g_{Cl^-}^0 \approx 3.2\text{ }\Omega^{-1}\text{m}^{-2}\text{ but } g_{Na^+}^0 \approx 0.08g_{Cl^-}^0. \quad \text{(resting)} \quad (11.13)
\]

A modern estimate of the momentary values is

\[
g_{K^+}' \text{ and } g_{Cl^-}' \text{ unchanged but } g_{Na^+}' \approx 160g_{Cl^-}'. \quad \text{(at the action potential peak)} \quad (11.14)
\]

What could change the ion conductance of sodium in just the right way? Hodgkin and Huxley realized that even a few millivolts across a nanometer-thick membrane amounts to a huge electric field, which could tug on charged residues in the proteins making up an ion channel. With the appropriate arrangement, a reversal in the direction of that tugging could mechanically pull open a channel that was normally closed. Hodgkin and Huxley therefore proposed that the conductance of the membrane to specific ions is voltage-dependent: We must use a function of potential $g_{Na^+}(\Delta\psi)$ in the cable equation. The hypothesized voltage gating modifies the cable equation to one that is nonlinear in $\psi$. Interesting things can happen with nonlinearity.

In particular, suppose that depolarization (making $\Delta\psi$ less negative than usual) causes sodium channels to open. Then a localized electrical disturbance that depolarizes a patch of membrane lets sodium ions rush in, which further depolarizes that patch.
The disturbance can then spread diffusively to a neighboring region, where the same sequence is repeated. Thus, the “resting” axon is actually poised to release stored free energy. Perhaps a disturbance at one end could lead to a propagating wave of depolarization, just as lighting a fuse leads to a propagating wave of combustion in some Hollywood blockbuster: Stored chemical energy is released in a controlled way, leading to a flame front that self-regulates to move at constant speed.

Does it really work? See Chapter 12.

FURTHER READING

**Semipopular:**

Undersea telegraph cables: Bodanis, 2005.
- https://en.wikipedia.org/wiki/Cable_theory
- https://en.wikipedia.org/wiki/William_Thomson,_1st_Baron_Kelvin#Transatlantic_cable
- https://en.wikipedia.org/wiki/Submarine_communications_cable#Bandwidth_problems


**Intermediate:**


**Technical:**


Gutted axon experiment: Baker et al., 1962.

**PROBLEMS**

11.1  *Fate of a wave*  
[[Not ready]]
CHAPTER 12

Vista: The Action Potential in Neurons

12.1 FRAMING: NONLINEARITY

Chapters 8 and 11 foreshadowed what we’d like to understand: Although a neural axon consists of a conducting interior wrapped in an insulator and bathed in a conducting solution, much like the early undersea cables, somehow the axon transmits signals over distances much longer than its diameter without amplitude loss nor waveform degradation—unlike those cables. Chapter 11 told us where to look for new physics (in the cell membrane), then suggested that we abandon the ohmic hypothesis, which states that all membrane conductances are fixed,$^1$ in favor of something more subtle: The observed temporary reversal of the sign of the membrane potential both reflects a sudden increase in $g_{Na}^+$ (Equation 11.14 instead of 11.13) and enhances that increase, via voltage gating. Thus, $g_{tot}$ temporarily becomes dominated by the sodium contribution instead of by potassium. This change counteracts the dissipative damping by driving the membrane potential still further away from the potassium Nernst potential and toward that of sodium (Your Turn 11F, page 178), thus continually extending the action potential to that it travels along the axon.

It’s time to see whether this nice story really works. We will follow pioneering work by several groups, shortly before and after the Second World War, who characterized real membrane behavior instead of assuming that it was ohmic, then fed the resulting phenomenological model of membrane conductance into a revised cable equation, whose solutions indeed resemble the action potential.

Electromagnetic phenomenon: A nerve axon carries signals that preserve their form and amplitude as they travel long distances.

Physical idea: Voltage-gating creates a nonlinearity in the cable equation, allowing continuous regeneration of a signal that would otherwise die out.

12.2 THE TIME COURSE OF AN ACTION POTENTIAL CONFIRMS THE HYPOTHESIS OF NON-OHMIC CONDUCTANCE

This chapter introduces a lot of notation. For reference, Table 12.1 lists some symbols already defined, and other defined below.

We can show directly from experimental data that the ohmic hypothesis breaks down. The observed action potential is a traveling wave of fixed waveform, moving at a constant speed $v$. (We would eventually like to understand why that should be so, but for now we

$^1$ Equation 11.9 (page 176).
regard it as an empirical fact.) For such a function, the entire history \( \psi_{\text{in}}(x, t) \) is completely known once we measure its time course at one point.\(^2\) Thus,

\[
\psi_{\text{in}}(x, t) = \tilde{\psi}(t - (x/\sigma)),
\]

where the waveform \( \tilde{\psi}(t) \equiv \psi_{\text{in}}(0, t) \) is shown in Figure 12.1b. Hence,

\[
\frac{\partial \psi_{\text{in}}}{\partial x} = -\frac{1}{\sigma} \left. \frac{d \tilde{\psi}}{dt} \right|_{t=(x/\sigma)},
\]

by the Chain Rule of calculus.

Instead of assuming an ohmic membrane conductance, as in Chapter 11, we can now test the ohmic hypothesis by determining the actual outward charge flux from experimental data. To do this, rearrange Equations 11.1–11.4 (page 172) to find the transmembrane (radial) charge flux:

\[
j_r = \frac{I_r - C \partial \psi / \partial t}{2\pi a \Delta x} = -\frac{1}{2\pi a} \left( -\frac{\partial}{\partial x} \frac{\partial \psi_{\text{in}}}{\partial x} \frac{x \pi a^2}{\Delta x} \right) - C \frac{\partial \psi_{\text{in}}}{\partial t}.
\]

For a traveling wave, Equation 12.2 lets us rephrase in terms of the measured time course \( \tilde{\psi}(t) \):

\[
j_r = \frac{\alpha x}{2\sigma^2} \frac{d^2 \tilde{\psi}}{dt^2} - C \frac{d \tilde{\psi}}{dt}.
\]
12.2 The Time Course of an Action Potential Confirms the Hypothesis of Non-Ohmic Conductance

Table 12.1: **Symbols used in this chapter.** See also Appendix B.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>axon radius</td>
</tr>
<tr>
<td>(\kappa)</td>
<td>conductivity of interior fluid</td>
</tr>
<tr>
<td>(g_\ell)</td>
<td>membrane conductance per area for ion species (\ell); (g_{\text{out}}), total</td>
</tr>
<tr>
<td>(C)</td>
<td>capacitance per area of membrane</td>
</tr>
<tr>
<td>(\psi_{\text{in}}(x, t))</td>
<td>interior electric potential</td>
</tr>
<tr>
<td>(\psi_{\text{out}})</td>
<td>exterior electric potential, (= 0) in our simplified model so (\Delta \psi = \psi_{\text{in}} - \psi_{\text{out}} = \psi_{\text{in}})</td>
</tr>
<tr>
<td>(\tilde{\psi}(t))</td>
<td>waveform of a traveling wave (Equation 12.1)</td>
</tr>
<tr>
<td>(\psi_{\text{Nernst}})</td>
<td>Nernst potential for ion species (\ell) (Section 10.2.1, page 139)</td>
</tr>
<tr>
<td>(\psi^0)</td>
<td>combination of Nernst potentials giving the resting potential (Your Turn 11F, page 178)</td>
</tr>
<tr>
<td>(v(x, t))</td>
<td>depolarization ((\Delta \psi) shifted by (\psi^0)); (v_1) and (v_2), special fixed-point values (Figure 12.4)</td>
</tr>
<tr>
<td>(\tilde{v}(t))</td>
<td>depolarization waveform of a traveling wave; (\tilde{v}), dimensionless rescaled form</td>
</tr>
<tr>
<td>(j_r)</td>
<td>radial charge flux (current per area) actually passing through axon membrane; (j_{\text{r,rf}}), its component from ion species (\ell)</td>
</tr>
<tr>
<td>(\lambda_{\text{cable}}, \tau_{\text{cable}})</td>
<td>space constant and time constant of axon (Equation 11.6, page 172)</td>
</tr>
<tr>
<td>(\vartheta)</td>
<td>speed of a traveling wave (Your Turn 11E, page 174)</td>
</tr>
</tbody>
</table>

Figure 12.2: [Sketch graphs.] **Membrane current inferred from action potential.** (a) The sketch shows the membrane potential \(\tilde{\psi}(t)\), measured at a fixed location \(x = 0\). \(\tilde{v}(t)\) refers to the difference between the membrane potential and its resting value \(\psi^0\). The dashed lines are six particular moments of time discussed in the text. (b) Reconstruction of the total membrane current from (a), using Equation 12.3. An ohmic stage \(A\) gives way to another stage \(B\). In \(B\), the membrane potential continues to rise but the current falls and then reverses; this is non-ohmic behavior. [Adapted from Benedek & Villars, 2000.]

The parameters \(a\), \(\kappa\), \(\vartheta\), and \(C\) in Equation 12.3 are all experimentally measurable, so applying it to the time course of an action potential will give us the corresponding time course for the membrane current (Figure 12.2). We can understand this result graphically,

---

2As in Chapter 11, we are considering a simplified model where the potential is everywhere zero outside the cable.
without any calculations. Note that the membrane current is particularly simple at the inflection points of panel (a) (the dashed lines labeled 1, 3, and 5): Here the first term of Equation 12.3 equals zero, and the sign of the current is opposite to that of the slope of \( \dot{\psi}(t) \). Similarly, at the extrema of panel (a) (the dashed lines labeled 2 and 4), we find that the second term of Equation 12.3 vanishes: Here the sign of the current is that of the curvature of \( \dot{\psi}(t) \), as shown in panel (b). With these hints, we can work out the sign of \( j_r \) at the points 0–6 and interpolate [panel (b)].

Comparing the two panels of Figure 12.2 shows what is happening during the action potential. Initially (stage A), the membrane conductance may indeed be ohmic: The cell’s interior potential begins to rise above its resting value, thereby driving an outward current flux, as predicted from your calculation of the potential of three resistor–battery pairs (Your Turn 11F, page 178). But when the membrane has depolarized by about 10 mV, something strange begins to happen (stage B): The potential continues to rise, but the net current falls. The ohmic hypothesis cannot account for that behavior.

Idea 11.12 made the key point needed for understanding the current reversal, in terms of a switch in the membrane’s permeabilities to various ions. Net current flows across a membrane whenever the actual potential difference \( \psi_{\text{in}} \) deviates from the “target” value. But the target value itself depends on the membrane conductances. If these suddenly change from their resting values, then so will the target potential; if the target switches from being more negative than \( \psi_{\text{in}} \) to more positive, then the membrane current will change sign. Because the target value is dominated by the Nernst potential of the most permeant ion species,\(^3\) we can explain the current reversal by supposing that the membrane’s permeability to sodium increases suddenly during the action potential.

So far, we have done little more than restate Idea 11.12 (page 179). As outlined in Section 11.3.4, Hodgkin and Huxley noted that the increase in sodium ion conductivity does not begin until after the membrane has depolarized significantly (Figure 12.2, stage B), supporting our hypothesis that\(^4\)

\[
\text{Membrane depolarization itself is the trigger that causes the sodium conductance to increase.}
\text{(12.4)}
\]

That is, they suggested that some collection of unknown molecular devices in the membrane (today called ion channels) allow the passage of sodium ions, with a conductance depending on the membrane potential. Idea 12.4 introduces an element of \textbf{positive feedback} into our picture: Depolarization begins to open the sodium gates, a process that increases the degree of depolarization. The increased depolarization opens still more sodium gates; and so on.

The simplest way to implement Idea 12.4 is to modify the ohmic hypothesis (Equation 11.9, page 176) by allowing each of the membrane’s conductances to depend on \( \psi_{\text{in}} \):

\[
j_r = \sum_{\text{species } \ell} (\psi_{\text{in}} - \psi_{\ell}^{\text{Nernst}}) g_{\ell}(\psi_{\text{in}}). \quad \text{prompt voltage-gating hypothesis}
\text{(12.5)}
\]

---

\(^3\)See Your Turn 11F (page 178).

\(^4\)Section 11.3.4 (page 179) introduced this idea.
In this formula, the potential drop $\Delta \psi$ across the membrane equals $\psi_{in}$ because we still neglect any exterior resistance (Figure 10.1, page 140).

The proposal Equation 12.5 certainly has a lot of content, even though we don’t yet know the precise form of the conductance functions appearing in it. For example, it implies that the membrane’s ion currents are still linear in $\ln(out/in)$ if we hold $\psi_{in}$ fixed with an external source but change the concentrations. However, the membrane current is now a nonlinear function of $\psi_{in}$, a crucial point for the following analysis.

Note that Equation 12.5 explicitly assumes that the conductances respond immediately to changes in membrane potential. Real neurons have a time delay, but Section 12.3 will show that even our prompt voltage-gating hypothesis already accounts for much of the phenomenology of the action potential.

12.3 VOLTAGE GATING LEADS TO A NONLINEAR CABLE EQUATION WITH TRAVELING WAVE SOLUTIONS

12.3.1 A purely mechanical system with traveling, solitary waves

We can now return to the apparent impasse reached in our discussion of the linear cable equation (Section 11.3.3): There seemed to be no way for the action potential to gain access to the free energy stored along the axon membrane by the ion pumps. The previous section motivated a proposal for how to get the required coupling, namely, Equation 12.5. However, it left an unanswered question: Who orchestrates the orderly, sequential increases in sodium conductance as the action potential travels along the axon? The full answer to this question is mathematically rather complex, involving multiple channel types and time delays. This section will implement a simplified version, in which we can explicitly solve the equations and see at least the outline of the full answer.

Consider first a mechanical analogy, a chain that progressively shifts from a higher to a lower groove (Figure 12.3a). This system exhibits traveling wave solutions of fixed speed and definite waveform. Some authors call such solutions trigger waves because the system can persist forever in the metastable upper channel, only releasing its stored energy if “triggered.” The initial state is sometimes called an excitable medium. Now we must translate our ideas into the context of axons, and do the math.

12.3.2 Voltage gating leads to bistability

The force needed to pull each successive segment of chain over its potential barrier comes from the previous segment of chain. But that sounds analogous to the proposal in Section 12.2 (page 181) for the axon, which said that even though the resting axon is in a

---

5See Equation 11.9 (page 176).
6Some authors restrict the phrase “excitable medium” to a more elaborate form that “resets” itself after a transient excursion. Real nerve axons have that property (Section 12.4’s, page 192), unlike the simplified models here and in Section 12.3.2.
Figure 12.3: [Schematic.] Mechanical analog of the action potential. A flexible chain lies in a tilted channel, with two troughs at heights differing by $\Delta h$. In the axon context, the upper trough represents the steady or quasisteady state prior to an action potential. (a) An isolated kink will move to the left at a constant speed $\dot{\varphi}$: successive chain elements are lifted from the upper trough, slide over the crest, and fall into the lower trough. (b) A disturbance can create a pair of such kinks if it is above threshold. The two kinks then travel away from each other at speeds $\pm \dot{\varphi}$. Media 4 shows a realization of this system.

stable steady state of the membrane,

- When one segment depolarizes, its depolarization spreads passively to the neighboring segment;
- Once the neighboring segment depolarizes by more than a threshold value, positive feedback sets in, triggering more depolarization; and
- The process repeats, spreading the depolarized region.

We begin by thinking only about the initial sodium influx. Our working hypothesis is that the membrane’s conductance per area for this ion, $g_{Na^+}(v)$, depends on the value of the depolarization $v \equiv \psi_{in} - \psi^0$.

A detailed model would use an experimentally measured form of the function $g_{Na^+}(v)$, as imagined in the dashed line of Figure 12.4a. We will instead use a mathematically simpler form (solid curve in the figure), namely, the quadratic function

$$g_{Na^+}(v) = g_{Na^+}^0 + Bv^2.$$  

Equation 12.7

Here $g_{Na^+}^0$ represents the resting conductance per area and $B$ is a positive constant. Equation 12.7 incorporates the key feature of increasing upon depolarization; moreover, it is always positive, as any conductance must be.

The total charge flux through the membrane, Equation 12.5, is then the sum of all

---

7Your Turn 11F (page 178) introduced the resting potential $\psi^0$; Equation 11.10 (page 178) introduced $v$. Our assumption of prompt response is not fully realistic; thus, our simple model will not capture all the features of real action potentials. See Further Reading for more realistic models.
12.3 Voltage Gating Leads to a Nonlinear Cable Equation With Traveling Wave Solutions

187

Figure 12.4: [Sketch graphs.] Voltage-gating hypothesis. (a) Dashed curve: The conductance $g_{Na^+}$ of an axon membrane to sodium ions, showing an increase as the membrane potential increases from its resting value ($v = 0$). Solid curve: Simplified form for membrane sodium conductance (Equation 12.7). This form captures the relevant feature of the dashed curve, namely, that it increases as $v$ increases and is positive. (b) Current-voltage relation resulting from the conductance model in (a) (Equation 12.9). The special values $v_1$ and $v_2$ are defined in the text. Arrows show the evolution if the potential is slightly disturbed from one of the three fixed points.

the ohmic terms plus the extra sodium contribution:

$$j_r = \left( \sum_{\text{species } \ell} (\psi_{\ell \text{ in}} - \psi_{\ell \text{ Nernst}}^0)g_{\ell}^0 \right) + (\psi_{\text{ in}} - \psi_{\text{ Nernst}}^{Na^+})Bv^2. \quad (12.8)$$

As in Your Turn 11F (page 178), the first term in Equation 12.8 can be rewritten as $g_{\text{ tot}}^0v$.

Letting $H$ denote the constant $\psi_{\text{ Nernst}}^{Na^+} - \psi^0$, we can also rewrite the last term as $(v - H)Bv^2$, obtaining

$$j_r = vg_{\text{ tot}}^0 + (v - H)Bv^2. \quad (12.9)$$

Figure 12.4b show the behavior of our model. The three points where the membrane current $j_r$ is zero are especially significant. Equation 12.9 says that these points are the roots of a cubic equation. We write them as $v = 0, v_1, \text{ and } v_2$, where $v_1$ and $v_2$ equal $\frac{1}{2}(H + \sqrt{H^2 - 4g_{\text{ tot}}^0/B})$, respectively. At small depolarization ($v \approx 0$), the sodium conductance stays small, so in that situation the last term of Equation 12.9 is negligible. A small positive $v$ then gives small positive (outward) current, as expected: We are in the ohmic regime (stage A of Figure 12.2). The outward flow of charge tends to reduce $v$ back toward zero.

A further increase of $v$, however, opens the voltage-gated sodium channels, eventually reducing $j_r$ to zero, and then below zero as we pass the point $v_1$. Now the net inward flow of charge tends to increase $v$, giving positive feedback—an avalanche. Instead of returning to zero, $v$ then increases toward the other root, $v_2$. At still higher $v$, we once again get a positive (outward) current, as the large outward electric force on all the ions finally overcomes the entropic tendency for sodium to flow inward.

In short, our model displays threshold behavior: Small disturbances get driven back to $v = 0$, but above-threshold disturbances drive to the other\textsuperscript{8} stable fixed point $v_2$. Our program is now to make the appropriate changes to the steps in Section 11.3.3 (page 177).

\textsuperscript{8}The value $v_1$ is an unstable fixed point, because small deviations above or below it get driven to larger deviations (Figure 12.4b).
12.3.3 The nonlinear cable equation

We first substitute Equation 12.9 into the charge balance equation (Equation 11.5, page 172). Some algebra shows that \( v_1 v_2 = \frac{g_{tot}}{B} \), so the equation becomes

\[
(\lambda_{cable})^2 \frac{\partial^2 v}{\partial x^2} - \tau_{cable} \frac{\partial v}{\partial t} = \frac{v(v - v_1)(v - v_2)}{(v_1 v_2)}. \tag{12.10}
\]

Unlike the linear cable equation, Equation 12.10 is not equivalent to a diffusion equation. In general, it’s very difficult to solve nonlinear, partial differential equations like this one. But we can simplify things, because our main interest is in finding whether there are any traveling wave solutions to Equation 12.10. Following the discussion leading to Equation 12.3, we can represent a wave traveling at speed \( \vartheta \) by a function \( \vartheta(t) \) of one variable, via \( v(x, t) = \vartheta(t - (x/\vartheta)) \). Substituting into Equation 12.10 leads to an ordinary differential equation:

\[
\left( \frac{\lambda_{cable}}{\vartheta} \right)^2 \frac{d^2 \vartheta}{dt^2} - \tau_{cable} \frac{d\vartheta}{dt} = \frac{\vartheta(\vartheta - v_1)(\vartheta - v_2)}{v_1 v_2}. \tag{12.11}
\]

We can tidy up the equation by defining the dimensionless quantities \( \tilde{\vartheta} \equiv \vartheta/v_2 \), \( \gamma \equiv -\vartheta/\lambda_{cable} \), \( s \equiv v_2/v_1 \), and \( \gamma \equiv \tau_{cable} \gamma/\lambda_{cable} \), finding

\[
\frac{d^2 \tilde{\vartheta}}{dy^2} = -\gamma \frac{d\tilde{\vartheta}}{dy} + s \tilde{\vartheta}^3 - (1 + s) \tilde{\vartheta}^2 + \tilde{\vartheta}. \tag{12.12}
\]

12.3.4 Solution

You could enter Equation 12.12 into a computer-math package, substitute some reasonable values for the parameters \( b \) and \( s \), and look at its solutions. But it’s tricky: The solutions are badly behaved (they blow up) unless you take \( \gamma \) to have one particular value (see Figure 12.5). This behavior is not surprising in the light of Figure 12.3: Our mechanical analog system selects one definite value for the pulse speed (and hence \( \gamma \)). You’ll find in Problem 12.1 that choosing

\[
\vartheta = \pm \frac{\lambda_{cable}}{\tau_{cable}} \sqrt{\frac{2}{s} \left( \frac{s}{2} - 1 \right)} \tag{12.13}
\]

does yield a traveling wave solution (the solid curves in Figure 12.5).

12.3.5 Interpretation

The hypothesis of voltage gating, embodied in the nonlinear cable equation, has led to the appearance of traveling wave solutions of definite speed and waveform. In particular, the amplitude of the traveling wave is fixed: It smoothly connects the two stable fixed-point values 0 and \( v_2 \) (Figure 12.4). We cannot excite such a wave with a very small disturbance, because for small enough \( v \), the nonlinear cable equation is essentially the same as the

---

\(^9\text{Contrast Section 11.3.3 (page 177).}\)
12.3 Voltage Gating Leads to a Nonlinear Cable Equation With Traveling Wave Solutions

![Figure 12.5: Traveling wave solution to the nonlinear cable equation](image)

**Figure 12.5**: [Mathematical functions.] **Traveling wave solution to the nonlinear cable equation** (see Problem 12.1). The membrane potential relative to rest, \( v(x,t) \), is shown as a function of time at three different fixed locations (three solid curves). Points at larger \( x \) see the wave go by at later times, so this wave is traveling in the +\( x \) direction. The parameter \( s \equiv v_2/v_1 \) has been taken equal to 3 for illustration. This simplified model qualitatively reproduces the leading edge of the action potential (Figure 12.2a). The dashed line shows a solution to Equation 12.11 with a value of the front velocity \( \delta \) different from that in Equation 12.13; this solution is singular. Time is expressed as multiples of \( \lambda_{\text{cable}}/\delta \). The depolarization \( v \) is expressed as multiples of \( v_2 \).

linear one (Equation 11.7, page 172), whose solution we have already seen corresponds to passive, diffusive spreading (electrotonus). Thus,

- **Voltage gating still leads to the observed graded, diffusive response for stimuli below a threshold, but**

- **An above-threshold, depolarizing stimulus yields a large, fixed-amplitude response.**

- **The above-threshold response can take the form of a traveling wave of fixed shape and speed.**

Our model, a mathematical embodiment of Idea 12.6, has captured many of the key features of real action potentials. We didn’t prove that the wave rapidly forgets the precise nature of its initial stimulus, remembering only whether it was above threshold or not, but such behavior should seem reasonable in the light of the mechanical analogy (Figure 12.3). We also get a quantitative prediction from Equation 12.13: The velocity \( \delta \) is proportional to \( \lambda_{\text{cable}}/\tau_{\text{cable}} = \sqrt{\alpha x_{\text{ion}}^2/(2C^2)} \). Thus, the model predicts that if we examine a family of axons of the same general type, with the same ion concentrations, we should find that the pulse speed varies with axon radius as \( \delta \propto \sqrt{a} \). This prediction is roughly borne out in experimental data.\(^{10}\) Moreover, the general magnitude of the pulse speed is approximately \( \lambda_{\text{cable}}/\tau_{\text{cable}} \). For the squid giant axon, our estimates give this quantity as about \( 12 \text{ mm}/2 \text{ ms} = 6 \text{ m s}^{-1} \), a value within an order of magnitude of the measured action potential speed of about \( 20 \text{ m s}^{-1} \).

In the mechanical analogy, the wave speed is proportional to the density of stored

---

\(^{10}\) Our result applies only to "unmyelinated" axons.

An axon carries electrochemical signals, but does so far more slowly than a coaxial cable.
energy divided by a friction constant. Both $k$ and $g_{\text{tot}}$ are inverse resistances, so $\sqrt{kg_{\text{tot}}}$ in our expression for $\dot{\vartheta}$ is indeed an “inverse friction”-type constant. In addition, the formula $\mathcal{E}/\Sigma = \frac{1}{2}q^2/(C\Sigma^2)$ for the stored electrostatic energy in a capacitor shows that it is proportional to $1/\mathcal{E}$. Thus, the prefactor in Equation 12.13 has the general form expected from the mechanical analogy.

Section 12.3’ (page 191) gives more details about how the nonlinear cable equation determines the speed of its traveling wave solution.

12.4 PLUS ULTRA

Although squid and humans diverged evolutionarily a very long time ago, the main outlines of their signaling mechanisms are remarkably similar. Indeed, the action potential, so critical for all multicellular animals, has turned out to be a physics problem. That is, a few classes of actors, obeying rules that can be characterized with simple functions, could be assembled as elements of a mathematical model that made testable, quantitative predictions about experiments different from the ones that characterized the elements.

Physicists like ideas with even wider applicability than the systems for which they were initially developed. Indeed, Hodgkin and Huxley’s work may be regarded as the opening moves in the vast field of excitable media, spanning from nerves to flame fronts to territorial invasions of species, and much more.11

Section 12.4’ (page 192) mentions more details about realistic axon models.

FURTHER READING

Semipopular:
Adee, 2023.
“Dancing Zombie Squid Explained” www.youtube.com/watch?v=JGPfSSU1ReM.

Intermediate:
Neurons: Raman & Ferster, 2021; see also Phillips et al., 2012.

Technical:
Even plants have action potentials: Hedrich & Neher, 2018.

11See Media 5 for one example.
12.3 Velocity selection in more general models

Problem 12.1 pulls an exact analytic solution out of a hat. The fact that any solution exists may seem a miracle, a pathology of our very specific illustrative form for the equations. To see that the behavior we found is actually general, here is a physically inspired argument. Begin with Equation 12.10 (page 188). We are interested in traveling wave solutions, representing the situation where the initial resting state \( (v = 0) \) is invaded by the excited state \( (v = v_2) \). Thus, we explore trial solutions of the form \( v(t, x) = \bar{v}(t - x/\partial_\theta) \) where \( \bar{v}(t) \) interpolates between 0 and \( v_2 \) as \( t \) runs from \( -\infty \) to \( \infty \).

As \( t \to -\infty \) and \( \bar{v}(t) \to v_2 \) as \( t \to +\infty \). The wave velocity \( \partial_\theta \) is not known yet, but let’s look for a solution with positive velocity (that is, moving to the right).

To get a recognizable equation, first change variables:

\[
y = -|\partial_\theta|t / \lambda_{\text{cable}} \quad \frac{d}{dt} = -|\partial_\theta| \frac{d}{dy}
\]

As a function of \( y \), our desired behavior is that \( \partial(y) \to 0 \) as \( t \to +\infty \) and so on. Now multiply both sides of Equation 12.11 by \( d\partial dy \) and rearrange to find

\[
\frac{d}{dy} \left[ \frac{1}{2} \left( \frac{d\partial}{dy} \right)^2 + U(\partial) \right] = -\gamma \left( \frac{d\partial}{dy} \right)^2 \tag{12.15}
\]

where

\[
U(\partial) = -\frac{1}{v_1 v_2} \left[ \frac{1}{4} \partial^4 - \frac{1}{3} (v_1 + v_2) \partial^3 \right] - \frac{1}{2} \partial^2, \quad \text{and} \quad \gamma = \tau_{\text{cable}} |\partial| / \lambda_{\text{cable}}. \tag{12.16}
\]

Similar manipulations continue to work for any voltage gating function with the general form in Figure 12.4 (page 187), but we’ll continue to use the illustrative quadratic function. The key point is that the zeros of the current flux function correspond to extrema of the function \( U \).

We can understand the behavior of Equation 12.15 by a mathematical analogy to mechanics. Think of a roller-coaster car, rolling with “position” \( \partial \) at “time” \( y \) on a “potential energy” landscape \( U \). On the left side of the equation, we have the time derivative of “kinetic plus potential energy.” On the right side we have “frictional loss” (in an imagined world where roller coasters are immersed in a viscous fluid). Our roller coaster starts at “time” \( y \to -\infty \) on top of a hill \( (\partial = v_2) \). After a long wait (set by the size of an initial small perturbation), it rolls off the hill toward the left. To see what happens next, examine Figure 12.6.

To get a value for \( B \) in Equation 12.7 (page 186), note that the sodium conductance rises from negligible to about 52 times the resting total conductance as membrane potential rises from resting to about 40 mV greater than that. These values and Equation 12.7 gave \( (B / g_0)^2 (40 \text{mV})^2 = 52 \), and then \( v_1 = 0.3 \text{mV} \) and \( v_2 = 100 \text{mV} \). Figure 12.6 shows the resulting quartic function \( U \) (Equation 12.16).

The generic behavior that ensues is that the roller coaster either rolls to \( \partial \to -\infty \) or comes to rest at the shallow trough at \( v_1 \), perhaps after some oscillations. Neither of those possibilities is what we want. But we get to select the value of the friction constant \( \gamma \), because it contains the unknown propagation speed \( \partial_\theta \). Imagine this system physically. If you bump it off the higher peak, it will roll down, gaining kinetic energy though losing some to friction. If the friction is too great, it will end up at \( \partial = v_1 \). If the friction is too small, it will overshoot \( \partial = 0 \) and end up at minus infinity. But in between, there will be a just right value of friction that glides our roller coaster precisely to a halt at the top of the lower hill \( (\partial = 0)! \)

---

\(^{12}\)The value of \( v_2 \) is higher than the actual maximum of an action potential, but we only want the leading edge; we are neglecting the later potassium currents and sodium channel inactivation that eventually cut off the rise of potential.
Figure 12.6: [Mathematical functions.] **Effective potential for Equation 12.16.** For illustration, total resting membrane conductance was set to 5 Ω⁻¹ m⁻² and \( H = \psi_N^0 - \psi^0 = 100 \, \text{mV} \). (a) The desired solution starts at \( \bar{v} = v_2 \). (b) Magnified form of the left side of (a). The desired solution coasts to a stop at the left hilltop (\( \bar{v} = 0 \)).

### 12.4’a More detailed models

Actually, all types of ion channels potentially have voltage-dependent conductance, not just sodium. We focused on sodium because it’s responsible for the switch to the high-conductance state (leading edge of an action potential). Later, potassium channels open, leading to the lagging front mentioned in Figure 8.6a (page 124), and later still the sodium channels “inactivate,” even if the membrane remains depolarized. Both potassium channel opening and sodium channel inactivation contribute to shutting down the conduction and returning the axon to its resting state after a transient overshoot (afterhyperpolarization; see Figure 12.1, page 182).

The word “later” reminds us that channels do not actually respond promptly to the current membrane potential as assumed in Equation 12.5 (page 184); they require time to overcome activation barriers and snap open. Instead of our prompt voltage-gating hypothesis, Hodgkin and Huxley acknowledged that individual channels open and close at random times. Really it is the rate constants for the opening and closing transitions that are functions of membrane potentials. In other words, Hodgkin and Huxley upgraded from prompt voltage gating to a full kinetic model.

### 12.4’b FitzHugh–Nagumo model

Introducing realistic kinetics leads to a much more complicated system. There is a useful intermediate theory, however, the **FitzHugh–Nagumo model**, in which the fastest ion channels (sodium) are assumed to respond promptly, and slower dynamics are merged into a single independent dynamical variable (Keener & Sneyd, 2009).

### 12.4’c Solitons

A nonlinear traveling wave is sometimes called a “solitary wave” or **soliton**. Here is will explore another context for them. The cables that send Internet between cities are not wires at all, but optical...
fibers. They can be formulated with ultra-low loss (absorption), but they still suffer from optical dispersion (mushing-out of signals). Modern optical fibers have nonlinear optical effects that make them transmit those ones and zeros as solitons, preserving their shape for hundreds of kilometers. [[Not ready]]
PROBLEMS

12.1  *Analytic solution for simplified action potential*
Show that the function \( \phi(y) = \frac{1}{\gamma} \left(1 + e^{\alpha y}\right)^{-1} \) solves Equation 12.12 (page 188), if we take the parameter \( \gamma \) to be given by \( \sqrt{2/\phi \left(\frac{\alpha}{2} - 1\right)} \). Hence derive the speed of the action potential (Equation 12.13, page 188). \( \alpha \) is another constant, which you are to find.

12.2  
[[Not ready]]
Examples of 3-Tensors in Physics

Rather than propose a new theory or unearth a new fact, often the most important contribution a scientist can make is to discover a new way of seeing old theories or facts.

— Richard Dawkins

13.1 FRAMING: ANISOTROPY

Ultimately our goal is to define and exploit a construction called “4-tensors.” Before we go there, let’s see some examples that may be familiar to you, at least implicitly, from previous work. Like the man who discovered he had been speaking prose all his life, you are probably already familiar with some tensors.

In fact, Chapter 3 already informally introduced a useful mathematical object called the quadrupole moment and introduced the term “tensor.” We now step back and generalize this concept, still informally, then more systematically in later chapters. This chapter discusses tensors in three-dimensional space, abbreviated 3-tensors.

Electromagnetic phenomenon: Molecular polarizability is in general anisotropic.

Physical idea: Molecular architecture can dictate specific directions of greater compliance; nonuniform distribution of charge in a molecule can give particular “handles” for electric fields to push or pull.

13.2 RANK ZERO; RANK ONE

A “3-tensor of rank 0” (also called a 3-scalar) is just a fancy term for a physical quantity that is a single number. More precisely, its value is the same when we work in any cartesian coordinate system. Electric charge is an example. It doesn’t need any coordinate index (that is, it carries zero indices).

A “3-tensor of rank 1” is just a fancy term for what we have been calling a vector. It is a geometrical object (an “arrow,” with direction and magnitude). It can be specified by giving three numbers (its components), after first choosing a cartesian coordinate system on space. The three components \( \{ r_i \} \) of a 3-vector \( \vec{r} \) carry one coordinate index, hence the name “rank 1.” We say that the components represent the vector in the chosen coordinate system.

---

1 Tensor calculus was developed gradually, starting with G. Ricci-Curbastro around 1890.
2 Also, Section 7.2.3 (page 111) constructed the fundamental forms of a two-dimensional surface; they are 2-tensor fields on a curved space.
Chapter 13  Examples of 3-Tensors in Physics

We also can think of a tensor of rank 1 as a function that eats a vector, returns a scalar, and is linear. For example, the projection \( f(\vec{v}) = \vec{a} \cdot \vec{v} \) is such a function, where \( \{\vec{a}_i\} \) is a set of three constants. Either way, we need three numbers to specify an object in this class.

When we use explicit index notation, we will usually drop summation symbols on tensor indices, relying on the convention that a repeated index is to be summed unless otherwise noted. Thus, \( \vec{a}_i \vec{v}_i \) is shorthand for \( \sum_i \vec{a}_i \vec{v}_i \) and so on.

13.3 RANK TWO

Three-tensors of rank 2 play several closely related roles in pre-Einstein physics.\(^3\)

- A tensor may specify a linear, vector-valued function of another vector.
- A tensor may specify a scalar-valued function of a vector that is quadratic in the vector’s components.
- A tensor may specify a scalar-valued function of two vector arguments that is linear in each one. For example, the function may be the second-order part of a Taylor expansion.

The following sections give details and concrete examples. As in the rank-1 case, we’ll also see that in either interpretation, a rank-2 tensor can be represented by an array of ordinary numbers (its components).

13.3.1 A tensor can specify a vector-valued, linear function of vectors

When your auto mechanic says that your car’s wheels need to be “balanced,” what do they mean? Clearly, it’s desirable for the wheel’s center of mass (CM) to lie on the axle. Otherwise, spinning the wheel would require the CM to move in a circular orbit. Circular motion implies acceleration, which requires a force. So as the wheel spins, the axle would be constantly subjected to sideways forces, which would wear out the bearings if not corrected.

But there is more. Suppose that the CM does lie on the axle, but the wheel is bent, so that its axis of symmetry, if it has one, does not coincide with the axle. Spinning the wheel about the axle, even at constant angular velocity, then requires torque, which is also bad for the car. Let’s see how to quantify this effect.

When a rigid body spins about any axis with angular frequency \( \omega \), we define the angular velocity \( \vec{\omega} \) as the product of \( \omega \) with a unit vector pointing along that axis, with sign chosen by a right-hand rule. Suppose that the body is subdivided into small masses \( m_r \) momentarily located at positions \( \vec{r}_r \) relative to a reference point fixed in the body. Then the resulting angular momentum \( \vec{L} \) has cartesian components that are linear

---

\(^3\)Physicists began to use tensors in four or more dimensions after H. Minkowski reformulated Einstein’s ideas in this language (Chapter 32).
functions of those of \( \mathbf{a} \), and therefore may be written\(^4\) as \( \mathbf{L}_i = \mathbf{\hat{J}}_{ij} \mathbf{a}_j \), or more simply \( \mathbf{L} = \mathbf{\hat{J}} \cdot \mathbf{a} \). Here the **moment of inertia tensor** \( \mathbf{\hat{J}} \) is a set of quantities with two indices, and hence is said to have rank 2.

We can compactly express \( \mathbf{\hat{J}} \) by the formula

\[
\mathbf{\hat{J}} = \sum \limits_{\mathbf{r}} m_{\mathbf{r}} [[\mathbf{r}, \mathbf{r}]^{1/2} \mathbf{\hat{I}} - \mathbf{r}(\mathbf{\hat{r}}_{\mathbf{r}}) \hat{r}(\mathbf{\hat{r}}_{\mathbf{r}})].
\]

(13.1)

The symbol \( \mathbf{\hat{I}} \) refers to the **identity tensor**, whose entries (components) are the identity matrix. Equivalently, \( \mathbf{\hat{I}} \) can be regarded as a machine that eats a vector and returns that same vector, which certainly is a linear operation. The second term of Equation 13.1 is called a **dyad product**,\(^5\) defined as the tensor that eats any vector \( \mathbf{a} \) and returns a rescaled version of a fixed vector \( \mathbf{v} \):

\[
(\mathbf{\hat{v}} \otimes \mathbf{v}) \cdot \mathbf{a} = \mathbf{\hat{v}}(\mathbf{\hat{v}} \cdot \mathbf{a}) 
\]

for any \( \mathbf{v} \) and \( \mathbf{a} \).

Each component of the new vector again depends linearly on \( \mathbf{a} \). Just as we can represent a vector by its cartesian components \( \mathbf{r} = [x, y, z] \), and so on, so also the dyad product has the nine components

\[
[\mathbf{\hat{v}} \otimes \mathbf{v}]_{ij} = \mathbf{e}(\mathbf{e}(\mathbf{\hat{v}} \otimes \mathbf{v}) \cdot \mathbf{e}(\mathbf{\hat{v}} \otimes \mathbf{v}))_{ij} = \mathbf{r}_i \mathbf{r}_j = \begin{bmatrix} x^2 & xy & xz \\
yx & y^2 & yz \\
2x & 2y & 2z \end{bmatrix}_{ij}, \quad i, j = 1, 2, 3. \tag{13.2}
\]

Each of the two terms of Equation 13.1 therefore has a \( 3 \times 3 \) array of components, and hence so does their sum \( \mathbf{\hat{J}} \).

Equation 13.2 illustrates a general idea:

**The components of a tensor are what emerge when we feed it the unit vectors corresponding to the axes of a cartesian coordinate system, then find the components of the result in that system.**

We again say that the components **represent** the tensor in the chosen coordinate system. Yet another view is to regard the components as forming a column vector \( \mathbf{r} \); then the usual rules of matrix multiplication give that \( [\mathbf{r}] [\mathbf{r}]^{1/2} \) is a \( 3 \times 3 \) matrix containing the components of the dyad product.

A tensor whose matrix of components is symmetric, for example \( \mathbf{\hat{J}} \), will itself be called a **symmetric** tensor. In fact, any tensor \( \mathbf{\hat{T}} \) of rank 2 can always be written as the sum \( \mathbf{\hat{T}}^{[S]} + \mathbf{\hat{T}}^{[A]} \), where the **symmetric and antisymmetric parts** are

\[
\mathbf{\hat{T}}^{[S]} = \frac{1}{2}(\mathbf{T} + \mathbf{T}^t) \quad \text{and} \quad \mathbf{\hat{T}}^{[A]} = \frac{1}{2}(\mathbf{T} - \mathbf{T}^t).
\]

---

\(^4\)As mentioned in Section 0.2.1, most authors omit the over-arrow when stating components, but we will retain it to emphasize the tensor status of the object that they describe.

\(^5\)Some books omit the symbol \( \otimes \) and use the ultra-concise convention that when two vectors are juxtaposed with no dot or cross joining them, this dyad product is implied. Some books call the dyad product the “outer product.” Later, we will introduce a generalization called “tensor product”; some books use this term for the dyad product as well.
Chapter 13 Examples of 3-Tensors in Physics

The operation “find the (anti)symmetric part of \( \tilde{T} \)” is often abbreviated to “(anti)symmetrize \( \tilde{T} \).”

If we set the two indices equal (consider only diagonal elements) and sum them, then the result is a single number called the \textbf{trace} of the tensor.\(^6\)

**Your Turn 13A**

a. Use first-year physics formulas to find an expression for \( \tilde{L} \) in terms of \( \tilde{\omega} \), \( \{\vec{r}_i(\ell)\} \), and \( \{m_\ell\} \). Then compare to \( \tilde{L}_I = \tilde{J}_{ij} \tilde{\omega}_j \) and rearrange as needed to obtain Equation 13.1.

b. Although \( \tilde{J} \) is symmetric, show that it need not be traceless (in contrast to the electric quadrupole moment tensor).

c. Show that if \( \tilde{\omega} \) is an eigenvector of \( \tilde{J} \), then the body can spin freely about that axis without wobbling (precessing).

Note that although \( \tilde{L} \) depends linearly on the components of \( \tilde{\omega} \), it need not point parallel to \( \tilde{\omega} \). If not, then in rigid rotation \( \tilde{L} \) will trace out a cone with \( \tilde{\omega} \) as its axis. That time dependence implies the unwanted torque mentioned in the automotive example.

Also note that, although \( \tilde{L} \) and \( \tilde{\omega} \) both change sign if we switch to a left-handed coordinate system, nevertheless, the relation between them is unaffected.

**Your Turn 13B**

a. Indeed, Equation 13.1 does not contain any cross products. Where did they go?

b. Work out the moment of inertia tensor of a solid cylinder with uniform mass density. Let its length be \( L \), its radius be \( R \), and use its center as the reference point. Once you’ve got it, make an Appropriate Comment about what its structure implies for spinning the cylinder about an axis that passes through its center but does not coincide with the axis of symmetry.

13.3.2 More general examples of rank-two tensors

So far, we have introduced the identity tensor and the dyad product of a vector with itself. The dyad product of \textit{any} two vectors is defined similarly: \( \tilde{u} \circ \tilde{v} \) is the tensor that eats any \( \tilde{a} \) and returns

\[
(\tilde{u} \circ \tilde{v}) \cdot \tilde{a} = \tilde{u}(\tilde{v} \cdot \tilde{a}).
\]

Each of the three components of the right side of this expression is a linear function of the component of \( \tilde{a} \). Notice that \( \tilde{u} \circ \tilde{v} \) is not necessarily the same function as \( \tilde{v} \circ \tilde{u} \): \textit{The dyad product is not commutative.}

As before, we can construct a \( 3 \times 3 \) array of ordinary numbers by letting \( \tilde{u} \circ \tilde{v} \) eat each of the coordinate axes and then expanding the three resulting vectors in components.

\(^6\)We encountered the trace in Sections 3.2 and 7.2.3.
13.3 Rank Two

13.3.1 Tensors arise naturally throughout physics: some examples

1. When we pull a rigid body through a viscous fluid, the fluid exerts a retarding drag force. If the body is spherical, then the drag force points oppositely to the velocity, but more generally, we get a linear relation $\vec{f} = \vec{n} \cdot \vec{v} + \cdots$, involving a viscous drag tensor. The fact that $\vec{f}$ need not be parallel to $\vec{v}$ is the key to bacterial locomotion (Figure 13.1).

2. A mass suspended on an array of springs has an equilibrium position. If we apply a small force to the object, then it responds by finding a new mechanical equi-

\textit{Figure 13.1:} [Schematic.] **Principle of flagellar propulsion in bacteria.** (a) A thin rod is dragged through viscous fluid. The force required to get velocity $\vec{v}$ is not parallel to $\vec{v}$, because the drag coefficient is larger in the perpendicular direction. (b) A thin, rigid, helical rod is cranked about its helix axis at angular speed $\omega$. For better visualization, a phantom cylinder has been sketched, with the rod lying on its surface. Two short segments of the rod have been singled out for study, both lying on the near side of the helix and separated by one turn. The rod is attached (black circle) to a disk, and the disk is rotated. The two short segments then move downward in the plane of the page (along $-\hat{x}$). The resulting $d\vec{f}$ lies in the $xz$ plane but is tipped slightly to the left as in (a). If $d\vec{f}$ were parallel to $\vec{v}$, then all forces would cancel. Instead, a net force with a negative $z$-component is required to keep the helix spinning in place; without such an external force, the helix will move to the right.

\textbf{Your Turn 13C}

- a. Show that the resulting matrix, which can be written as $[\vec{u} \otimes \vec{v}]_{ij}$, has entries given by the products $\vec{u}_i \vec{v}_j$ in row $i$ and column $j$. Compare this definition to the special case Equation 13.2. This matrix will only be symmetric if $\vec{u}$ is a scalar multiple of $\vec{v}$ (or if either $\vec{u}$ or $\vec{v}$ is zero).
- b. Show that the components of $\vec{v} \otimes \vec{u}$ form a matrix that is the transpose of $[\vec{u} \otimes \vec{v}]_{ij}$.

But not every rank-2 tensor can be written as a dyad product, for example, the moment of inertia tensor. Even the sum of two dyad products will not itself be a expressible as a dyad product.

\textit{Note:}

\textit{For a small and/or slowly moving body in viscous fluid, higher-order terms indicated by the ellipsis are negligible. You may be more familiar with a formula for wind resistance that is quadratic in velocity; that term can dominate for large bodies moving rapidly through a low-viscosity medium (for example, for wind resistance on a car).}
librium displaced by some \( \vec{r} \), such that \( \vec{f} = -\vec{K} \cdot \vec{r} \). Here the \textbf{spring tensor} \( \vec{K} \) summarizes the spring system as far as its linear response is concerned. Conversely, \( \vec{r} = -[\vec{K}]^{-1} \cdot \vec{f} \). The tensor whose components are the inverse matrix of \([\vec{K}]\) is called the \textbf{compliance tensor}.

3. Continuing example 2, suppose that the object is electrically charged; for example, it could be part of a molecule. Then it may respond to an applied electric field by deforming, which in turn gives an \textbf{induced dipole moment}. In the linear regime, \(^8\)

\[
\vec{\mathcal{D}}_E = \vec{\alpha} \cdot \vec{E},
\]

where \( \vec{\alpha} = q^2 [\vec{K}]^{-1} \) is called the \textbf{polarizability tensor}. Because molecular polarizability gives rise to dielectric susceptibility and hence to a change in permittivity, \(^9\) those quantities are also in general tensors.

4. Some electrically conductive media are ohmic but anisotropic. This means that although charge flux is a linear function of electric field, those vectors need not be parallel, analogously to example 1. Thus, instead of \( \vec{j} = \sigma \vec{E} \), we have \( \vec{j} = \vec{\sigma} \cdot \vec{E} \), where \( \vec{\sigma} \) is called the \textbf{conductivity tensor}. Similarly, any molecule or ion that moves diffusively has a \textbf{mobility tensor}, which need not be a scalar if the medium is anisotropic.

5. The net force \( df_{1\to2} \) exerted by a small element of fluid 1 on its adjacent neighbor 2 depends linearly on the area of the interface between them, but might not be directed perpendicular to that surface. More precisely,

\[
df_{1\to2} = T \cdot d\Sigma_{1\to2}, \tag{13.4}
\]

where \( T \) is a symmetric rank-2 tensor and \( d\Sigma_{1\to2} \) is directed along the perpendicular from 1 to 2.

6. \( \text{T} \) The order parameter describing the state of a nematic liquid crystal can also be regarded as a traceless symmetric tensor of rank 2.\(^{10}\)

Some further explanation of \( T \) will be useful later. We have an intuitive picture of what it means for our hands to exert force on a rock, but what could it mean for two regions of a fluid, even one as insubstantial as the air in a room, to do this? To answer, recall that force is the rate of momentum transfer. Imagine a small rectangular frame in the middle of a room, separating regions 1 and 2. Even if the average molecular velocity is zero (no overall flow of air), molecules of air are still in random thermal motion. They constantly pass through the frame, \textit{carrying their momentum} at some rate per time.

In equilibrium, the net momentum transfer is zero; molecules passing from 1 to 2 across the surface are canceled by those passing from 2 to 1. But Equation 13.4 involves a

---

\(^{8}\)Even if polarization is nonlinear in applied field, we can use the first-order part of its Taylor expansion to define a tensor.

\(^{9}\)See Sections 6.5.2–6.5.3. The tensor character of permittivity will enter our discussions of birefringence in Chapter 52.

\(^{10}\)See Problem 14.3.
subtly different quantity, signaled by the phrase “net exerted by 1 on 2.” What this means is that each contribution is weighted by +1 if it is carried by a molecule crossing from 1 to 2, or by −1 if the molecule crosses in the opposite sense. With this weighting, the two kinds of contribution need not cancel, even in equilibrium. We will call $\vec{T}$ the **momentum flux 3-tensor**.\(^1\)

**Your Turn 13D**

a. Still thinking about the air in a room, suppose that $d\Sigma_{1\to2}$ is oriented along the +z axis. Make a connection between $\vec{T}_{zz}$ and air pressure.

b. Without making any physical change, reverse the roles of regions 1 and 2. Show that two minus signs arise in Equation 13.4, and hence $\vec{T}_{zz}$ is the same as in (a).

c. Now imagine a box of air with real, not imaginary walls, and pump out all the air inside of it. Exterior air molecules now collide elastically with the walls, each transferring twice its original momentum in the perpendicular direction. Show that, although there are no air molecules inside, this factor of two implies that the force per unit area perpendicular to the walls with constant $z$, in equilibrium, is again $\vec{T}_{zz}$.

We can state that last result more invariantly by defining the **pressure** of a fluid in equilibrium, in a coordinate system where it is overall at rest, as the trace of $\vec{T}$ divided by 3.\(^2\)

Similarly, in an elastic continuum, like a lump of jello or steel, each volume element exerts forces on its neighbors, again described by a momentum flux 3-tensor. Unlike in a fluid, even a *static* deformation can lead to stresses in an elastic body.

### 13.3.4 A symmetric tensor can also express a scalar, quadratic function of a vector

If $\vec{T}$ is any tensor of rank 2, then $f(\vec{u}) = \vec{v} \cdot \vec{T} \cdot \vec{u}$ defines a corresponding quadratic function of $\vec{v}$. For example, the length-squared function, $f(\vec{v}) = ||\vec{v}||^2$, is a scalar-valued function that is quadratic in the components of $\vec{v}$. We’ll call it the 3D **metric tensor**.\(^3\)

Its components in any cartesian system are called the **Kronecker symbol** $\delta_{ij}$.

Here are some more examples of this idea:

---

\(^1\)Many authors instead use the phrase **stress tensor**, but beware that a minority use that same phrase to denote a different quantity. This book will avoid confusion by using the descriptive name “momentum flux 3-tensor.”

\(^2\)Other contributions to $\vec{T}$ can arise if the fluid is in motion. For example, if the fluid is in rigid motion with velocity $\vec{u}$, then there will be an additional flux of momentum given by $\rho_m \vec{u} \otimes \vec{u}$. More generally, fluid in nonuniform motion can have additional contributions called **shear stresses**. The alternate meaning of “stress tensor,” for fluids only, mentioned earlier is similar to our $\vec{T}$ except that the $\rho_m \vec{u} \otimes \vec{u}$ has been subtracted away.

\(^3\)This same tensor, regarded as a linear function of a vector, was called $\hat{n}$ in Equation 13.1.
Your Turn 13E

a. Show that the kinetic energy of a spinning rigid body is 
\[ \frac{1}{2} \vec{\omega} \cdot \vec{T} \cdot \vec{\omega}, \]
where \( \vec{T} \) is the moment of inertia tensor introduced earlier.

b. Show that the rate at which work is done pulling a rigid object through viscous fluid equals 
\[ \vec{v} \cdot \vec{\eta} \cdot \vec{v}, \]
where \( \vec{\eta} \) is the drag tensor.

c. Show that the potential energy stored by a spring system is 
\[ \frac{1}{2} \vec{r} \cdot \vec{K} \cdot \vec{r}, \]
where \( \vec{K} \) is the spring tensor.

d. Show that the dissipated power density in a general ohmic material is 
\[ \vec{E} \cdot \vec{\kappa} \cdot \vec{E}, \]
where \( \vec{\kappa} \) is the conductivity tensor. Show how the units work in this formula.

A tensor that specifies a quadratic function must be symmetric, because any antisymmetric part would cancel in the expressions appearing above. To see this, suppose that 
\[ f(\vec{u}) = \vec{u} \cdot \vec{T} \cdot \vec{u}. \] If \( \vec{T} \) is antisymmetric, then this expression equals 
\[ -\sum_{i,j} T_{ij} \vec{u}_i \vec{u}_j. \] But we may rename the summed indices in any way we please, so long as their names remain different from each other. Let us rename “(i)” as “(j)” and vice versa. Then
\[ f(\vec{u}) = -\sum_{j} T_{ij} \vec{u}_i \vec{u}_j = -f(\vec{u}). \]

Any quantity equal to minus itself is zero! Indeed, the moment of inertia, quadrupole, and metric tensors, all of which specify quadratic functions of a vector, are all symmetric.

No such restriction applies to a tensor used to specify a bilinear function; \( g(\vec{u}, \vec{v}) \) may or may not be equal to \( g(\vec{v}, \vec{u}) \).

The electric quadrupole moment also defines a contribution to the far potential that depends quadratically on \( \vec{r} \): Equations 3.2 and 3.4 (page 38) give this as
\[ (\text{quadratic function of } \vec{r}) / r^5. \]

The numerator is specified by a tensor \( \vec{Q}_e \), which has a coordinate representation as a \( 3 \times 3 \) matrix. That matrix changes when we change coordinates (or rotate the object) in the same way as any of the other tensors described above. In contrast to some of the preceding examples, we found that \( \vec{Q}_e \) is always traceless; that is, \( \text{Tr} \vec{Q}_e = 0 \).

13.3.5 Some linear vector functions, but not all, arise as the derivative of a quadratic scalar function

In ordinary calculus, any linear function can be written as the derivative of a quadratic function: \( \alpha x = (\frac{1}{2} \alpha x^2)' \). Some vector-valued functions of a vector can similarly be written as the gradient of a quadratic function. For example, the Hooke-law force is the gradient of minus the potential energy. Unlike in one dimension, however, not every linear \( \vec{f}(\vec{r}) \) can be expressed in this way.

For example, consider again a rigid body. When we rotate it about the \( z \) axis, the
position of each mass element $\ell'$ changes from $\vec{r}_I(\ell')$ to $\vec{r}_I(\ell') + d\vec{r}_I(\ell')$, where\(^{14}\)

$$d\vec{r}_I(\ell') = d\vec{T} \cdot \vec{r}_I(\ell'), \quad \text{with} \quad \vec{T}_{ij} = \delta^j_i \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$  \hspace{1cm} (13.5)

This linear function of $\vec{r}_I(\ell')$ is specified by an *antisymmetric* tensor, whereas anything arising as derivatives of a quadratic function would have to be expressed by a *symmetric* matrix.

Section 13.3 has outlined the many useful roles played in physics by tensors of rank 2. Next, we’ll extend these ideas to a new class of geometrical objects.

Section 13.3’ (page 207) offers a connection to quantum mechanics.

### 13.4 TENSORS VERSUS MATRICES

Why do we introduce a separate concept of tensor, when they turned out to be “just matrices?” For one thing, we would in any case need new words to describe rank three and higher. But there is a deeper reason, already hinted.

A vector has some intrinsic geometric meaning, independent of any choice of coordinate system. For example, it could be the tangent to a curve at a point. Tensors of higher rank as introduced in this chapter also have intrinsic meaning, even if it is harder to visualize. For example, “vector-valued, linear function of a vector” makes sense even if we have not settled on a coordinate system yet. So even though in practice we ultimately do choose coordinates and work with matrix representations, still the fact that a tensor is intrinsic tells how its matrix representation will change when we change coordinates. That understanding will be useful because relativity theory, for example, is the study of how tensor quantities appear different in different coordinate systems.

### 13.5 RANK THREE

#### 13.5.1 Handedness is conventional in classical physics

Generally we introduce dynamical variables whose values are determined operationally by physical measurements. Some quantities, however, require one bit of extrinsic information before they can be defined: an arbitrary convention for which coordinate systems we will deem “right handed.”\(^ {15}\) Ultimately we will eliminate such quantities from our formulation of electrodynamics, but they are still widely used, so for now suppose that such a choice has been made. Another scientist, observing the same reality but with the opposite convention, would report different values for such observables.

---

\(^{14}\)We encountered this relation earlier in Equation 3.12 (page 46).

\(^{15}\)The choice of “right-hand” convention has nothing to do with your hands, which side of your body your heart is on, nor even the shape of your DNA. Any cartesian coordinate system may be singled out and given this status.
Chapter 13 Examples of 3-Tensors in Physics

Figure 13.2: A parallelepiped is a six-sided solid with three pairs of parallel faces, each of which is a parallelogram. The edge vectors $\vec{U}$, $\vec{V}$, and $\vec{W}$ shown form a right-handed triad, so $\varepsilon(\vec{U}, \vec{V}, \vec{W})$ in Equation 13.6 is positive.

13.5.2 Levi-Civita as a vector-valued bilinear function of vectors

Here is a recipe from Section 0.2.2 (page 5): Given two vectors, return zero if they are parallel (or if either is zero). Otherwise, find the vector $\vec{n}$ that is perpendicular to the plane that they span and is chosen using the right-hand rule. Let $\Sigma$ be the area of the parallelogram with the two given vectors as edges, and define the cross product as $\vec{n}\Sigma$. This new vector is linear in each of the two that we began with; for example:

- If we double either vector, then $\Sigma$ doubles and $\vec{n}$ is unchanged, so $\vec{n}\Sigma$ doubles.
- If we replace either vector by its negative, then $\Sigma$ is unchanged but $\vec{n}$ reverses, so $\vec{n}\Sigma$ also changes sign.

So the operation of cross product is itself some kind of tensor.

The operation just defined eats two vectors and returns another vector, so we need to generalize Section 13.3.1 by introducing an array of numbers with three indices to express it. Instead of regarding its components as a matrix (grid of cells addressed by row and column), imagine it as an apartment building with “rooms” addressable by row, column, and floor. Each room is inhabited by a numerical value. Those 27 numbers, the components of the Levi-Civita tensor, are given by the Levi-Civita symbol defined earlier (Figure 0.3), as you can check by examples (try substituting $\hat{x}$, $\hat{y}$, and $\hat{z}$ into the preceding definition). Chapter 14 will discuss this rank-3 tensor in more detail.

13.5.3 Levi-Civita as a scalar-valued trilinear function of vectors

Here is another way to look at the Levi-Civita tensor, by extending Section 13.3.4. Given three vectors $\vec{U}$, $\vec{V}$, and $\vec{W}$, construct the parallelepiped that has these vectors as three edges (Figure 13.2). Compute the volume $v$ of this solid and multiply by $\sigma = +1$ if the three given vectors form a right-handed triad (otherwise $\sigma = -1$):

$$\varepsilon(\vec{U}, \vec{V}, \vec{W}) = v\sigma.$$  \hfill (13.6)

Exchanging any two of the three vectors leaves $v$ unchanged while reversing the sign of $\sigma$, so we say that $\varepsilon$ is totally antisymmetric.\(^{16}\)

---

\(^{16}\)A totally symmetric rank-3 tensor would be unchanged under exchange of any of its inputs, just as in rank two. Mixed symmetry is also allowed, for example, antisymmetric on only the first two arguments and so on.
Equation 13.6 yields a function that is linear in all three of its vector arguments. For example:

- If we double the length of any input vector, then volume doubles and \( \sigma \) is unchanged, so \( \sigma \) doubles.
- If we replace any input vector by its negative, then volume is unchanged but \( \sigma \) is replaced by its negative, so \( \sigma \) changes sign.

In fact, Equation 13.6 is just \( \vec{U} \cdot (\vec{V} \times \vec{W}) \), similar to the relation between the two interpretations of rank-1 tensors in Section 13.2.

**Your Turn 13F**

a. Clearly \( \vec{U} \cdot (\vec{V} \times \vec{W}) \) changes sign if we exchange \( \vec{V} \) and \( \vec{W} \). But show that it also changes sign if we exchange \( \vec{U} \) and \( \vec{V} \).

b. What happens upon exchange of \( \vec{U} \) and \( \vec{W} \)?

c. What happens under a cyclic permutation, \( \vec{U}, \vec{V}, \vec{W} \rightarrow \vec{W}, \vec{U}, \vec{V} \)?

Again, substituting \( \vec{U} = \hat{x}, \vec{V} = \hat{y}, \) and \( \vec{W} = \hat{z} \) shows that the 1,2,3 component of this tensor equals 1, which agrees with the 1,2,3 entry of the Levi-Civita symbol. We also see that permuting the three vectors leaves \( \nu \) unchanged but changes \( \sigma \) by the sign of the permutation, again like \( \varepsilon_{ijk} \), so again we find that the Levi-Civita symbols are components of a totally antisymmetric, 3-tensor of rank three.

When there are more than 2 indices (rank higher than 2), it’s too cumbersome to put any arrow or other glyph above the symbol to indicate its tensor character. Also, in this situation we will never drop the indices, so their presence suffices to announce that \( \varepsilon_{ijk} \) are the components of a 3-tensor of rank three.

### 13.6 Tensor Fields

Chapter 7 introduced a quadratic function of small displacements describing how a curved 2D surface bends away from its tangent plane. The matrix \( B \) defined in Equation 7.3 (page 103) is generally not a constant; it defines a tensor associated with each point of the surface. Just as we can have vectors that depend on position, so also there are tensor fields. Later chapters will make extensive use of this concept. In fact, the permittivity and conductivity of a nonuniform medium, and the momentum flux 3-tensor of a fluid, are all local state variables that are tensor fields.

**Further Reading**

*Semipopular:*
This video is worthwhile: [www.youtube.com/watch?v=f51iq0kOCsw](http://www.youtube.com/watch?v=f51iq0kOCsw).

*Intermediate:*
Chapter 13  Examples of 3-Tensors in Physics

General: Neuenschwander, 2015; Arfken et al., 2013; Cahill, 2019; Fleisch, 2012; Stone & Goldbart, 2009.
Physical examples and particularly the momentum flux 3-tensor: Feynman et al., 2010a, chap. 31.
Tensors in other areas of physics and engineering: Schobeiri, 2021.
13.2’a Vectors and their duals
The main text described both vectors, and linear machines that convert vectors to scalars, as being examples of 3-tensors of rank 1. Mathematicians distinguish these two kinds of object and call each kind the other one’s “dual.” They also often refer to the linear machines as “covectors.” For example, the gradient of a function can be regarded as a linear machine that eats a vector at some point and returns the directional derivative of the function along that vector at that point. This chapter neglected the distinction, but Chapter 34 will return to it.

13.2’b Tensor properties of probability density functions
[[Not ready]]

13.3’a Tensors in quantum mechanics
Quantum mechanics introduces an operator on state space called the “density matrix.” If it can be expressed as a dyad product, then it represents a “pure state”; otherwise, it is a “mixed state.” Even the sum of two pure-state density matrices will not in general represent any pure state (it’s usually “entangled”), echoing the corresponding statement in the main text about dyad products.

Quantum mechanics also has operator families that transform as tensors under rotation, and tensor representations of internal symmetry groups in high energy physics. All are subject to similar analyses. A more general class of fields, designed to handle intrinsic particle spin, are called “spinors” (Section 34.7b, page 542).

13.3’b Another concept of rank
In this book, the “rank” of a tensor T always means the number of indices, a convention followed by most physicists. Some mathematicians reserve “rank” for a notion from linear algebra: the dimension of the image space when T is fed all possible input vectors. In this sense, a dyad product always has rank less than three, because its matrix of components always has determinant zero.
13.1 Octahedron I
A mass distribution consists of six equal point masses $m$ placed at the vertices of an octahedron: $\vec{r}_{(±1)} = (±\alpha, 0, 0)$, $\vec{r}_{(±2)} = (0, ±\alpha, 0)$, $\vec{r}_{(±3)} = (0, 0, ±\alpha)$. Find the moment of inertia tensor of this mass distribution about the origin. Does it have any surprising feature?

13.2 Octahedron II
In both parts below, use the origin of coordinates as the basepoint for computing multipole moments.\(^{17}\)

a. A charge distribution consists of six single point charges all of magnitude $e$ placed at the vertices of the octahedron introduced in Problem 13.1. A neutralizing charge $-6e$ is placed at the origin. Find the electric dipole and quadrupole moments.

b. A charge distribution consists of four single charges $e$ placed at the vertices of a square: $\vec{r}_{(±1)} = (±\alpha, 0, 0)$, $\vec{r}_{(±2)} = (0, ±\alpha, 0)$. A neutralizing charge $-4e$ is placed at the origin. Find the electric dipole and quadrupole moments.

---

\(^{17}\)See Section 3.6.4, page 42.
14.1 FRAMING: INTRINSIC STRUCTURES

The preceding chapter gave many examples of tensors in physics. A little thought shows that they fall into two main classes:

- Most of the examples were contingent; they describe properties of an object. If we rotate a mass distribution, its moment of inertia tensor in general changes (unless we rotate about a symmetry axis). Even total mass, which is rotationally a scalar, changes if we consider a different object.
- Two of the examples were different: The 3D metric tensor (Section 13.3.4, page 201) is a property of space itself, not contingent on anything. And we’ll see that the Levi-Civita tensor is almost equally intrinsic to space: It depends only on a binary choice of which coordinate systems we have chosen to call “right-handed.”

Let’s explore these last two tensors “from Heaven” a bit more. Along the way, we will also examine how any tensor’s representation changes if, instead of changing the physical objects under consideration, we merely change our choice of coordinate system. This understanding will prove useful when we start to construct more elaborate things, and then again when we upgrade everything to four dimensions.

**Phenomenon:** Although nematic liquid crystals are made from complicated molecules, only a few physical constants are needed to describe their overall behavior.

**Physical idea:** Rotational invariance permits only a few terms in the free energy function.

14.2 THE COMPONENTS OF A TENSOR TRANSFORM UPON LINEAR CHANGE OF COORDINATES

14.2.1 An example from mechanics

Section 13.3.4 said that we may think about a spring tensor \( K \) as a function that eats a displacement vector and returns a number, the stored potential energy \( \frac{1}{2} \Delta \mathbf{r} \cdot K \cdot \Delta \mathbf{r} \). This function is quadratic in the components of \( \Delta \mathbf{r} \). It can be represented in any coordinate system by a matrix of ordinary numbers. We called those numbers the **components** of \( K \) in the chosen coordinate system, and denoted them by \( K_{ij} \). It’s important that the nine numbers \( K_{ij} \) depend not only on the physical object (system of springs), but also on a choice of **coordinate system** on space. That is, the same tensor can have different representations when referred to different coordinate systems.
Suppose that we define new coordinates by

\[ \vec{r}'_a = S_{al} \vec{r}_l. \]  

(14.1)

Then the same spring potential energy function as before can also be written \( \frac{1}{2} \Delta \vec{r}' \cdot \vec{K}' \cdot \Delta \vec{r}' \), where the new components are determined by

\[ \vec{r}'_i \vec{K}'_{ij} \vec{r}'_j = \vec{r}'_a \vec{K}'_{ab} \vec{r}'_b = [\vec{r}]^t S^t [\vec{K}'] S [\vec{r}]. \]

This must hold for any spring displacement, so \([\vec{K}'] = S^t [\vec{K}] S\), or

\[ \vec{K}'_{ab} = S_{al} S_{bj} \vec{K}_{ij}. \]  

(14.2)

14.2.2 Cartesian coordinates are connected via orthogonal matrices

In euclidean geometry, there are always some special ways to associate numbers to points in space (that is, to choose a coordinate system\(^1\)). What’s special about these cartesian coordinate systems is that the distance-squared between two points always takes the pythagorean form

\[ ||\Delta \vec{r}||^2 = \sum_i \Delta \vec{r}_i \Delta \vec{r}_i \]  

in cartesian coordinates. (14.3)

Certainly we can find other coordinate systems for euclidean space in which the metric tensor doesn’t have the simple form Equation 14.3, for example, polar coordinates. What makes euclidean space special is that at least one such set of “good” coordinates does exist (unlike, say, on the surface of a sphere).

If one cartesian coordinate system exists, then many others, equally good, will exist also. To see this, again define new coordinates via Equation 14.1, where now \( S \) is specifically an orthogonal matrix, that is, one for which

\[ SS^t = S^t S = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \]  

(14.4)

Then the new coordinates again have the property that the length-squared of a vector equals \( \sum \Delta \vec{r}'_a \Delta \vec{r}'_a \), which has the same form as Equation 14.3. For future use, note that Equation 14.4 implies

\[ (\det S)^2 = 1, \text{ and hence } \det S = \pm 1 \]  

for an orthogonal matrix. (14.5)

The case with positive determinant contains all rotations; the other case consists of spatial inversion combined with a rotation.\(^3\)

---

\(^1\) Later we will upgrade to coordinate systems on space and time (four dimensions). For now, we consider three-dimensional space only.

\(^2\) If you are worried about up- versus down-indices, we’ll get to that fine point later. It’s traditional to forget about this distinction when we work on euclidean 3-space in cartesian coordinates, and always write coordinate indices as subscripts. (But when we use tensors on a non-euclidean space, or with curvilinear coordinates, the distinction becomes essential.)

\(^3\) Some books say “proper” or “improper rotations” for the cases with positive and negative determinant, respectively.
14.2.3 The components of the 3D metric are the same in any cartesian system

Section 13.3 defined tensors as functions involving vectors. The preceding section showed that their components have certain transformation rules generalizing those of vectors. Alternatively, we could turn things around and instead define a general rank-2 tensor, such as a spring tensor $\hat{K}$, as any set of nine numbers that transform like the components of $\hat{F} \otimes \hat{F}$. Similar relations can be used to define a 3-tensor of any rank $p$: There will be $p$ copies of the transformation matrix on the right-hand side of Equation 14.2.

Let’s look at the metric tensor from this new viewpoint. Instead of the geometric definition, we can say

Choose any cartesian coordinate system. Define $\tilde{T}$ to be that 2-tensor whose components in this coordinate system are $\delta_{ij}$.

The corresponding quadratic function defined by $\tilde{T}$ is then the usual length-squared.

The formulation Equation 14.6 may worry you: What if you and your friend start out with different cartesian coordinate systems? Will you both agree on the meaning of $\tilde{T}$? To investigate, let’s see how the components of your tensor look in your friend’s (primed) coordinate system, by applying Equation 14.2:

$$\tilde{T}_{ij}' = S_{ai}S_{bj}\delta_{ij} = S_{ai}S_{bj} = [SS']_{ab} = \delta_{ab},$$

the same nine constants as before. That is, it doesn’t matter what coordinate system we started with, as long as it’s cartesian: the components of the metric tensor are always the same. So the tensor we defined is not contingent on coordinates chosen; it is a property of euclidean space itself. Admittedly, in this case that conclusion is a tautology, not a surprise (it just restates the definition of cartesian coordinate system). However, we can now use the same logic to get a more nontrivial result.

14.3 COMPONENTS OF THE LEVI-CIVITA TENSOR

14.3.1 The components of $\varepsilon$ are the same in any right-handed cartesian system

Section 13.5 gave two geometric definitions of the Levi-Civita tensor, then noted that its components are given by the Levi-Civita symbol (that is, the constants $\pm 1$ or zero). As in Section 14.2.3, one may worry: What if you and your friend choose different coordinate systems when defining it? Following the preceding paragraph, we’d like to show that if we start in one cartesian system, then transform to any other, then the components are numerically the same as before. If that’s the case, then the tensor that they define won’t actually depend on the original choice of coordinates—it’s an intrinsic property of space itself, a tensor “from Heaven.” We know this must work out somehow, because we started with a geometric definition (Figure 13.2), but the details are interesting (in part because we will later use the same approach to generalize to four dimensions).

Following Sections 0.2.2 and 13.5.1 (page 203), suppose that we work in euclidean space, and moreover, we have chosen one of the cartesian coordinate systems and declared it to be “right-handed.” We now define a 3-tensor by stating its components as in

Fig. 13.2 (page 204)
Section 0.2.2 (page 5):

\[ \varepsilon_{ijk} = 0 \text{ if any two of the indices are equal;} \]
\[ \varepsilon_{ijk} = +1 \text{ if } i, j, k \text{ are an even permutation of 1,2,3;} \]
\[ \varepsilon_{ijk} = -1 \text{ if } i, j, k \text{ are an odd permutation of 1,2,3.} \]  
(14.8)

Next, we must calculate the new components

\[ \varepsilon'_{abc} = S_{ai} S_{bj} S_{ck} \varepsilon_{ijk}. \]  
(14.9)

and show that they are the same 27 numbers as in Equation 14.8. First, note that

\[ \varepsilon'_{112} = S_{1i} S_{1j} S_{2k} \varepsilon_{ijk}. \]

The sums over \( i \) and \( j \) involve something antisymmetric under exchange (that is, \( \varepsilon_{ijk} \)) times something symmetric under exchange (that is, \( S_{1i} S_{1j} \)). Altogether, the expression is therefore antisymmetric, so it gives zero when summed over \( i, j \). Indeed, we get zero when any two indices of \( \varepsilon'_{abc} \) are equal, as required by Equation 14.8.

All that remains, then, is to check the case where \( i, j, k \) are all different. In fact, you can readily show that \( \varepsilon'_{abc} = -\varepsilon'_{bac} \) and so on, as desired, so we only need to check a single permutation, for example, \( \varepsilon'_{123} \). And of the 27 terms being summed in Equation 14.9, all but six are zero:

\[ \varepsilon'_{123} = S_{11} S_{22} S_{33} + S_{12} S_{23} S_{31} + S_{13} S_{32} S_{21} - S_{11} S_{23} S_{32} - S_{13} S_{22} S_{31} - S_{12} S_{21} S_{33} \]
\[ = \det S. \]  
(14.10)

But we know that \( \det S = \pm 1 \) for any orthogonal matrix (Equation 14.5). Moreover, any two right-handed coordinate systems are related by a rotation. Any rotation can be continuously obtained from the identity operator, whose determinant is +1. Thus,

\[ \cdot \det S = +1; \]
\[ \cdot \det S \text{ cannot change discontinuously;}^4 \text{ and} \]
\[ \cdot \text{Any rotation can be continuously reached starting from the identity by a chain of rotations with increasing angle.} \]

Those facts are enough to conclude that the determinant must always be \( +1 \). Thus, Equation 14.10 implies \( \varepsilon'_{123} = +1 \), completing the proof that all components are the same in any right-handed system.

### 14.3.2 Components only specify a unique \( \varepsilon \) after a right-hand convention is chosen

However, had we used Equation 14.8 in conjunction with an oppositely-handed system, then we would have defined a different Levi-Civita tensor, \( \bar{\varepsilon} \), equal to minus the one defined below.

---

^4After all, the determinant of a matrix is just a polynomial in the entries of that matrix.
by Equation 14.8. To confirm that claim, re-express \( \varepsilon \) in terms of a right-handed system \( \{e''_a \} \); then Equation 14.10 says that \( \varepsilon_{abc} = -\varepsilon_{abc} \). So the definition of the Levi-Civita tensor, as well as anything defined with its help (cross product, curl, vector representation of an area element \( d^2 \Sigma \)) requires that we first commit to a convention about which is our “right” hand. For this reason, some books refer to the “Levi-Civita pseudo tensor.” We’ll instead take the viewpoint that

\( \varepsilon \) is a perfectly well-defined 3-tensor, once we have chosen a convention for which coordinate systems we will call right-handed.

Even the purely geometric, coordinate-independent definition of \( \varepsilon \) in Section 13.5.3 (page 204) depended on a choice of right-hand convention.

14.3.3 Ne Plus Ultra

Remarkably, when mathematicians studied this problem they found that there were essentially no more new 3-tensors “from Heaven.” You can build up higher-rank examples by sticking together some metric and Levi-Civita tensors (for example, the 3-tensor of rank 4 with components \( \delta_{ij} \delta_{k\ell} \)), but that is all.

14.4 CONNECT TO FAMILIAR THINGS

Although the above reasoning is a model for more complicated things to come, it’s also good to see how it connects to things you already know.

14.4.1 Dot product

Besides telling us how long a vector is, the metric tensor can tell us the angle between two vectors \( \vec{v} \) and \( \vec{w} \). Define the dot product as

\[
\frac{1}{2} (||\vec{v} + \vec{w}||^2 - ||\vec{v} - \vec{w}||^2)
\]

It’s a machine that eats two vectors and returns a number that is separately linear in each one (it is “bilinear”). You can quickly see that in any cartesian coordinate system, the invariant definition just given implies that it’s given by the usual formula

\[
\vec{v}_i \delta_{ij} \vec{w}_j = \vec{v}_i \vec{w}_i.
\]

The same derivation as the one above then assures us that we get the same answer regardless of which cartesian coordinate system we chose.

For example, choose a system with \( \hat{x} \) parallel to \( \vec{u} \) and \( \hat{y} \) lying in the \( xy \) plane (Figure 14.1). Let \( \theta \) be the angle between \( \vec{u} \) and \( \vec{v} \). Thus, \( \vec{u} = (1, 0, 0) \) and \( \vec{v} = (\nu \cos \theta, \nu \sin \theta, 0) \). The sum \( \vec{u}_i \vec{v}_i = \nu^2 \cos \theta \) as stated in Section 0.2.1 (page 4).

14.4.2 Cross product

Because we proved the rotation invariance of the Levi-Civita tensor, we know that we can compute \( \varepsilon_{ijk} \vec{u}_i \vec{u}_j \vec{u}_k \) using any right-handed coordinate system we like. The three resulting numbers, interpreted as vector components in the same system, will then define a vector that does not depend on which system we chose. We will call that vector \( \vec{u} \times \vec{v} \).
If \( \vec{v} \) is parallel to \( \vec{u} \), for example \( \vec{v} = \beta \vec{u} \), then the cross product becomes \( \beta \varepsilon_{ijk} \vec{u}_i \vec{u}_k \). This is the sum (“contraction”) of something antisymmetric on \( jk \) times something symmetric on \( jk \), so it’s zero.

If \( \vec{v} \) and \( \vec{u} \) are not parallel, then we may again choose a right-handed, cartesian coordinate system with \( \hat{x} \) along \( \vec{u} \), and with \( \vec{v} \) lying in the \( xy \) plane (Figure 14.1). This time, however, we must be careful to specify that \( \theta \) is the angle from \( \vec{u} \) to \( \vec{v} \), and that \( \theta \) is taken to be positive if that angle is anticlockwise when viewed along the \( z \) axis from positive toward negative values of \( z \) (Figure 14.1). Then again \( \vec{u} = (u, 0, 0) \), \( \vec{v} = (v \cos \theta, v \sin \theta, 0) \), and

\[
(\vec{u} \times \vec{v})_3 = \varepsilon_{31k} u \vec{u}_k = \varepsilon_{312} u (v \sin \theta) = uv \sin \theta,
\]

as stated in Section 0.2.2 (page 5). (You should show that the other two components of \( \vec{u} \times \vec{v} \) equal zero in this coordinate system.)

### 14.5 USEFUL IDENTITIES

#### 14.5.1 Swap dot and cross

The geometrical interpretation of \( \vec{u} \cdot (\vec{v} \times \vec{w}) \) as a volume (Figure 13.2) makes it clear that this quantity equals \( (\vec{u} \times \vec{v}) \cdot \vec{w} \). For practice, you should derive this algebraically by using the properties of the Levi-Civita symbol.

#### 14.5.2 Triple cross product

First, let’s consider \( \varepsilon_{ijk} \varepsilon_{ijk} \). There are three implied summations, for a total of \( 3^3 = 27 \) terms being summed, but most of them are zero. The Levi-Civita symbol is only nonzero if \( i, j, \) and \( k \) all have different values, which requires that they be some permutation of \( 1,2,3 \). There are six such permutations, each contributing a term of the form \( (\pm 1)^2 \), so

\[
\varepsilon_{ijk} \varepsilon_{ijk} = \sum_{\text{permutations}} (\pm 1)^2 = 6.
\]

Next, try the same expression but don’t set the last indices equal nor sum them: \( \varepsilon_{ijk} \varepsilon_{ij\ell} \).

In this expression, the loose indices \( k \) and \( \ell \) are held fixed; for example, suppose they are...
both. Then $\varepsilon_{ij1} \varepsilon_{ij1}$ is the sum of nine terms, again mostly zero. Only two terms survive, so the result is 2. If we instead fix $k$ and $\ell$ to be, say, 1 and 2, then in every term either $\varepsilon_{ij1} = 0$ or $\varepsilon_{ij2} = 0$, so this time $\varepsilon_{ijk} \varepsilon_{ij\ell} = 0$. All told:

$$\varepsilon_{ijk} \varepsilon_{ij\ell} = 2 \delta_{k\ell}. \quad (14.13)$$

Finally, try not setting the last two indices equal: $\varepsilon_{ijk} \varepsilon_{jlm\ell}$. Again, some particular cases are illuminating. We need only consider cases with $i \neq j$ and $m \neq \ell$, for example,

$$\varepsilon_{12} \varepsilon_{12} = 1 \quad \text{and} \quad \varepsilon_{12} \varepsilon_{21} = -1.$$

Considering other special cases gives the general result:

$$\varepsilon_{ijk} \varepsilon_{jlm\ell} = \delta_{jm} \delta_{k\ell} - \delta_{j\ell} \delta_{km}. \quad (14.14)$$

**Your Turn 14A**

Use one of the three preceding identities to get a familiar formula for $\vec{u} \times (\vec{v} \times \vec{w})$.

**Your Turn 14B**

Check that our formulas are mutually consistent as follows:


b. In Equation 14.13, set $\ell = k$, sum $k$, and compare to Equation 14.12.

14.6 PLUS ULTRA

It may seem that we have gone the long way round the barn to reconstruct things you already knew. But when calculations start to get complicated, the benefits of using $\varepsilon$ to express cross products will become clear. Also, the approach of this chapter continues to work in any number of dimensions: For example, we will find it useful to know that a metric space of dimension 4, with a choice of handedness, gets a rank-4 Levi-Civita tensor “from Heaven,” despite the fact that there is no direct analog of cross product. The argument is the same as the one in Section 14.3.1.

Finally, understanding $\varepsilon$ as a tensor will prove valuable as we seek to reformulate electrodynamics without any cross products, thereby making its inversion invariance obvious.

**FURTHER READING**

When referring to a vector quantity that depends on a choice of handedness, some authors add the deprecatory prefix “pseudo-” or the qualifier “axial.” For example, Chapter 15 will package the three numbers needed to represent a magnetic field as $\vec{B}$, which is sometimes
called a “pseudovector” or “axial vector,” and similarly for the components of angular momentum, angular velocity, and torque. There are even pseudoscalars, single quantities that change sign upon change of handedness, including the field that, when quantized, represents the pion. Sometimes a true vector is called a “polar vector” to emphasize that it is not “axial.”

Intermediate:
Neuenschwander, 2015; Arfken et al., 2013; Cahill, 2019; Fleisch, 2012; Stone & Goldbart, 2009.

14.3' Spatial inversion invariance

1. The appearance of the Levi-Civita tensor in a law of classical physics should bother you! Classical mechanics and electrodynamics are supposed to be invariant under spatial inversions, so why do we need any right-hand convention to formulate them? The answer is: We don’t. Both classical mechanics and electrodynamics can be expressed completely without ever introducing cross products or other quantities that depend on a right-hand convention. In fact, doing this for electrodynamics, and hence making its invariance under spatial inversions (“parity invariance”) manifest, is one of this book’s goals. 6

Gravitation and the strong nuclear interaction are also invariant under inversions, but the weak nuclear interactions are not. For example, when a neutron decays, the outgoing antineutrino has a preferred helicity. There is a spin operator analogous to that appears in the weak interaction, that changes under spatial inversion, and that cannot be removed by redefining things.

2. Mathematicians call the binary choice of a right-hand convention an “orientation,” but beware: That term can lead to confusion with the everyday sense of that word (a continuous variable describing which way a rigid object is pointing in space). Similarly, a physicist normally understands the words “change the orientation” to mean “rotate [an object],” not “reverse the right-hand convention for space.”

3. What does “from Heaven” mean? Our constructions all relied on choosing cartesian coordinates. In fact, with a bit more work they can all be generalized to curvilinear coordinates on flat space, or even to curved space; for example, all we need in order to construct an invariant analog of \( \mathbf{\hat{r}} \) is a local distance function. That’s the first step to formulating electrodynamics on curved space(time), for example, to study diffraction effects in gravitational lensing.

Similarly, the Levi-Civita tensor can be defined on any space with a metric and a choice of right-handed coordinate systems: The geometric construction of Section 13.5 works on any such space and does not require any coordinate choice. (It did, however, require a metric to define the volume of the parallelepiped in Figure 13.2, page 204.)

14.4' Twisted tensors

The main text takes the following attitude:

1. Vectors and tensors are real objects with concrete geometrical meaning independent of any choice of coordinate system (for example, vectors “point”).

2. The Levi-Civita tensor, and things constructed with its assistance, are ambiguous (ill-defined) until we choose a right-hand convention. Once such an overall sign choice has been made, however, they become ordinary vectors and tensors.

Actually, however, in three dimensions there is an intriguing reinterpretation of “pseudo” objects that is just as intrinsic (independent of coordinate choice) as ordinary vectors and tensors. For this reason, some authors replace the deprecatory “pseudo” by the more neutral “twisted” to specify these objects. 6 Thus, angular momentum and magnetic induction \( \mathbf{\hat{B}} \) are interpreted as twisted

---

5Equation 15.1 (page 222) will begin this process. Doing the same thing for rigid-body dynamics is similarly rewarding; see Problem 15.4.

6Others speak of “tensor densities.”
Chapter 14  Tensors from Heaven

Figure 14.2: [Diagrams.] Twisted vectors and their operations. (a) An ordinary vector. The reflection \( y \rightarrow -y \) turns it into minus itself.
(b) A twisted vector. The reflection \( y \rightarrow -y \) leaves it unchanged.
(c) Cross product of vector with vector yields a twisted vector (see text).
(d) Cross product of vector with twisted vector yields a vector (see text).
(e) Cross product of twisted vector with twisted vector yields a twisted vector (see text).

vectors, whereas velocity and force are ordinary vectors. This book will not use that approach, but here it is for reference.

To visualize an ordinary vector, we draw a line segment, choose one end, and draw an arrowhead on that end. To visualize a twisted vector, we again draw a line segment, but with no arrowhead. Instead, draw a directed loop encircling the segment. That loop can run in one of two ways, similar to the fact that we can draw the arrowhead on an ordinary vector in two ways. But contrast the objects in Figure 14.2a–b: One changes sign upon a particular reflection, whereas the other does not.

Of course, if we make a choice of which hand to call “right” then we can associate an ordinary vector to any twisted vector and vice versa. If we don’t make any such choice, we must keep these two categories distinct.

We can now define an intrinsic cross product that does not require any choice of right hand, as long as we keep track of the fact that it adds or removes “twistedness”:

- Given two ordinary vectors, return zero if they are parallel or antiparallel. Otherwise, the vectors determine a plane. Construct a line segment perpendicular to the plane with length \( ||\vec{v}|| ||\vec{w}|| \sin \theta \). Imagine a rotation in the plane that turns from the first to the second vector. Instead of trying to put an arrowhead on the perpendicular segment, define the loop encircling it that turns from \( \vec{v} \) toward \( \vec{w} \) (Figure 14.2c). That choice of loop converts the segment into a twisted vector, which we call \( \vec{v} \times \vec{w} \).
- Given a vector \( \vec{v} \) and twisted vector \( \vec{B} \), return zero if they are parallel or antiparallel. Otherwise, proceed as above to draw a perpendicular line segment. This time, however, we place an arrowhead on one end of the segment, as follows: Rotate the arrow representing \( \vec{v} \) about the segment representing \( \vec{B} \) in the sense determined by the loop around it. This brings the arrowhead on \( \vec{v} \) closer to one end of the segment (dashed line in Figure 14.2d); place the arrowhead on the
other end.\(^7\)

- Given two twisted vectors \(\vec{B}\) and \(\vec{\Omega}\), return zero if they are parallel or antiparallel. Otherwise, proceed as above to draw a perpendicular line segment. There will be a rotation in the plane spanned by the two twisted vectors that superimposes \(\vec{\Omega}\)'s loop onto that of \(\vec{B}\). That rotation defines a direction for a loop about the perpendicular segment (Figure 14.2e), allowing us to define it as a twisted vector.

Higher rank twisted tensors can also be defined, but it’s harder and less useful to make visualizable metaphors for them. For more details see Burke, 1985.

Because our goal is to move away from three dimensions, we will not pursue these constructions further. Chapter 15 will regard \(\vec{B}\) as an ordinary vector defined with the help of some choice of right-hand convention, and similarly for cross products. Eventually, later chapters will eliminate such quantities altogether from our formulation of electrodynamics.

\(^7\)This construction also lets us associate an antisymmetric rank-2 tensor \(\vec{\omega}\) to any twisted vector \(\vec{B}\) and vice versa: The tensor takes any vector \(\vec{v}\) and returns the vector \(\vec{v} \cdot \vec{\omega} = \frac{1}{2} \vec{v} \times \vec{B}\), which is Equation 15.3 (page 222).
14.1  Dots and crosses
Prove the identity $(\mathbf{A} \times \mathbf{B}) \cdot (\mathbf{C} \times \mathbf{D}) = (\mathbf{A} \cdot \mathbf{C})(\mathbf{B} \cdot \mathbf{D}) - (\mathbf{A} \cdot \mathbf{D})(\mathbf{B} \cdot \mathbf{C})$.

14.2  Only one rank-2 tensor from Heaven
Chapter 49 will argue that the propagation of light through a medium of randomly-oriented molecules involves the average of the polarizability tensor over rotations. Perhaps it already seems reasonable to add, “That average will always be a constant times the identity tensor,” but let’s prove this.

Note first that the rotational average must itself be a rotationally-invariant, symmetric 3-tensor. Call it $\mathbf{A}$; then its matrix of components in some coordinate system must have the property that $S^I \mathbf{A} S^J = \mathbf{A}$ for any rotation matrix $S$. In particular, this property holds for any infinitesimal rotation. Recall from Equation 13.5 (page 203) that an infinitesimal rotation is given by $S = \mathds{1} + \epsilon \mathbf{T} + O(\epsilon^2)$, where $\mathbf{T}$ is an antisymmetric matrix and $\mathds{1}$ is the identity matrix.

Work out the consequences of invariance under such transformations (to order $\epsilon$) and prove that $\mathbf{A}$ can only be a constant times $\mathds{1}$.

14.3  Liquid crystals
Background: Let’s illustrate the utility of tensor methods in another branch of physics. Suppose someone tells you that some kind of matter (an “isotropic ferromagnet”) has states characterized by a spatially varying 3-vector field $\mathbf{u}(\mathbf{r})$ (the “order parameter”). The energy cost to be in one of these states is some analytic, local, rotationally invariant function of $\mathbf{u}$ and its derivatives, integrated over space. Because it’s analytic, we can expand the energy in Taylor series as a polynomial in the components of $\mathbf{u}$. Clearly the part of this function with no spatial derivatives must involve only even powers of the components $u_i$. This simple observation has profound consequences for the phase-transition behavior of ferromagnets.

Now suppose someone else tell you that some other kind of matter (a “nematic liquid crystal”) has states characterized by a spatially varying, symmetric, traceless rank-2 tensor $\mathbf{M}$. The free energy cost to be in one of these states is some analytic, local, rotationally invariant function of $\mathbf{M}$ and its derivatives, integrated over space. Because it’s analytic, we can expand the energy in Taylor series as a polynomial in the components of $\mathbf{M}$. The part of this function with no derivatives must be at least quadratic in the components of $\mathbf{M}$ (why?).

Do: Now find all possible contributions to the free energy cost function (if any) that are quadratic or cubic in the components of $\mathbf{M}$ (again, only consider terms with no spatial derivatives). Your answer has profound consequences for the phase-transition behavior of nematic liquid crystals.

14.4  $T_2$ Nematic–polar interaction
[[Not ready]]
Ørsted received his PhD in 1799 in the medical faculty of Copenhagen; his topic dealt with Kant's philosophy. [His] discovery, easy to reproduce, was the first direct demonstration of the connection between electricity (a current) and magnetism, and it was first done by accident at the end of a lecture demonstration. Interestingly, Ørsted was apparently all “thumbs” in the lab, and all his experiments had to be carried out by his students and assistants.

— R. M. Clegg

For twenty years no one tried the action of the voltaic pile on a magnet. . . . Coulomb’s hypothesis on the nature of magnetic action . . . rejected any idea of action between electricity and the so-called magnetic wires. This prohibition was such that when Arago spoke of [Ørsted’s] phenomena at the Institute, they were rejected . . . . Everyone decided that they were impossible.

— Ampère, to a friend

15.1 FRAMING: INTEGRABILITY

We have already started thinking about charges in motion, but we have not yet considered the magnetic fields that they create. That simplification is justified if charges are motionless, or in disordered motion. It may also extend to situations where they move slowly, so that any magnetic fields they create if any do not react back on them, nor create significant electric fields. Also, sometimes we assumed that the charges were constrained to specified motions, so that any forces they might get from magnetic fields were unimportant.

Nevertheless, magnetic fields generated by even slowly-moving charges can be significant if those charges are sufficiently numerous. So let's begin studying that situation. We'll invent a formulation, involving a new kind of potential function, whose existence will again follow from (a new kind of) integrability lemma. This vector potential will prove to be just as useful as the corresponding construction was in electrostatics.

Electromagnetic phenomenon: Tiny magnetic field disturbances can reveal brain activity without requiring invasive probes.

Physical idea: [[Not ready]].
15.2 A NEW FORCE AWAKENS

Imagine a steady current through a long, straight wire. There is no net charge anywhere to create any electric field. A test charge $q$ outside that wire will feel a kind of force that we have not yet encountered. It differs from the electrostatic force in that:

- The force is zero unless the test charge is moving; and
- The force (a vector) is a linear function of the velocity (a vector).

Section 13.3.1 called such a function a 3-tensor of rank two:

$$ (\text{force}) = 2q\hat{\omega} \cdot \vec{v}, \quad \text{that is,} \quad f_i = 2q\hat{\omega}_{ij}\vec{v}_j. \quad (15.1) $$

After someone sets up a current distribution in the lab, we can operationally measure the resulting field $\hat{\omega}$ by throwing a lot of charged test bodies and seeing how they accelerate. Moreover, experimentally the new force has another unusual property:

- The force is always perpendicular to the test charge’s velocity.

This observation implies that the current specifically creates an antisymmetric rank-two tensor $\hat{\omega}$. To see this, think about two velocities $\vec{v}$ and $\vec{u}$. Then $\hat{\omega} \cdot (\vec{v} + \vec{u})$ must be perpendicular to $(\vec{v} + \vec{u})$:

$$ 0 = (\vec{v} + \vec{u}) \cdot \hat{\omega} \cdot (\vec{v} + \vec{u}) = \vec{v} \cdot (\hat{\omega} \cdot \vec{v}) + \vec{u} \cdot (\hat{\omega} \cdot \vec{u}) + \vec{v} \cdot \hat{\omega} \cdot \vec{u} + \vec{u} \cdot \hat{\omega} \cdot \vec{v}. $$

The first two terms are zero by assumption, so the last two must always sum to zero, regardless of what $\vec{u}$ and $\vec{v}$ may be. That requires $\hat{\omega}$ to be antisymmetric.

Prior to now, well-meaning but misguided people may have thought you weren’t ready for tensors, so they repackaged the magnetic field by defining the three quantities

$$ \tilde{B}_i = \varepsilon_{ijk} \hat{\omega}_{jk}. \quad (15.2) $$

In a sense, we lose nothing by this reformulation, because it is invertible: We can always recover $\hat{\omega}$ from $\tilde{B}$:

**Your Turn 15A**

Show that

$$ \hat{\omega}_{im} = \frac{1}{2}\varepsilon_{kim}\tilde{B}_k. \quad (15.3) $$

(Where did the factor of $1/2$ come from?)

Introductory texts formulate magnetism in terms of $\tilde{B}$, and so will we at first. But there is a price to pay for this approach:

---

1. We will see later that this formula remains valid in relativistic situations, if we interpret the left side as the time derivative of particle momentum. Putting the factor of 2 in the definition Equation 15.1 is convenient because this convention eliminates another 2 in Equation 15.2.

2. See Media 7.
• Equation 15.2 involves a Levi-Civita tensor, and hence requires us to choose a handedness on space before we can even say what is “the” magnetic field in some experimental situation. In contrast, Equation 15.1 defines $\vec{\omega}$ in terms of two directly measurable physical quantities (velocity and force).

• Using $\vec{B}$ instead of $\vec{\omega}$ also introduces Levi-Civita tensors (via cross product and curl) into our equations of physics, obscuring their inversion symmetry. For example, the force law Equation 15.1 becomes $\vec{f} = q\vec{\omega} \times \vec{B}$.

• Later, we'll see that $\vec{B}$ also obscures the Lorentz invariance of electrodynamics, which is one reason why it took a genius (Lorentz) to see that property, another genius (Einstein) to see the implications, and a third genius (Minkowski) to make sense of it! Indeed, we will abandon $\vec{B}$ later, in order to construct a formulation in which even mortals can see the full invariance at a glance.\(^3\)

• There is nothing physical that points along $\vec{B}$! Certainly not the force. So $\vec{B}$ is no less abstract than $\vec{\omega}$.

• Section 15.3 will obtain a useful result whose full generality is apparent only in the $\vec{\omega}$ language.

Despite those criticisms, we do need to be able to talk to people who use $\vec{B}$. So we need to be able to switch between both representations, by using Equations 15.2 and 15.3.

**Your Turn 15B**

The magnetic field outside a cylindrical bar magnet is traditionally represented as arrows that sprout from the N pole, then return to dive into the S pole. Nothing about that picture seems to choose clockwise over counterclockwise rotation. Moreover, the metal casing in Media 6, and the fluid in Media 7, carry current radially from the center, again not privileging either sense of rotation. And yet both demonstrations show no indecision about which way to whirl around. How can this be?

Section 15.2\(^4\) (page 235) raises a puzzle about velocity-dependent forces.

### 15.3 VECTOR POTENTIAL

#### 15.3.1 No scalar potential this time

In electrostatics, the four equations $\vec{V} \cdot \vec{E} = \rho_0/\varepsilon_0$ and $\vec{V} \times \vec{E} = \vec{0}$ boiled down to just one equation for one potential function (the Poisson equation). That was handy. It worked because we found a general solution to Faraday’s law, $\vec{V} \times \vec{E} = \vec{0}$, in terms of $\psi$, so we could just substitute $\vec{E} = -\nabla \psi$ into the Gauss law and forget Faraday. Can we duplicate that victory?

---

\(^3\) It may seem that abandoning a vector description of magnetism would obscure electric/magnetic duality. On the contrary, when we unify electric and magnetic fields into a single object, there will be a “duality” transform on that object under which the Maxwell equations in vacuum are invariant (Section 34.9\(^4\), page 544).
At first it looks bad. The magnetic field is not curl-free: Ampère’s law says $\nabla \times \vec{B} \neq 0$. It’s true that sometimes we want to solve for magnetic fields throughout a current-free region, and in such a case we may get some success by introducing a “magnetic scalar potential.” But let’s instead try to exploit the magnetic Gauss law, $\nabla \cdot \vec{B} = 0$, because it’s always true.

15.3.2 Lemma to a lemma

Let’s brush up on a point we’ll need soon. Suppose that $f$ is a scalar function of $\vec{r}$. We can construct a function of four variables, $g(u, \vec{r})$, by evaluating $f$ at the point $(u \vec{r})$. Make sure that you understand how the Chain Rule implies that

$$\frac{\partial g}{\partial u} = \frac{\partial f}{\partial r_i} \bigg|_{u \vec{r}} \frac{\partial (u \vec{r})}{\partial u} = \vec{\nabla} f \bigg|_{u \vec{r}}$$

(15.4)

and similar results for $\frac{\partial g}{\partial r_2, 3}$. Think about how the indices match on each side of these formulas.

15.3.3 Revisit electrostatics

The magnetic Gauss law $\nabla \cdot \vec{B} = 0$ looks pretty different from $\nabla \times \vec{E} = 0$, but surprisingly there is a close analogy. To bring it out, let’s return briefly to electrostatics. Previously we invoked Stokes’s theorem, along with the static Maxwell equation $\nabla \times \vec{E} = 0$, to conclude that the line integral of $\vec{E}$ was independent of the path chosen to $\vec{r}$. Then a clever choice of path made it easy to find the gradient of $\psi$. 4

What if we didn’t know Stokes’s theorem? We could instead make a standard choice of path, for example, “the straight line from the origin to $\vec{r}$.” Then $\psi$ is well defined. Computing its gradient is a bit more tricky than before, but working it out also sets us up for the generalization we need in magnetism.

**Ex.**

Show that the curl-free condition is equivalent to

$$\nabla_i \vec{E}_j - \nabla_j \vec{E}_i = 0 \text{ for any } i \text{ and } j \text{ (stationary case).}$$

(15.6)

**Solution:** One way is to write out explicitly each component of the equation, for example, $(\nabla \times \vec{E})_1 = 0$ and so on. But let’s get some practice with Levi-Civita identities: Take the curl and contract it with $\varepsilon$:

$$0 = \varepsilon_{ijk}(\nabla \times \vec{E})_k = (\varepsilon_{ijk} \delta_{kl} \varepsilon_{jm}) \nabla_l \vec{E}_m.$$

Now use the identity Equation 14.14 (page 215) to simplify the factor in parentheses:

$$0 = (\delta_{ik} \delta_{jm} - \delta_{im} \delta_{jk}) \nabla_l \vec{E}_m = \nabla_i \vec{E}_j - \nabla_j \vec{E}_i.$$

---

4Section 2.2.1 (page 28).
We are exploring the proposal to integrate $\vec{E}$ along a straight line from a reference point (for simplicity the origin) to a desired field point $\vec{r}$. We can express that path in parametric form by the formula $\vec{r} = u \vec{r}$ where $\vec{r}$ is fixed and $u$ ranges from 0 to 1. Substituting into Equation 2.2 (page 28) gives

$$\psi(\vec{r}) = - \int_0^1 (\vec{r} \, du) \cdot \vec{E}(u \vec{r}).$$

(15.7)

In this expression, $\vec{r}$ is held constant during the integration over $u$. Then the negative gradient is (see Equation 15.5)

$$- \frac{\partial \psi}{\partial r_i} = \int_0^1 \frac{du}{u} \left[ \vec{E}_m(\vec{u} \vec{r}) \frac{\partial \vec{r}_m}{\partial \vec{r}_i} + \vec{r}_m \frac{\partial \vec{E}_m}{\partial \vec{r}_i} \right] \left[ \frac{\partial (u \vec{r}_k)}{\partial \vec{r}_i} \right].$$

(15.8)

In the last term, we may replace $\frac{\partial \vec{E}_m}{\partial 

\frac{\partial \vec{r}_i}}$ by $\frac{\partial \vec{E}_m}{\partial \vec{r}_i}$, thanks to Equation 15.6. We can now use Equation 15.4 and the Fundamental Theorem of Calculus to find

$$- \nabla_i \psi |_{\vec{r}} = \int_0^1 \frac{du}{u} [u \vec{E}_i(u \vec{r})] = u \vec{E}_i(u \vec{r}) |_{u=0}^1 = \vec{E}_i(\vec{r}).$$

Summarizing:

- Once again, we have established the potential representation for electrostatics.
- It still relies on the curl-free condition, even though we did not explicitly use Stokes’s theorem.
- There is an ambiguity in $\psi$: Adding a constant to $\psi$, for example, by choosing a different reference point, won’t change its gradient.
- The payoff for the potential formulation is again that we have fewer and simpler equations to solve.\(^5\)
- The caveat is that we’ll need to rethink when we go beyond statics, because then $\nabla \times \vec{E} \neq 0$. Chapter 18 will pick up this loose thread.

### 15.3.4 The magnetic Gauss law expresses an integrability condition

We’d like an integrability lemma like the one just given, but applicable to magnetism. First we’ll uncover a hidden analogy to electrostatics.

**Your Turn 15C**

Use Equation 15.2 to show that the magnetic Gauss law is equivalent to

$$\varepsilon_{imk} \nabla_k \omega_{im} = 0.$$  

(15.9)

\(^5\)In fact, Chapter 2 found a complete, general solution to electrostatics with a specified charge distribution.
That is, when we take all the first derivatives of \( \dot{\omega}_{lm} \) and antisymmetrize, the result is always zero. This resembles Equation 15.6, albeit with an extra index.

**Your Turn 15D**  
Show that, of the six nonzero terms on the right side of Equation 15.9, half are redundant; that is, it may be written as

\[
\vec{\nabla}_k \ddot{\omega}_{lm} + (2 \text{ cyclic permutations}) = 0 
\]

(15.10)  
for any \( k, i, \) and \( m \).

15.3.5 The Poincaré lemma applies in any number of dimensions, and to tensors of any rank

With this preparation, we’re ready to generalize Section 15.3.3. Analogously to Equation 15.7, define

\[
\vec{A}_i(r) = 2 \int_0^1 (u^m \frac{d}{du}) \ddot{\omega}_{ml}(u \hat{r}).
\]

(15.11)

We want the curl of this new vector field, or equivalently

\[
\vec{\nabla}_k \dddot{A}_i - \vec{\nabla}_i \dddot{A}_k = 2 \int_0^1 du \left[ \frac{\partial \ddot{r}_m}{\partial \ddot{r}_k} \ddot{\omega}_{ml}(u \hat{r}) + \ddot{r}_m \frac{\partial \ddot{\omega}_{ml}}{\partial \ddot{r}_n} \bigg|_{ur \to \ddot{r}} \ddot{r}_n \right] - (i \Leftrightarrow k)
\]

\[
= 2 \int_0^1 du \left[ \left( \ddot{\omega}_{ki} - \ddot{\omega}_{ik} \right) + u \dddot{r}_m \left( \frac{\partial \ddot{\omega}_{ml}}{\partial \ddot{r}_k} - \frac{\partial \ddot{\omega}_{mk}}{\partial \ddot{r}_i} \right) \bigg|_{ur} \right].
\]

The first two terms can be written as \(2 \ddot{\omega}_{ki}\). The last two terms can be simplified by using Equation 15.10: They equal \(-\vec{\nabla}_m \ddot{\omega}_{ik} \bigg|_{ur}\).

Analogously to Equation 15.8, we therefore get

\[
= 2 \int_0^1 du \left[ 2u \ddot{\omega}_{ki}(u \hat{r}) - u \dddot{r}_m \frac{\partial \ddot{\omega}_{ik}}{\partial \ddot{r}_m} \bigg|_{ur} \right]
\]

\[
= 2 \int_0^1 du \left[ \dddot{u} \ddot{\omega}_{ki}(u \hat{r}) \right] = 2u^2 \dddot{\omega}_{ki}(u \hat{r}) \bigg|_0^1 = 2\dddot{\omega}_{ki}(r).
\]

(15.12)

Now tidy things up by recalling the formula for curl and Equation 15.2:

\[
(\vec{\nabla} \times \vec{A})_m = \varepsilon_{mkl} \vec{\nabla}_k \vec{A}_l = \frac{1}{2} \varepsilon_{mkl} (\vec{\nabla}_k \dddot{A}_l - \vec{\nabla}_l \dddot{A}_k) = \varepsilon_{mkl} \dddot{\omega}_{kl} = \dddot{B}_m.
\]

Indeed, Equation 15.11 constructs a vector field \( \vec{A} \) whose curl equals \( \dddot{B} \). So we’ll call \( \vec{A} \) the magnetic vector potential.

---

6See Section 13.3.1 (page 196).

7Beware that most books also use the same letter \( A \) to denote a different quantity, the 4-vector potential. Later, we will disambiguate by using \( \vec{A} \) for the 3-vector and \( A \) for the 4-vector potentials.
Our payoff for this level of abstraction is that the result we proved works in \textit{any number} of dimensions:

\textit{Any antisymmetric rank-two tensor with the property that its antisymmetrized first derivatives vanish (Equation 15.9) may be written as the antisymmetrized tensor of derivatives of some vector field (Equation 15.12).} \vphantom{15.13}

Later, when we need this result in four dimensions, we won’t need to prove it again.\footnote{It even works for tensors of rank different from two. For example, applied to rank 1, it’s just what we proved in Section 15.3.3.}

Notice that Equation 15.11 makes no use of the Levi-Civita tensor. Hence \( \vec{A} \) is meaningful without requiring any convention about which hand is “right.”

\section*{15.4 GAUGE INVARIANCE AND COULOMB GAUGE}

We have found the general solution to the magnetic Gauss law, so we can just substitute \( \vec{V} \times \vec{A} \) for \( \vec{B} \) into Ampère’s law and forget about Gauss. However, there is an ambiguity in this representation. After all, if we add the gradient of anything, \( \vec{A} \rightarrow \vec{A} = \vec{A} + \vec{\nabla} \Xi \), then the curl of \( \vec{A} \) doesn’t change. So \( \vec{B} \) doesn’t fully determine its vector potential \( \vec{A} \). This fact is known as \textbf{gauge invariance}, although perhaps “indeterminacy” would have been a better term. The substitution \( \vec{A} \rightarrow \vec{\tilde{A}} \) is called a \textbf{gauge transformation} of the vector potential. This is much more freedom than what we had in electrostatics, where adding a \textit{constant} to \( \psi \) left \( \vec{E} \) unchanged.

Gauge invariance sounds like a nuisance, but it can be helpful. We can use that freedom to represent a magnetic field by a vector potential that additionally satisfies some extra condition (\textit{gauge fixing}). For example, we can always insist that \( \vec{\tilde{A}} \) obey

\begin{equation}
\vec{\nabla} \cdot \vec{\tilde{A}} = 0. \quad \textbf{Coulomb gauge condition} \end{equation}

To see this, suppose that we represent a \( \vec{B} \) field by some vector potential that doesn’t satisfy Coulomb gauge. After a gauge transformation, \( \vec{\tilde{A}} \) must obey \( \vec{\nabla} \cdot \vec{A} = \vec{\nabla} \cdot \vec{A} + \vec{V} \Xi \). We just need to choose \( \Xi \) to be a function of position that solves the Poisson equation\footnote{Section 2.3.2 (page 31) found a general solution to the Poisson equation.} with source given by \( \vec{\nabla} \cdot \vec{A} \). After that gauge transformation, \( \vec{\tilde{A}} \) is in Coulomb gauge.

\section*{15.5 BACK TO PHYSICS}

\subsection*{15.5.1 Steady currents}

To avoid distraction from electrostatics, let’s temporarily assume that there is no free net charge (\( \rho_\text{q}(\vec{r}) = 0 \)); charge may nevertheless be moving (\( \vec{j} \neq 0 \)).

The results in Sections 15.3–15.4 are valid regardless of whether the fields are time-dependent or not. But before we work up to full dynamics, the remainder of this chapter...
will temporarily restrict to situations with \textit{steady} motion ($\partial \vec{J}/\partial t = 0$). Thus, our system will be unchanged upon time \textit{translation} (it is \textbf{stationary}), though not under time \textit{reversal} (it is not \textbf{static}). Somewhat inconsistently, the study of such situations is often called \textbf{magnetostatics}.

Stationarity is only an idealized, approximate situation. Really each electron or proton is pointlike, so as any one of them passes any point, the electric and magnetic fields pulse. We replace discrete charges by a continuous “river of charge,” an approximation that certainly makes sense in a macroscopic apparatus. The overall river can be considered as flowing steadily if we neglect its granular character in this way. (Later chapters will upgrade to a fully dynamic formulation.)

The approach in this book is to take the Maxwell equations as a physical hypothesis and explore their testable consequences. In the stationary situation just described, they simplify to just

\begin{align*}
\nabla \cdot \vec{E} &= \rho_\text{q}/\varepsilon_0 = 0 \quad \text{Gauss (no net charge)} \\
\nabla \cdot \vec{B} &= 0 \quad \text{Gauss} \\
\nabla \times \vec{B} &= \mu_0 \vec{J} \quad \text{Ampère (stationary case)} \\
\nabla \times \vec{E} &= 0 \quad \text{Faraday (stationary case)}
\end{align*}  

These equations have decoupled into two that involve $\vec{E}$ only, and whose solution is $\vec{E} = 0$, plus two that involve $\vec{B}$ only. They have falsifiable content because $\vec{B}$ has an independent definition: We can measure it throughout space by looking at the motions of test charges, which feel the force given in Equation 15.1. Once $\vec{B}$ is measured, we can check if it does or does not obey the above equations for a steady current distribution.

\subsection*{15.5.2 Axial symmetry suggests a solution to the Ørsted problem}

In the most basic situations, we can guess a trial solution to Equations 15.15 and adjust it until it works: Imagine an infinite, straight wire along the \textit{z} axis carrying steady current $I$ uniformly distributed across its cross-section and directed along $+\hat{z}$. This situation has so much symmetry that we can try a trial solution where $\vec{B}$ is everywhere pointing radially outward from the wire. That fails. But the next possibility, in which $\vec{B}(\vec{r}) = f(r)\hat{\phi}$, is also axially symmetric and more promising. We integrate Ampère’s law over a disk of radius $w$ perpendicular to and centered on the wire:

\begin{align*}
\text{Ampère: } &\int d^2 \Sigma \cdot (\nabla \times \vec{B}) = \mu_0 \int d^2 \Sigma \cdot \vec{J} = \mu_0 I \\
\text{Stokes: } &\oint \vec{B}(r_e) \cdot d\vec{r}_e = \int_0^{2\pi} (w d\phi) \vec{B} \cdot \hat{\phi} = 2\pi w f(w).
\end{align*}

We conclude that $f(w) = \mu_0 I/(2\pi w)$ for any $w$ larger than the wire’s radius, and hence that

$$\vec{B}(\vec{r}) = \phi \frac{\mu_0 I}{2\pi r},$$

the famous answer.
15.5 Back to Physics

Other problems are harder than this one, however. We need a more systematic approach.

15.5.3 The same Green function from electrostatics also solves the magnetostatic equations

Sections 15.3–15.4 showed that any magnetic field can be represented in terms of a divergence-free vector potential.

Your Turn 15F

To see the power of this observation, first show that Ampère’s law may be written as

\[ \nabla^2 \vec{A} = -\mu_0 \vec{j} \]  

in Coulomb gauge  \hspace{1cm} (15.17)

That scary vector partial differential equation has separated nicely into three independent copies of the Poisson equation. And we already know how to solve the Poisson equation, from electrostatics (Equation 2.7, page 31)! For each component of \( \vec{j} \), compute

\[ \vec{A}_i(\vec{r}) = \mu_0 \int d^3r' \frac{\vec{j}(\vec{r}')}{4\pi ||\vec{r} - \vec{r}'||} \]  

(15.18)

So we just finished magnetostatics, for situations where we are given the current distribution: Evaluate Equation 15.18 for the three components of \( \vec{A} \). Then compute the curl to get \( \vec{B} \).

15.5.4 Self-consistency

Before we accept Equation 15.18, we should check that it really is a potential in Coulomb gauge. If not, then the fact that it solves Equation 15.17 would be irrelevant, because Equation 15.17 is not Ampère’s law except in Coulomb gauge. To do this, first consider a function of two vectors that depends only on the magnitude of their difference: \( G(\vec{r}, \vec{r}') = g(||\vec{r} - \vec{r}'||) \).

---

\(^{10}\)The notation \( \nabla^2 \vec{A} \) means that we apply the Laplace operator to each component of \( \vec{A} \) and interpret the results as the components of a vector. This operation only makes sense in cartesian coordinates; a more elaborate form of the derivation is needed in curvilinear coordinates.
Your Turn 15G

a. Show that
\[ \frac{\partial G}{\partial r} = -\frac{\partial G}{\partial r'} . \]

b. Differentiate under the integral in Equation 15.18 and apply your result in (a).

c. Use the continuity equation to show that \( \vec{V} \cdot \vec{J} \) must be zero in a steady situation.

d. Use your results in (a,b) to establish that \( \vec{A} \) is in Coulomb gauge.

15.5.5 Some of the equations are automatically satisfied, resolving a counting puzzle

The equations of electro- and magnetostatics (Equation 15.15) appear to be overdetermined: eight equations in just six unknown functions \( \vec{E}_i, \vec{B}_i \), an issue first raised in Hanging Question #D (page 14). And yet, we previously reformulated electrostatics as one equation in one unknown: the Poisson equation (Equation 2.4, page 29). Also, magnetostatics boiled down to three Poisson equations for the three components of \( \vec{A} \) (Equation 15.17). So at least in statics, our puzzle has disappeared: We really have a total of four equations in the four unknown potential functions (\( \psi \) and \( \vec{A} \)).

To reconcile the two approaches, note that two of the eight Equations 15.15 are identities; they do not constrain the fields and hence should not be included in our count. For example, taking the divergence of both sides of the Faraday law gives \( \nabla \cdot \vec{E} = 0 \) identically, regardless of what \( \vec{E} \) may be. Similarly, taking the divergence of both sides of Ampère’s law (Equation 15.15) gives the single equation
\[ \vec{V} \cdot (\nabla \times \vec{B}) = \vec{V} \cdot \vec{J} . \] (stationary case)

The left side is identically zero, regardless of what \( \vec{B} \) may be; the right side is also automatically zero because Equations 15.15 assume time-independence (recall the continuity equation). So again, we end up with equal numbers of unknowns (the six components of \( \vec{E} \) and \( \vec{B} \)) and equations (the remaining six Maxwell equations). Reformulating in terms of potentials just made this consistency more evident.

15.6 BIOT–SAVART FORMULA

15.6.1 Second solution to Ørsted, via vector potential

The previous sections got a bit abstract. Let’s see how the story plays out in some familiar problems. First, we’ll revisit the Ørsted problem: Suppose that a thin, straight, infinite wire carries current \( I \) directed along \( +\hat{z} \), as in Section 15.5.1. Thus, its charge flux is
\[ \vec{J}(\vec{r}) = I \delta^{(2)}(\vec{r}_\perp) \hat{z} . \] (15.19)

Here \( \vec{r}_\perp \) denotes the two-component vector \( \begin{bmatrix} x \\ y \end{bmatrix} \). Each delta function contributes a dimension \( L^{-1} \), so this expression has dimensions appropriate for a 3D charge flux.\(^{11} \) We already

\(^{11} \)See Section 0.3.8 (page 11).
found the resulting magnetic field in Section 15.5.1.

**Your Turn 15H**

a. Do it again, this time by using potentials: Solve Equation 15.17 with source given by Equation 15.19. [*Hint:* The Green-function solution given in Section 15.5.3 isn’t the easiest way to do this problem, which has lots of useful symmetry. Instead, make a trial solution, then check and adjust it.]
b. Confirm that the vector potential you found really is in Coulomb gauge, as we argued generally must be the case.
c. Finally, work out the curl of your answer and confirm it’s what was already found in Section 15.5.1.

**15.6.2 \( \vec{B} \) for a general current distribution**

We can now go beyond the Ørsted problem and find the magnetic field created by an arbitrary current distribution.

**Your Turn 15I**

Show that the curl of Equation 15.18 is

\[
\vec{B}(\vec{r}) = \frac{\mu_0}{4\pi} \int d^3\vec{r}_s \vec{j}(\vec{r}_s) \times \frac{\vec{r} - \vec{r}_s}{||\vec{r} - \vec{r}_s||^3}, \quad \text{stationary case} \quad (15.20)
\]

This is a generalization of the usual Biot–Savart formula to cover an arbitrary current distribution (not necessarily confined to a thin wire).

**15.6.3 More about thin wire approximation**

Sometimes it is appropriate to consider a limiting case in which \( \vec{j} \) is everywhere zero except along a mathematical curve (a “thin wire”). We already considered the simplest case in Equation 15.19.

In a static situation, the continuity equation implies that the total current \( I \) has the same value when computed through any cross-section of the wire. Suppose that the wire is described by a parameterized curve in space \( \vec{\ell}(s) \). For example, we could choose \( s \) to be arc length along the curve. Then at any point \( s_0 \), the current is flowing parallel to the tangent vector, that is, to the unit tangent \( d\vec{\ell}/ds|_{s_0} \).

Start by considering just one chunk of wire, of length \( ds \) and centered at \( s_0 \) (Figure 15.1). Choose a coordinate system centered on \( \vec{\ell}(s_0) \), and rotated so that the tangent lies along \( z \). Chapter 8 explained how to find the \( z \)-component of the charge flux: Find the net charge crossing the surface element shown in the figure, from smaller to larger \( z \), during time \( dt \). That charge equals \( I dt \) if the element \( dx dy \) includes the wire (at the origin); otherwise, it’s zero. Idea 8.5 (page 117) defined \( \vec{j}_z \) as a function that, when integrated
Chapter 15 Magnetostatics

Figure 15.1: Thin-wire idealization.

Think about why this formula has the units appropriate for a 3D charge flux.

We can make our formula less dependent on a specific choice of coordinates. First, notice that the one chunk of wire we considered is also confined to a limited range $dy = (dy/ds)ds$ near $y = 0$. With that observation, we get the more general form

$$\delta J(\vec{r}) = I\delta(\vec{r} - \vec{r}(s)) \frac{d\vec{r}}{ds} ds.$$  \hspace{1cm} (15.21)

This formula has the same dimensions as Equation 15.19, but it’s no longer restricted to any special coordinate system, nor to one particular point on the wire. To get the charge flux set up by the entire wire, integrate Equation 15.21 over its entire arc length.

Your Turn 15J

Substitute Equation 15.21 into Equation 15.20 and recover the usual form of the Biot–Savart law.

15.7 BOUNDARY CONDITIONS

Regardless of whether we use the potential formalism, the magnetic Gauss law implies a no-jump condition for the magnetic field across a boundary, similar to the one in electrostatics but without any dependence on the behavior of charges or currents at the surface:

$$\Delta \vec{B} = 0.$$  \hspace{1cm} (15.22)

This fact can be especially useful if we know that the magnetic field is zero on one side. For example, superconductors exclude magnetic fields, so $\vec{B} = 0$ just outside as well.

Similarly, integrating Ampère’s law around a loop near the surface gives a condition on the tangential component of $\vec{B}$ (Figure 15.2b). We must allow for the possibility of

---

12See Figure 15.2a; compare Section 6.10 (page 90).
Boundary conditions near an interface. (a) The short red cylinder has one end cap just outside a material and the other just inside. Integrating the magnetic Gauss law over it, and using the divergence theorem, shows that the component of $\mathbf{B}$ perpendicular to the surface must be the same just inside and outside the material (Equation 15.22).
(b) The shallow red rectangle has one of its longer edges just outside a material and the other just inside. Integrating Ampère’s law, and using Stokes’s theorem, shows that any $\mathbf{B}$ component parallel to the surface may jump if there is a surface current layer (Equation 15.23).

Your Turn 15K

Show that

$$\Delta \mathbf{B}_1 = \mu_0 j^{[2D]} \times \hat{n}. \tag{15.23}$$

Here $\Delta \mathbf{B}_1 = (\mathbf{B}^{[2]} - \mathbf{B}^{[1]})_//$, and $\hat{n}$ is the unit perpendicular vector pointing from region 1 to region 2.

15.8 MAGNETOENCEPHALOGRAPHY

[[Not ready]]

15.9 PLUS ULTRA

Section 15.5.1 found the general solution to magnetostatics with specified, steady currents. But we actually got much more: We also found a simplified formulation of the equations

---

13See Section 8.3.1 (page 116). Some books call this quantity “surface current density,” but it is neither a volume density nor an areal density.
that involves a potential (in this case a vector potential), and it works, even in nonstationary situations (because \( \nabla \cdot \vec{B} = 0 \) always). Chapter 34 will find an analogous object that combines the vector potential with electric potential, and that, unlike our previous construction of \( \psi \), remains valid beyond statics.

Section 15.9' a (page 235) mentions hypothesized magnetic monopoles. Section 15.9' b says more about \( \vec{B} \) versus \( \vec{\omega} \). Section 15.9' c connects our constructions to more advanced, and general, mathematics.

**FURTHER READING**

*Intermediate:*
Sections 15.3.3–15.3.5 follows the explicit construction in Spivak, 1999, vol. 1.

Differential forms and Poincaré lemma: Stone & Goldbart, 2009; Spivak, 1999, vol. 1; Hubbard & Hubbard, 2007; Burke, 1985. (Note that some authors call this result the “converse of the Poincaré lemma.”)

*Technical:*

15.2′ Puzzle about angular momentum conservation

Maybe you recall from first-year physics that the proof of angular momentum conservation, as presented even in the Feynman Lectures, involves the assumption that every force on any particle is directed along the line joining that particle to another one. That certainly is not guaranteed with magnetic forces, whose direction depends on the velocity of the particles. What happens to angular momentum conservation? Chapter 35 will get back to this, but the spoiler is: It survives, once we correctly attribute angular momentum to the fields themselves.

15.9’a About magnetic monopoles

The magnetic monopoles predicted by grand unified theories, if observed, would seem to invalidate the discussion in Section 15.3.1: A point source of \( \mathbf{B} \) implies that \( \nabla \cdot \mathbf{B} \neq 0 \) somewhere. Indeed, inside such hypothetical objects there is always a region in which classical electrodynamics breaks down altogether (other fields like the ones associated to the W and Z bosons have nonzero expectation values).\(^{14}\) But magnetic monopoles haven’t (yet) been observed experimentally in free space.\(^{15}\)

Quite apart from such theoretical concerns, E. Parker realized that the observed filamentous structures in distant galaxies is evidence for large-scale magnetic fields, and that this observation in turn implies a severe bound on the hypothetical existence of magnetic monopoles. Just as free electric charges terminate electric field lines in a conductor, so also free magnetic charges (if they existed) would terminate magnetic field lines. The fact that cosmic magnetic fields are observed then implies a limit on the abundance of free magnetic charges. For a review of magnetic monopoles and flux limits, see, for example, Preskill, 1984.

15.9’b Elimination of \( \mathbf{B} \)

The main text pointed out the conceptual benefits of formulating magnetic effects in terms of the tensor \( \mathbf{\omega} \), not the traditional \( \mathbf{B} \). For example, to measure \( \mathbf{\omega} \) we need not first choose any coordinate system and arbitrarily anoint it as “right-handed.” In fact, every “pseudovector” quantity in classical physics, including angular velocity and angular momentum, can be eliminated in favor of tensor quantities, whereupon all the cross products appearing in rigid body dynamics and so on disappear and everything is manifestly inversion-invariant (Problem 15.4).

Is this distinction just Puritanical fussiness? First, notice that you rarely see any physics formulas involving the sum \( \mathbf{E} + c\mathbf{B} \), any more than you ever see people adding momentum to angular momentum (or temperature to velocity). Temperature and velocity have different tensor structures; it’s not meaningful to add them, and the same for electric and magnetic fields.\(^{16}\) It’s a quirk of three dimensions that they happen to have the same numbers of independent components, but nevertheless they are incompatible objects. Second, Section 15.3.5 showed a deep analogy that only becomes apparent when we abandon the superficial analogy obtained by representing magnetism by \( \mathbf{B} \). Third, and most important, everybody does agree that \( \mathbf{B} \) has got to be scrapped when we unify

\(^{14}\)See Problem 17.2.

\(^{15}\)There may be collective excitations in condensed matter with this character.

\(^{16}\)One exception will arise in Chapter 51, when we consider light propagating in a medium that breaks inversion symmetry. Even in vacuum, some authors do introduce the “Riemann-Silberstein vector” \( \mathbf{E} + ic\mathbf{B} \), but the inversion invariance of the resulting formulas is then hidden.
electricity and magnetism and reformulate the theory relativistically in Chapter 34. Our destination is a formulation in which invariance under Lorentz transformations is explicit; when we arrive there, we’ll find that explicit invariance under inversions has come along for free.

Until that happy day, notice that in Equation 15.15, the Gauss law doesn’t care about the ambiguous sign of $\vec{B}$. The right-hand side of Ampère’s law involves only the true vector $\vec{j}$, but the left side has two sign changes if we switch handedness conventions, so it, too, is invariant, albeit secretly. Nevertheless, it is desirable to make the invariance explicit in each object separately.

### 15.9c Differential forms

Totally antisymmetric tensors are so useful that mathematicians have a separate name for them: differential forms of rank $p$, or just “$p$-forms” for short. Standard mathematical notation abbreviates by omitting the indices and over-arrows; you must remember the tensor character of each symbol from its original definition. The totally antisymmetrized first derivatives of the components of such a tensor form the components of a similar object of rank $p+1$. The operation just described is called the exterior derivative and denoted by the very concise symbol $d$. The exterior derivative operator has the property that $d^2 = 0$. Thus, applying $d$ to anything of the form $dA$ always yields zero.

The Poincaré lemma is a limited converse to the preceding statement: If $d\omega = 0$, then we may locally write $\omega = d\eta$ for some $(p-1)$-form $\eta$. There is an important caveat: “Locally” means that this result is valid only on a region of space that can be smoothly shrunken to a point. (On a torus, for example, we would not be able to choose an unambiguous path to each $\vec{r}$ as we did in Section 15.3.5, nor are different choices of path guaranteed to give answers that agree.) The study of exactly how the Poincaré lemma fails on a topologically nontrivial space is called deRham cohomology.

In this language:

- The existence of an electrostatic potential is the case $p = 1$. The Maxwell equation $dE = 0$ implies that we may write $E = -d\psi$. There’s an ambiguity: We may add any constant to the scalar potential $\psi$ without altering $d\psi$.
- For magnetostatics, we need the case $p = 2$: We found that the Gauss law for magnetic fields can be elegantly written as $d\omega = 0$, which implies that we may write $\omega = dA$ (Equation 15.12, page 226). There’s an ambiguity: Because $d^2 = 0$, we may add any gradient $d\vec{\xi}$ to the vector potential $A$ without altering $dA$. That’s gauge invariance (Section 15.4, page 227).

**Your Turn 15L**

Find equally elegant forms of Ampère’s law and the Stokes theorem.

**Your Turn 15M**

Show that changing the base point used in Equation 15.11 results in a gauge-transformed $\vec{A}$. 
15.1  *Jaws*

Let us explore a possible mechanism for sharks to navigate using Earth’s magnetic field. Given that a shark can detect an electric field strength of 0.5 $\mu V/m$, how fast would it have to swim through Earth’s magnetic field to experience an equivalent force on a charged test particle? Can sharks really swim that fast?

15.2  *Simplest possible electric motor*

Media 6 shows an electric motor consisting of a button magnet suspended from a 1.5 V cell, by a frictionless pivot. A wire brushes along the magnet’s rim, creating a circuit with the other terminal of the current source. Discuss why this magnet spins, and what determines the direction of its spin.

15.3  *Salt and pepper*

*Background:* Figure 15.3a shows a demonstration experiment involving salt, pepper, and an overhead projector. When an electric current is sent through the solution, ions migrate toward and away from the electrodes (electrophoresis), but there is no visible bulk flow of the water. Applying a static magnetic field, however, does lead to bulk (macroscopic) motion. It may seem implausible that the tiny little ions could pull hard enough on the surrounding water to get it moving! Let’s make some estimates.

The setup shown has a circular geometry. But to simplify the math, in this problem instead imagine a rectangular geometry (Figure 15.3b): Current passes between two parallel plates separated horizontally by distance $L = 5$ cm. The plates have width $w = 5$ cm and are immersed in a solution with depth $h = 1$ cm. The water between the plates contains sodium and chloride ions, each at number density (ions per volume) $c_{\text{ion}}$. Each ion carries electric charge $\pm e$, where $e$ is the charge on a proton. The solution consists of about one

![Figure 15.3: An experiment to demonstrate magnetoelectrophoresis.](image1)

(a) Pepper is sprinkled on the surface of a salt-water solution to visualize bulk flow. A central electrode sends direct current radially outward to a ring-shaped electrode at the rim of the dish. A magnet pole can be brought toward the surface (red). See also Media 7. (b) Simplified geometry for Problem 15.3.
gram of NaCl dissolved in volume \( Lhw \) of water. (The pepper is irrelevant.)

A total current of \( I = 1 \) A passes through the solution. In time \( dt \), ions of each species migrate an average distance \( v_\pm dt \) toward or away from the + electrode. Thus, all – charges originally in a layer with that thickness near the electrode arrive there and deposit negative charge; similarly, all + charges originally in that layer move away and get replaced by new + charges from the electrode. In all, net charge \( Idt \) leaves the + electrode.

a. Write a formula that connects \( v_\pm, I, \) and other quantities in the problem, and solve it for \( v_\pm \). (Don’t evaluate it numerically yet.)

b. Now imagine applying a uniform magnetic field \( \vec{B} \) perpendicular to the plane of the picture, with strength \( B = 0.03 \) T. Write a formula for the resulting magnetic force on a single ion of each species. Then convert this to a formula for the total force per unit volume. (Still don’t evaluate yet.)

c. Multiply your result for (b) by the volume of the chamber to get the total force and evaluate it. Is it big enough to drive the slow but visible motion seen in Media ??

15.4 Parity I

Background: The static Maxwell equations in traditional form (Equation 15.15, page 228) are manifestly invariant under spatial rotations, because they involve constructions that are themselves invariantly defined (dot product, cross product, curl, divergence). The equations are also invariant under spatial inversions (parity), but this is not quite so obvious, because:

- They involve the \( \vec{B} \) field, whose definition involves a choice of which hand is deemed “right”; and
- They involve cross products, which also depend on the same conventional choice.

This chapter found reformulation of some of these equations that involve no Levi-Civita tensors (Equations 15.6 and 15.10, plus the Lorentz force law Equation 15.1), by re-expressing the magnetic field \( \vec{B} \) in terms of \( \vec{\omega} \).

Do:

a. Finish the job by reformulating the magnetostatic Ampère law in terms of \( \vec{\omega} \) and simplifying with identities such as the ones in Section 14.5.2 (page 214). Hence, render the equations in a form that is manifestly invariant under inversion.

b. Similar criticisms can be raised for rigid body mechanics, which is also parity invariant, yet full of cross products:

\[
\vec{u}_\ell(\tau) = \vec{\omega} \times \vec{r}_\ell(\tau) \quad \text{velocity from angular velocity and position}
\]
\[
\vec{L} = \vec{\omega} \cdot \vec{r} \quad \text{angular momentum from angular velocity}
\]
\[
\vec{\tau} = \sum_\ell \vec{r}_\ell(\tau) \times \vec{f}_\ell(\tau) \quad \text{torque from force and position}
\]
\[
\frac{d\vec{L}}{dt} = \vec{\tau} \quad \text{Newton’s law of motion}.
\]

Construct second-rank tensors \( \vec{\Omega}, \vec{\Lambda}, \) and \( \vec{T} \) that are dual to \( \vec{\omega}, \vec{L}, \) and \( \vec{\tau} \), respectively. Re-express the preceding equations in these new quantities, and show that all the cross products are gone. Hence, render the equations in manifestly parity-invariant form.

November 14, 2023; Contents Index Notation Glossary
[**Hint:** Equation 13.1 for the moment of inertia tensor already has the desired form, so there's no need to reformulate it.]

15.5  
[[Not ready]]

15.6 **Helmholtz coils**

**Background:** Sometimes it’s desirable to have a very uniform $\vec{B}$ field, for example, to minimize net force on a molecular dipole (Chapter 17).

Two flat, circular coils each have $N$ turns of wire, each has radius $a$ and lies parallel to the $xy$ plane, and each carries current $I$ in the same direction (for example, both clockwise when viewed along the central axis). But the coils are displaced from each other, with centers at $(0, 0, \pm w/2)$.

a. Explain why $\vec{B}$ evaluated at points on the $z$ axis is always directed strictly along the $z$ axis. Explain why $d\vec{B}_z/dz = 0$ on the $z$ axis at the midpoint $0$ between the coils.

b. There is a special value $w_*$ for which the second derivative $d^2\vec{B}_z/dz^2 = 0$ is also zero at the midpoint. Find that value, then with that choice find the third derivative $d^3\vec{B}_z/dz^3 = 0$ at the midpoint. So far, everything can be done analytically.

c. Now switch to numerical evaluation: For the geometry you found in (b), use a computer to graph $||\vec{B}(x, 0, z)||$ relative to its value at the center, for $x, z$ throughout an interesting region of the $xz$ plane. Then repeat but with $w = 0.7w_*$, and comment on the qualitative difference.

15.7  
**2D and 3D magnetic field line plots**

Consider a circular loop of wire in the $xy$ plane, of radius $a$ and carrying a steady current. The magnetic field that it creates, when evaluated anywhere in the $xz$ plane, itself lies in that plane. Hence the streamline passing through any point in that plane remains confined to it.\(^{17}\)

a. Learn how to use a computer to create 2D streamplots, and use it to show a representative collection of magnetic field lines in the $xz$ plane. Although this system has a magnetic dipole moment, don’t use the dipole approximation; find $\vec{B}$ by numerical integration.

b. Repeat but this time in 3D: Instead of a cross-section, show a representative sample of the field lines throughout space for the same system. Try various viewing angles until you find one that is most informative.

[**Hint** for both parts: Replacing $\vec{B}$ by $\vec{B}/||\vec{B}||$ will not change the streamlines, though it will change the parameterization of the curves in space that you’ll find. Specifically, this transformation will ensure that your streamlines are parameterized by arc length, which may help your computer to find them more readily.]

By now you have probably noticed a more general statement than the one mentioned earlier: At any point in space, $\vec{B}$ points in a direction lying in the plane that includes that point and also the $\hat{z}$ axis, so the field line through that point stays confined to that plane.

c. Repeat (a) for a system consisting of two identical, circular current loops, both centered

\(^{17}\)Section 0.3.1 (page 7) describes streamlines.
at the origin, both carrying the same current, but one lying in the $xy$ plane and the other in the $yz$ plane. Before you begin numerical work, formulate some expectations: (i) What do you expect the field lines to look like far from the source? (ii) What do you expect the field lines to look like up close to each wire? (iii) Will there be any plane with the property that field lines starting there remain there? If so, make your 2D streamplot graph in that plane, and discuss more generally whether it has the expected properties.

d. Repeat (b) for the two-loop system.
CHAPTER 16

Units and Dimensional Analysis

The gardeners had told the Prince that you couldn’t have pigs and flowers, so he decided to have pigs.

— “Saki” 1870–1916

16.1 FRAMING: COMMUNICATION

The fundamental equations of physics are valid in any set of units. So you have a right to be puzzled when an author writes down the Maxwell equations in a different looking form from the one in Equations 0.1–0.4 (page 2), and explains the difference by saying “I’m working in gaussian units.” We sometimes need to communicate with such people, or at least read their work. This chapter will explain that “gaussian units” is really three different sets of conventions, only one of which involves the choice of base units.

The three conventions are (i) choice of base units, (ii) choice of what physical quantity we use to represent magnetic induction, and (iii) choice of whether to eliminate charge units.

Once you understand that there are three distinct points, conversions (and conversations) become more straightforward.

Electromagnetic phenomenon: Eddy currents slow the fall of a magnet through a conducting tube.
Physical idea: Although the full analysis is complex, dimensional analysis immediately tells us the general magnitude of the effect.

16.2 BASE UNITS IN MECHANICS AND THERMODYNAMICS

Just about every useful thing you’ve ever learned about units in mechanics can be systematized via a simple maxim:\(^1\)

Most physical quantities should be regarded as the product of a pure number times one or more units.\(^2\) A unit can be regarded as a symbol representing an unknown quantity, just as we use the letter \(x\) for an unknown number.

---

\(^1\) Appendix A discusses background to this chapter.

\(^2\) “Most” because a few are dimensionless. Also, one quantity (temperature) is sometimes expressed with an offset that complicates its conversions.
Again: The units are part of the quantity. We carry the unit symbols along throughout our calculations. They behave just like any other multiplicative factor; for example, a unit can cancel if it appears in the numerator and denominator of an expression. Although they are unknowns, we do know certain relations among them; for example, we know that 1 inch \(\approx 2.54\) cm. Dividing both sides of this formula by the pure number 2.54, we find 0.39 inch \(\approx 1\) cm, and so on.

We refer collectively to m, cm, inches...as different units for the same “dimension,” which we denote generically by L. Similarly time and mass have generic dimensions called T and M respectively.

Suppose that you encounter a quantity that’s incommensurable with anything that’s already got a unit. To express quantities of the new type, you first need to choose a new base unit.

Long ago it seemed obvious that one should take the most important person (the king), and use his foot as a unit of length. Later when kings were less fashionable in France, scientists chose the size of the Earth as a more universal length, and so defined the meter in terms of that. Earth isn’t easy to measure accurately, so instead a single object (a metal bar) became the standard meter. But even that object was subject to tiny changes, for example with temperature, and anyway it was inconvenient to travel to Paris to calibrate one’s apparatus. (What if we wanted to discuss physics with extraterrestrials?)

Today’s SI system instead begins with time. We could try defining an hour as 1/24 of a day, but that again seems Earth-centric (and which day?). But anyone anywhere in the Universe can examine isolated atoms of \(^{133}\text{Cs}\), so the second is now defined as the duration of exactly 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of that atom’s ground state.\(^3\) It’s as arbitrary as a king’s foot, but it’s a universal standard.

Once the second is defined, we can get back to length. For example, anyone anywhere in the Universe can define the light-second as the distance traveled by light in one second, but that’s an inconveniently large unit for laboratory work. Instead, the SI currently defines the meter by requiring that the speed of light be

\[
c = 299792458 \text{ m/s } \quad \text{exact. (16.2)}
\]

It similarly defines the kilogram in terms of the meter and second by requiring that the (unreduced) Planck constant \(h\) be exactly \(6.62607015 \cdot 10^{-34} \text{ kg m}^2 \text{ s}^{-1}\). (The kelvin is similarly defined by requiring that the Boltzmann constant have a specified exact numerical part.)

### 16.3 UNITS IN ELECTRODYNAMICS

\(^3\)That peculiar numerical value, and others appearing below, were chosen to make the SI base units as close as possible to their older definitions.
16.3.1 A new dimension

Charge is incommensurable with time, length, and mass, so we have to make an additional arbitrary choice of base unit, and also assign a new dimension symbol for it (extending $T$, $L$, and $M$).

Here are the Maxwell equations as stated in the Prologue:

\[
\begin{align*}
\nabla \cdot \vec{E} &= \rho / \varepsilon_0 \quad \text{Gauss} \\
\nabla \cdot \vec{B} &= 0 \quad \text{Gauss} \\
\n\nabla \times \vec{E} + \partial \vec{B} / \partial t &= \vec{0} \quad \text{Faraday} \\
\n\nabla \times \vec{B} - \mu_0 \varepsilon_0 \partial \vec{E} / \partial t &= \mu_0 \vec{j} \quad \text{Ampère}.
\end{align*}
\]

and the Lorentz force law:

\[
d\vec{p} / dt = q(\vec{E} + \vec{v} \times \vec{B}).
\]

Two constants of Nature, $\varepsilon_0$ (the electric permittivity of vacuum) and $\mu_0$ (magnetic permeability of vacuum), were needed in order for the dimensions to work out in Eqns. 0.1–0.5.

The dimensions of the electric and magnetic fields follow from the Lorentz force law:

\[
\vec{E} \sim \frac{M L}{T^2 Q} \quad \text{and} \quad \vec{B} \sim \frac{M}{T Q}.
\]

Here “~” means “has the same dimensions as.”

The Gauss and Ampère laws then give the dimensions of $\varepsilon_0$ and $\mu_0$:

\[
\varepsilon_0 \sim \frac{Q}{L^3 M} \sim \frac{Q^2 T^2}{L^3 M} \quad \text{and} \quad \mu_0 \sim \frac{M}{T Q L} \sim \frac{M L}{Q^2}.
\]

Because these physical constants involve charge dimensions, their numerical parts will depend on what we choose as our base unit of charge. We can use this freedom to arrange that either $\varepsilon_0$ or the proton charge $e$ has an exactly specified numerical part. Once we do that, then there’s no more freedom; the other one has numerical part set by Nature, which we can only measure and quote to a certain number of significant figures.

16.3.2 The SI base unit of charge is the coulomb

To set a universal standard unit of charge, we can imitate the preceding discussion by asking for a constant of Nature whose value we can require to have a specific, exact value.

Because protons are available everywhere in the Universe and appear to all have the same charge, the SI defines the coulomb by requiring that the proton charge is

\[
e = 1.602 176 634 \cdot 10^{-19} \text{ coul.} \quad \text{exact} \quad (16.3)
\]

---

4 The symbol “~” means “is approximately equal to.”

5 Then $\mu_0$ will follow whatever status we gave to $\varepsilon_0$, via Equations 16.2 and 16.4.

6 Again that strange multiple was chosen to make this definition nearly equivalent to an older one. See Further Reading.
The large exponent has the convenient consequence\(^7\) that typical atomic and molecular-bond energies are around \(1\text{eV} = (e)(1\text{ J coul}^{-1})\).

The choice in Equation 16.3 exhausts our freedom to set units, so any other constant of Nature must be measured in the lab, not set to an exact value by definition. Thus, the numerical values of\(\varepsilon_0\) and\(\mu_0\) do not have declared exact values. They are not independent, however; Chapter 18 will show that they are related by

\[
\varepsilon_0 = \frac{(\mu_0 \varepsilon_0)^{-1/2}}{nT} \quad (16.4)
\]

Thus, measuring one gives the other one, because \(c\) is exact. Recent values are

\[
\mu_0 \approx (4\pi)(1.000 000 000 82 \cdot 10^{-7}) \text{ m kg coul}^{-2} \quad (16.5)
\]

and hence, via Equations 16.4 and 16.2,

\[
\varepsilon_0 \approx 8.854187817 \cdot 10^{-12} \text{ coul}^2 \text{ N}^{-1} \text{ m}^{-2}.
\]

Section 16.3.2' (page 249) asks, “Why the proton?”

### 16.3.3 Derived SI units

Starting from the base units coul, m, kg, and s, various useful combinations have been given names (Table 16.1).\(^8\)

Hence \(\varepsilon_0\) can also be written as \(8.85 \cdot 10^{-12} \text{ F/m}\), and \(\mu_0 \approx 4\pi \cdot 10^{-7} \text{ H/m} = 4\pi \cdot 10^{-7} \text{ N/A}^2\).

---

\(^7\)Also, it implies a convenient magnitude for the SI unit of current (ampere): It’s approximately the current through a 100W light bulb (in the USA system of 110 V mains). Also, the total charge delivered in a lightning strike is of order 1 coul.

\(^8\)SI style guides say to use C for coulomb; this book instead uses coul rather than risk the confusion of a one-letter abbreviation. In handwriting, C could look like a variable representing capacitance, concentration, or the speed of light.
16.4 THE GAUSSIAN SYSTEM

16.4.1 First point: The gaussian base unit of charge is the statcoulomb

The gaussian system uses base units cm, g and s. But instead of defining a charge unit that gives \( e \) an exact value, this system makes the equally valid choice to instead give \( \varepsilon_0 \) an exact, and convenient, numerical part: The base unit of charge (the “statcoulomb”) is defined by requiring that \( \varepsilon_0 \) (not \( e \)) shall have a specified numerical part:

\[
\varepsilon_0 = \frac{1}{4\pi} \text{statcoul}^2 \text{s}^2 \text{g cm}^3.
\]

Therefore in this system, it’s \( e \) that has an approximate, measured value. We again determine \( \mu_0 \) by using Equation 16.4:

\[
\mu_0 = \frac{4\pi}{c^2} \frac{\text{g cm}^3}{\text{statcoul}^2 \text{s}^2}.
\]

Combining Equations 16.5, 16.4, and 16.6 yields

\[
1 \text{statcoul} \approx (0.1 \text{m/s})c^{-1} \text{coul} \approx \frac{1}{3 \cdot 10^9} \text{coul}.
\]

We then can express charge density in statcoul/cm\(^3\) and charge flux in statcoul/(cm\(^2\) s), and so on.

Another useful unit conversion involves electrostatic potential. The SI unit is \( \text{volt} = \text{J/coul} \). The corresponding gaussian unit is \( \text{statvolt} = \text{erg/statcoul} \).

**Your Turn 16A**

Find the relation between these units.

16.4.2 Second point: The gaussian system uses a modified \( B \) field

We can deal more briskly with point (ii) in Idea 16.1. “Gaussian” authors also redefine the magnetic induction, introducing a physically different quantity that we will call

\[ \vec{B} \equiv c \vec{B}. \]

Confusingly, they call this new quantity “the magnetic induction” and use the symbol \( \vec{B} \) for it! We won’t do that; we’ll just call it \( \vec{B} \) with no particular identifying phrase.

We can use \( \vec{B} \) in any system of units, and indeed, we’ll occasionally find it convenient even in SI units, because it has the same dimensions as \( \vec{E} \). Similarly, we will sometimes use a modified magnetic dipole moment\(^{11}\) defined by \( \vec{D}_M = \vec{D}_M/c \).

---

\(^9\)Maxwell and F. Jenkin had more to do with developing this system than Gauss (Maxwell & Jenkin, 1865). Actually, there are several cgs-based systems; gaussian is the most commonly used.

\(^{10}\)To see why this is convenient, note that a spatial region \( A \), of uniform electric field \( \vec{E}_A \), and no magnetic induction, will have the same energy density as a region \( B \) of uniform magnetic induction \( \vec{B}_B \), and no electric field, if \( ||\vec{E}_A|| = ||\vec{B}_B|| \).

\(^{11}\)Chapter 17 will introduce magnetic dipole moment. Section 51.3.2 will define a similarly modified density of moment.
The Maxwell equations can then be written as:

\[ \vec{\nabla} \cdot \vec{E} = \rho / \varepsilon_0 \]  
(16.8)
\[ \vec{\nabla} \cdot \vec{B} = 0 \]  
(16.9)
\[ \vec{\nabla} \times \vec{E} + \frac{1}{c} \frac{\partial \vec{B}}{\partial t} = 0 \]  
(16.10)
\[ \vec{\nabla} \times \vec{B} - \frac{1}{c} \frac{\partial \vec{E}}{\partial t} = \frac{1}{\varepsilon_0} \frac{\vec{j}}{c}, \]  
(16.11)

and the Lorentz force law says

\[ \frac{\partial}{\partial t} \vec{p} = q \left( \vec{E} + \frac{\vec{v}}{c} \times \vec{B} \right). \]  
(16.12)

These equations are still valid in any system of units; in gaussian base units, we have the numerical values \( \varepsilon_0 = \frac{1}{4\pi} \text{statcoul}^2 \cdot \text{s}^2 \) and \( c \approx 3 \cdot 10^{10} \text{cm} \cdot \text{s}^{-1}. \)

The electric and modified magnetic fields have the same dimensions, but it’s traditional to call the unit of \( \vec{B} \) the gauss, and that of \( \vec{E} \) the statvolt/cm. In fact, these (and the oersted) are all the same as \( \text{g cm}/(\text{s}^2 \text{ statcoul}). \)

**Your Turn 16B**

Use Equation 16.7 to show that the SI equivalents of these units are

1 gauss \( \approx \) \( c \cdot 10^{-4} \) T

1 statvolt/cm \( \approx \) \( 3 \cdot 10^4 \) V/m.

More precisely, a field \( \vec{B} = 1 \) T corresponds to \( \vec{B} \approx 10^4 \) gauss.

16.4.3 Third point: The gaussian system eliminates charge dimensions

The gaussian system has a third key difference from SI (point (iii) in Idea 16.1). To motivate it, let’s drop back to mechanics for a moment.

**Elimination of length dimensions**

Suppose that we chose to measure time in seconds and length in light-seconds. Then the value of \( c \) is exactly 1 light-second/s. Although its numerical part is 1, we cannot drop this constant from formulas, because it still has dimensions \( \text{L}/\text{T} \). But here is a procedure we could use to accomplish something similar.

For each physical quantity \( X \), define \( \bar{X} \) to be \( X \) divided by as many powers of \( c \) as are needed to eliminate the length dimensions. Thus, all barred quantities have dimensions that are powers of \( \bar{\text{T}} \) and \( \bar{\text{M}} \) only. In particular, \( \bar{c} \) is shorn of all dimensions; in fact it equals 1, so we really can drop it, simplifying formulas.

For example, we have force \( \bar{f} = f / c \), mass \( \bar{m} = m \), acceleration \( \bar{a} = a / c \), and energy \( \bar{\varepsilon} = \varepsilon / c^2 \). Then \( f \sim \text{kg}/\text{s}, \varepsilon \sim \text{kg}, \) and some very famous formulas become

\[ \bar{f} = \bar{m} \bar{a}; \quad \bar{\varepsilon} = \bar{m}. \]
Suppose that we were asked to find a force. After we do our calculations, we wind up with a value for \( f \). Knowing the meaning of force, we can then find the actual force (newtons) as \( f \cdot c \). We get the same final answer as usual—if we didn’t make any errors along the way.

*This book does not advocate elimination of dimensions*, because the price paid for more compact formulas is that we lose some of the precious error-checking opportunities given by dimensional analysis.

### Elimination of charge dimensions

Analogously to the preceding discussion, for each physical quantity \( X \), “gaussian” authors define \( \tilde{X} \) to be \( X \) divided by as many powers of \((\text{statcoul})(\text{s})(\text{g cm}^{-1})^{-1/2}\) as are needed to eliminate the \( Q \) dimensions. This step does not change the numerical part of \( X \) if we’ve expressed it in the base units \( \text{cm}, \text{g}, \text{s}, \text{and statcoul}, \) because this factor’s numerical part equals one in those units. Why this crazy choice? With this choice, \( \varepsilon_0 \) is shorn of all dimensions: \( \varepsilon_0 = 1/(4\pi) \).

All barred quantities then have dimensions that are powers of \( \text{L}, \text{M}, \) and \( \text{T} \) only. The vacuum Maxwell equations now take the elegant form

\[
\begin{align*}
\vec{\nabla} \cdot \vec{\tilde{E}} &= 4\pi \tilde{\rho}_q \\
\vec{\nabla} \cdot \vec{\tilde{B}} &= 0 \\
\vec{\nabla} \times \vec{\tilde{E}} + \frac{1}{c} \frac{\partial \vec{\tilde{B}}}{\partial t} &= 0 \quad \text{(gaussian system)} \quad (16.13) \\
\vec{\nabla} \times \vec{\tilde{B}} - \frac{1}{c} \frac{\partial \vec{\tilde{E}}}{\partial t} &= \frac{4\pi}{c} \tilde{j} \\
\frac{d\tilde{\rho}}{dt} &= \tilde{q}(\vec{\tilde{E}} + \frac{\vec{u}}{c} \times \vec{\tilde{B}}).
\end{align*}
\]

Coulomb’s Law now has the ultra-simple form \( \tilde{q}(\vec{r}) = \tilde{q}/||\vec{r}|| \), and so on. The price we pay is that the above equations are valid only in the gaussian system (unlike Equations 0.1–0.5, which are valid in any units). And as in the mechanical example, we sacrifice some error-checking ability.

“Gaussian” authors confuse us by omitting all the bars and checks! That explains a lot of bizarre-sounding assertions like “\( 1 \text{ F} = 9 \cdot 10^{11} \text{ cm} \),” which one sometimes hears. More precisely, this statement says that “a capacitance of \( \tilde{C} = 1 \text{ F} \) corresponds to the barred quantity \( \tilde{C} = 9 \cdot 10^{11} \text{ cm} \).”

#### 16.4.4 What is an “esu”?

We have seen that the gaussian system eliminates the dimension \( Q \), while still retaining the familiar \( \text{L}, \text{T}, \) and \( \text{M} \).

Some authors go all the way and do not state any specific units. Sometimes they just write something like “esu” for everything, which roughly means “whichever gaussian base units are appropriate for this quantity in the system I’m using.” You’re supposed to

---

12See Problem 16.4.
supply the appropriate unit using context. It works if you never make any errors, and you always communicate with people who use the same unit system as you do.

### 16.5 REMARKS

It is humbling to note that electrodynamics was only a small part of Maxwell’s short professional life (think kinetic theory of gases; mathematical theories of color vision and of feedback control, management of a large laboratory…). On top of all that, he (with F. Jenkin) invented dimensional analysis in nearly its current form!

By now, the difference between unit systems should be starting to seem like, say, the difference between French and Spanish. You need to talk like the natives, wherever you’re going, but the equations have the same physical content in any language.

The gaussian system does reduce the number of constants of Nature that appear in the Maxwell equations: Instead of \( \varepsilon_0 \) and \( \mu_0 \), only \( c \) remains in Equations 16.13. Some people find this beautiful. If you instead think that making fewer errors in your work is beautiful, then don’t eliminate dimensions.

Some say that gaussian units make the duality of the electric and magnetic field clearer. It’s true, but in a trivial way. We will get the same benefit just by expressing the Maxwell equations in terms of \( \vec{B} \) instead of \( \vec{\mathbf{B}} \) (Equations 16.8–16.11), regardless of whether we measure \( \vec{B} \) in gauss or in \( T \cdot c \). Ultimately we’ll construct a single, unified “Faraday tensor” out of the components of \( \vec{E} \) and \( \vec{B} \).

Finally, don’t try looking on Amazon for a “statvoltmeter” or an “statammeter.” Using SI units in our math keeps us connected to the real world of experiments, where people use volts and amperes.

### FURTHER READING

**Semipopular:**
Basic dimensional analysis: Mahajan, 2014.

**Intermediate:**
Historical: Maxwell & Jenkin, 1865.

**Technical:**
Current definitions of SI units: [www.bipm.org/en/publications/si-brochure](http://www.bipm.org/en/publications/si-brochure). (Prior to 2019, the SI defined the coulomb by giving \( \mu_0 \), not \( \varepsilon_0 \), an exact conventional value: The constant in Equation 16.5 was exactly 1).
16.3.2' Why base the SI on the proton?
The SI essentially measures charge as multiples of the proton charge, a seemingly arbitrary choice. Why is the proton privileged among all the many fundamental particles? Remarkably, every known, isolable, fundamental particle has charge that is an exact integer multiple of $e$. (Even quarks, which are not isolable, and quasiparticles in condensed matter have charges that are exact rational multiples of $e$.) The Standard Model of particle physics offers no necessary reason for this numerical coincidence; explaining it was one of the original motivations behind grand unification, which however has not been confirmed experimentally.

16.4.1'a Planck units
The SI sets one base unit arbitrarily (the second), then fixes the others by requiring that fundamental constants have exactly specified values. In principle, one need not stop there, because there is another fundamental constant: Newton’s gravitational constant. Rather than employ a frequency associated to cesium atoms (Section 16.2, page 241), we can therefore require that $G_N$ have an exact value. (The SI does not do this because of technical limitations on the accuracy of determining $G_N$.)

The simplest possible approach would be to require that $c$, $h$, and $G_N$ all have numerical part equal to one. The resulting system is called Planck units. Amazingly, Max Planck intuited the existence of such universal units before even coming up with his black body spectrum formula (and decades before the meaning of $h$ was understood).

16.4.1'b Other schemes
Two other unit schemes deserve mention here, each of which stops short of the Planck system but in different ways:

- In gravitational physics, many authors set one base unit arbitrarily (the meter), then agree to measure time in units of $(1 \text{ m})/c$ (not in seconds) and mass in $(1 \text{ m})c^2/G_N$ (not in kilograms). Barred quantities are obtained by dividing physical quantities by enough powers of $c$ and $G_N$ to eliminate both $\text{ m}$ and $T$, leaving only $L$. In this scheme, $G_N = 1$ and $c = 1$, so both can be dropped from formulas.
- In high-energy physics, many authors instead set a different base unit arbitrarily (the electron volt), then eliminate both $L$ and $T$ by demanding that the numerical parts of $c$ and $h$ both equal 1. There remains only the dimension $M$, which they typically measure in GeV/$c^2$.

---

15Generally authors using this system also use the remaining freedom to make the numerical part of $\varepsilon_0$ equal to $1/4\pi$ as in the gaussian system.
16.1 Dimensional shortcut

*Background:* A magnet is dropped through a vertical channel in a block of nonmagnetic, but conducting, material. Friction with the walls is negligible. But instead of increasing without bound, the magnet’s velocity saturates\(^ {16}\) at a surprisingly small value \(v_s\). At this terminal velocity, the release of gravitational potential energy does not go into increasing the magnet’s translational kinetic energy; instead, it all goes into ohmic heating of the conductor, via induced “eddy” currents.

We could try to set up and solve a lot of equations, but it would be a long road. Instead, obtain an estimate for the terminal velocity as follows. Before you begin, note that:

(i) The effect depends on the strength of the magnet, that is, on its dipole moment \(\mathcal{D}_M\). We’ll see later that the moment always appears in induction formulas multiplied by \(\mu_0\), so let \(X = \mu_0 ||\mathcal{D}_M||\). In the limit \(X \to 0\), there’s no effect and (in vacuum) the falling magnet’s velocity increases without limit, that is, \(v_s \to \infty\).

(ii) The effect depends on the electrical conductivity \(\kappa\) of the conducting material. In the limit \(\kappa \to 0\), eddy currents are suppressed, so again \(v_s \to \infty\).

(iii) The terminal velocity must also depend on the weight \(F\) of the magnet (a force). We expect that pulling harder on the magnet will let it achieve larger terminal velocity, by analogy to the case of pulling on an object immersed in a viscous fluid, that is, \(v_s\) is an increasing function of \(F\).

(iv) The effect may depend on the size scale of the apparatus, for example, on the diameter \(L\) of the channel.\(^ {17}\)

Now take these steps:

a. Find a combination of the relevant constants \(X, \kappa, F,\) and \(L\) that has the dimensions of a velocity.

b. Confirm that the formula you found in (a) has the expected trends as parameters change [(ii–iii) above].

c. Here are some typical values: A stack of button magnets like the one used in Media 8 has magnetic moment \(\approx 0.3\) A m\(^2\) and mass \(\approx 7\) g. The conductivity of aluminum is \(\kappa \approx 5 \cdot 10^7\) \(\Omega^{-1}\) m\(^{-1}\). A typical demo apparatus has diameter \(L \approx 1\) cm. Evaluate your formula for \(v_s\) numerically with these values and comment.

16.2 Units: conductivity

a. Infer the units of conductivity \(\kappa\) from the formula \(\vec{j} = \kappa \vec{E}\). Infer the units of resistance from the formula \(\Delta \phi = IR\).

b. Use dimensional analysis to guess the relation between \(\kappa\) and \(R\) for a long wire of length \(L\) and cross section \(\Sigma\). Compare to Equation 8.8 (page 119).

\(^{16}\)See Media 8.

\(^{17}\)The demonstration is often done with a copper or aluminum tube, which introduces another length scale (the wall thickness). You can neglect such complications.
c. Substitute SI base units into the dimensions of $R$ to find the definition of the SI unit of resistance (the ohm) in terms of base units.

16.3 Units: Polarizability

Explain the apparently paradoxical utterance of “gaussian” people when they say: “Electric polarizability is the ratio of the electric dipole moment of a molecule to the applied electric field. Its units are cm$^3$.”

16.4 Unit fun

Explain the paradoxical-sounding utterances of “gaussian” people, when they say:

a. “1 $\Omega = ? s/cm.$”

b. “1 $H = ? s^2/cm.$”

c. “1 farad = ? cm.”

Also fill in the missing numbers (that is, derive them).
CHAPTER 17

Magnetostatic Multipole Expansion

17.1 FRAMING: DISTILLATION AGAIN

Analogously to electrostatics, we consider a stationary, localized distribution of electric current and seek to distill just a few numbers from the distribution that characterize the far fields. Thus, \( \vec{j} = 0 \) outside a region of size \( a \), and we wish to know the fields as an expansion in powers of \( a/r \). We’ll again exploit Taylor’s theorem for \( ||\vec{r} - \vec{r}_o||^{-1} \), but there are some tricky tensor things to get right.

Electromagnetic phenomenon: Just three constants suffice to characterize the far fields of even a complicated stationary current distribution.

Physical idea: Again a multipole expansion organizes the far fields according to their falloff with distance.

17.2 TENSOR PRELIMINARIES

As in the electrostatic case (Chapter 3), first place the origin of coordinates at some fixed point inside the source. Next, recall\(^1\) that a stationary source must have \( \nabla \cdot \vec{j} = 0 \). So for \( i = 1, 2, 3 \) the divergence theorem gives

\[
0 = \int d^3r (\vec{\nabla})(\vec{\nabla} \cdot \vec{j}) = -\int d^3j \vec{j}_k (\vec{\nabla}_k \vec{r}) = -\int d^3j \vec{j}_i.
\]  

(17.1)

(The boundary term is zero because we assumed a localized source.) We conclude that each cartesian component of \( \vec{j} \), yields zero when integrated over the source.

Similarly, for every pair of indices \( k, i \) we have

\[
0 = \int d^3r (\vec{\nabla}_k \vec{r}_i)(\vec{\nabla} \cdot \vec{j}) = -\int d^3r \vec{j}_m \vec{\nabla}_m (\vec{\nabla}_k \vec{r}_i)
\]

\[
0 = \int d^3r (\delta_{mk} \vec{r}_j \vec{j}_m + \delta_{mj} \vec{r}_k \vec{j}_m) = \int d^3r (\vec{r}_j \vec{j}_k + \vec{r}_k \vec{j}_j).
\]  

(17.2)

\(^1\)Section 8.4 (page 118).
Define the **magnetic dipole moment tensor** as the collection of first moments of the components of \( \vec{j} \):

\[
\vec{I} = \int d^3 r \vec{r} \otimes \vec{j}.
\]  

(17.3)

The charge flux is a vector field, but after integrating, \( \vec{I} \) is a constant tensor. Equation 17.2 says that it is **antisymmetric**.

From now on, we will change notation from \( \vec{r} \) to \( \vec{r}_s \) to refer to the location of a point inside the source. The notation \( \vec{r} \) will now refer to the position of an observer (“field point”), as in the cartoon above.

### 17.3 Far Fields of a Steady, Localized Current Distribution

#### 17.3.1 The magnetic dipole vector potential is the leading term in a series expansion

Suppose that we wish to talk about a continuously distributed current source, maybe some interstellar plasma or the flow of ions outside a neuron.\(^2\) Section 15.5.3 showed that each component of the vector potential obeys the Poisson equation. Applying a Taylor expansion to Equation 15.18 (page 229), much as we did in electrostatics, thus gives

\[
\vec{A}(\vec{r}) = \frac{\mu_0}{4\pi r} \int d^3 r_s \vec{j}(\vec{r}_s) \left( 1 + \frac{\vec{r} \cdot \vec{r}_s}{r^2} + \cdots \right).
\]

(17.4)

In principle we’re done. But some further observations are useful.

Equation 17.1 says that the first term of Equation 17.4 is zero: There is no contribution at order \( r^{-1} \), that is, a localized, stationary current distribution never creates a “magnetic monopole” field.\(^3\)

The definition Equation 17.3 lets us rephrase the second term:

\[
\vec{A}(\vec{r}) = \vec{A}^{MD}(\vec{r}) + \cdots \text{ where } \vec{A}^{MD}(\vec{r}) = \frac{\mu_0}{4\pi r^3} \vec{r} \cdot \vec{I}.
\]

(17.5)

Similarly to the electrostatic case, we have accomplished our usual goal of expanding the potential in a systematic power series and, at the lowest nontrivial order, separating the vector potential into a product of universal, standard functions of \( \vec{r} \) (here the three functions \( \mu_0 \vec{r}_s/(4\pi r^3) \)) multiplied by some constants characterizing the source (here the components of \( \vec{I} \)).

Although \( \vec{I} \) appears to be a rank-two tensor with nine independent entries, actually we have seen that it is always antisymmetric, and hence has only three independent entries. We can make this fact more obvious by manipulating a bit to cast our result into a traditional form.

---

\(^2\)Section 8.7 (page 121) introduced this problem.

\(^3\)A magnetic monopole field is, however mathematically imaginable; see Problem 17.2.
Recall that any antisymmetric, rank-two, 3-tensor can be rewritten in terms of a vector (as we already did when we introduced $\vec{B}$ in Chapter 15). Thus, we get relations analogous to Equations 15.3 and 15.2 (page 222):

$$\vec{r}_{ln} = \varepsilon_{lnk} \vec{D}_{M,k} \quad \text{where} \quad \vec{D}_{M} = \frac{1}{2} \int d^3r_s (\vec{r}_s \times \vec{j}(\vec{r}_s)).$$

(17.6)

The three numbers $\vec{D}_{M}$ are called the components of the magnetic dipole moment vector. In terms of them, the leading term of our expansion, Equation 17.5, takes the form

$$\vec{A}^{\text{MD}}(\vec{r}) = \frac{\mu_0}{4\pi} \frac{\vec{D}_{M} \times \hat{r}}{r^2}.$$ 

(17.7)

Thus, the leading nonzero term of the vector potential far from a general local current distribution falls like $r^{-2}$, similarly to the electrostatic dipole potential in electrostatics.

17.3.2 A familiar example

**Your Turn 17A**

a. To make sure you understand how it all works, consider a thin, circular loop of wire of radius $a$ in the $xy$ plane, centered on the origin of coordinates and carrying current $I$. Work out $\vec{D}_{M}$ for this current distribution. [Hint: Use Equation 15.21 to find the charge flux and substitute into Equation 17.6.]

b. Also, compute the curl of Equation 17.7 to find the $\vec{B}$ field far away from a source, to leading nontrivial order in $a/r$. Comment on the parallel between your answer and Your Turn 3B (page 41).

17.4 HIGHER MOMENTS

17.4.1 The magnetic quadrupole fields fall faster than the dipole

As in electrostatics, there are higher magnetic multipole fields controlled by higher magnetic multipole moments. For example, consider a pair of circular wire loops, lying in parallel planes but shifted perpendicular to those planes and carrying opposite currents. The total magnetic dipole moment is zero, but there will nevertheless be magnetic fields outside this source. Those magnetic quadrupole fields fall off with distance faster than those of a magnetic dipole.

**Your Turn 17B**

Work out the next-order term in Equation 17.4 (the first term in the ellipsis).

Your answer involves a 3-tensor of rank three, the second moment of the charge flux. Dropping the stars, it’s $\int d^3r \vec{r}_i \vec{r}_j \vec{r}_k \vec{r}_m$. This tensor is clearly symmetric on its first two indices.

---

4Like $\vec{B}$ itself, the definition of $\vec{D}_{M}$ requires that we make a convention for which hand is “right” (Section 15.2, page 222); Equation 17.3 shows that $\hat{r}$ is free from this subtlety.
so we might imagine that it would have \( \frac{3(3+1)}{2} \times 3 = 18 \) independent entries. But our experience with the dipole term should make us suspicious of this conclusion: it initially looked like a generic tensor, but then proved to be always antisymmetric. Let’s look more closely at the corresponding situation at quadrupole order.

Begin by extending the argument in Section 17.2: For every \( k, i, n \) we have

\[
0 = \int d^3 r (\vec{r}_k \vec{r}_n) \vec{V} \cdot \vec{J} = -\int d^3 r \vec{j}_m \vec{V}_m (\vec{r}_k \vec{r}_n)
\]

\[
= \int d^3 r \left( \delta_{mk} \vec{r}_n \vec{j}_m + \delta_{mn} \vec{r}_k \vec{j}_m + \delta_{mn} \vec{r}_k \vec{j}_m \right) = \int d^3 r \left( \vec{r}_k \vec{r}_n \vec{j}_k + \vec{r}_k \vec{r}_n \vec{j}_k + \vec{r}_k \vec{r}_n \vec{j}_k \right). \tag{17.8}
\]

That is, the totally symmetrized part of the second moment of the charge flux is zero, extending our earlier result about the first moment (Equation 17.2).

Next, Equation 17.6 expressed the rank-two tensor \( \vec{I} \) in terms of a lower-rank object. Imitating that step with one higher rank suggests that we define the magnetic quadrupole moment tensor via

\[
\vec{Q}_M = \frac{2}{3} \int d^3 r_s (\vec{r}_s \times \vec{j}(\vec{r}_s)) \otimes \vec{r}_s. \tag{17.9}
\]

**Your Turn 17C**

Show that \( \vec{Q}_M \) is a traceless tensor.

Certainly \( \vec{Q}_M \) has fewer independent entries than the full second moment of \( \vec{j} \), but we must now see whether it really contains all the information we need to get the far fields. Extending the analogy to the dipole term further:

**Your Turn 17D**

Simplify the expression

\[
\frac{1}{2} \left( \varepsilon_{int} \vec{Q}_{M,nk} + \varepsilon_{ink} \vec{Q}_{M,nt} \right) \tag{17.10}
\]

by using identities in Chapter 14. Then use Equation 17.8 to show that the expression equals \( \int d^3 r_s \vec{j}_s \vec{r}_s \vec{r}_s \vec{r}_s \). That is, \( \vec{Q}_M \) does fully determine the second moments of the current. We can therefore substitute Equation 17.10 into your result from Your Turn 17B to find the next term in Equation 17.5:

\[
\vec{A}_j^{MQ}(\vec{r}) = \frac{\mu_0}{8\pi r^3} \left( 3\vec{r}_j \vec{r}_j - r^2 \delta_{k}\right) \varepsilon_{int} \vec{Q}_{M,nk}. \tag{17.11}
\]

Again, this result separates the far fields into a linear combination of several universal functions of \( \vec{r} \), which do not depend on any aspect of the source, weighted by a handful of expansion coefficients, the constants that make up the quadrupole moment tensor. Those
few constants completely characterize the source for the purposes of finding the far fields at this order in the expansion.

Problem 17.6 asks you to find the corresponding $\vec{B}$ field. Only the symmetric part of the magnetic quadrupole tensor enters the final answer.\(^5\) Also, because $\vec{D}_M$ is traceless, we see that just five independent numbers determine the magnetic quadrupole fields in magnetostatics, analogously to the five independent entries of the electric quadrupole moment tensor.

**Your Turn 17E**

A circular loop of wire carries current $I$ and sits in the $xy$ plane centered on the origin. Find a symmetry argument that saves us the trouble of having to compute the magnetic quadrupole moment. [*Hint: Recall a similar situation in electrostatics (Section 3.6.6, page 43).]*

---

**17.4.2 All moments after the first nonzero one are basepoint-dependent**

**Your Turn 17F**

Returning to Equation 17.6, show that, had we chosen a different origin of coordinates shifted by a constant vector $\vec{h}$, we would have ended with the same values for $\vec{D}_M$.

Similarly to the electrostatic case, higher moments may depend on the choice of basepoint; more precisely, only the first nonzero moment is unambiguously defined.\(^6\)

---

**17.5 MAGNETIC DIPOLE IN AN EXTERNAL FIELD**

We now find magnetic analogs of some results in Section 3.7 (page 44).

**17.5.1 Force and torque on a dipole**

**Force**

Consider a stationary distribution of charge flux, for example, a loop of wire fed by a constant-current source. For a more microscopic example, some individual molecules can have persistent currents in their electron state.

This distribution is immersed in an external static magnetic field $\vec{B}^{\text{ex}}$, which we suppose varies with a characteristic length scale much bigger than the size of the distribution itself. For example, that scale could be in the centimeter range for a lab apparatus, whereas the charge flux is zero outside a molecular-scale region.

---

\(^5\)In Problem 17.5 you’ll see why we may just discard any antisymmetric part of $\vec{D}_M$.

\(^6\)See Problem 17.7.
Choose an origin of coordinates somewhere inside the current distribution. Any internal forces must sum to zero, so we need only consider forces from the external field. To find them, first imagine the charge flux to consist of small packets in motion. Near some point \( \vec{r}_s \), the packets have a mean velocity \( \vec{v} \). Next, consider a small surface element \( d\Sigma \) perpendicular to \( \vec{v} \). Then in time \( dt \), all packets within a distance \( \nu dt \) of the surface cross it. The total charge that they carry is the volume \( \nu dt d\Sigma \) times the charge density \( \rho_q \). The charge flux is that net charge divided by \( d\Sigma dt \), or \( \vec{f} = \frac{\nu \rho_q}{d\Sigma} \). Indeed this expression has the units appropriate for a charge flux. Applying the Lorentz force law to the charges in a small region of space, and summing the forces, then gives\(^7\)

\[
\vec{f} = \int d^3 \vec{r}_s \vec{j}(\vec{r}_s) \times \vec{B}_{\text{ext}}(\vec{r}_s).
\]

Similarly to Section 3.7, we now make a Taylor expansion of the external field near the reference point: \( \vec{B}_{\text{ext}}(\vec{r}_s) = \vec{B}_{\text{ext}}(\vec{0}) + \cdots \). Then

\[
\vec{f}_i = \varepsilon_{ijk} \left[ \frac{\partial B_{\text{ext}}}{\partial r_m} \right]_{\vec{r}=\vec{0}} \varepsilon_{mnj} \vec{D}_{\text{M},j} + \cdots.
\]

The first term on the right equals zero by Equation 17.1. The second involves the magnetic dipole moment, which we again express as in Equation 17.6:

\[
\vec{f}_i = \varepsilon_{ijk} \frac{\partial B_{\text{ext}}}{\partial r_m} \varepsilon_{mnj} \vec{D}_{\text{M},j} \left|_{\vec{r}=\vec{0}} \right.
\]

\[
= \frac{\partial B_{\text{ext}}}{\partial r_m} \left|_{\vec{r}=\vec{0}} \right. (\delta_{in} \delta_{kj} - \delta_{ij} \delta_{mk}) \vec{D}_{\text{M},j}
\]

\[
= \vec{D}_{\text{M},j} \varepsilon_{ijk} \vec{B}_{\text{ext}} - \vec{D}_{\text{ext}} \varepsilon_{ijk} \vec{B}.
\]

Similarly to the electric dipole case, so too a magnetic dipole feels no net force in a uniform, stationary magnetic field.

If the dipole moment is fixed (unchanged by rigid translation), then we can rewrite the last expression as

\[
\vec{f} = \nabla (\vec{B}_{\text{ext}} \cdot \vec{D}_{\text{M}}).
\]

Equation 17.13 expresses the force as the gradient of a scalar function. It is tempting to interpret this result by saying that the dipole “wants” to move to minimize a “potential energy” equal to \(-\vec{B}_{\text{ext}} \cdot \vec{D}_{\text{M}}\), similarly to an analogous result with electric dipoles, but we must be careful. Unlike in the electric case (Section 3.7.1, page 44), allowing the dipole to move, while assuming a constant moment, may require the assumed constant-current source to inject (or remove) some energy.

\(^7\)Equation 35.5, page 558 will revisit this argument.
Your Turn 17G

Equation 17.13 expresses the force in terms of the product of two quantities that each change sign if we change our arbitrary convention about which hand is “right.” Remedy this defect by starting over from Equation 15.1 (page 222) and expressing the force in terms of the tensors $\mathbf{\tilde{\omega}}$ and $\mathbf{\tilde{l}}$. [Hint: The magnetic Gauss law again enters, in the form Equation 15.9 (page 225).]

Torque

We can also work out the torque on this rigid current distribution:

$$\mathbf{\tau} = \int d^3r_e \mathbf{r}_e \times (\mathbf{j} (\mathbf{r}_e) \times \mathbf{B}^{\text{ext}}).$$

For example,

$$\tau_3 = \int d^3r_e \left[ j_3 (\mathbf{r}_e) (\mathbf{r}_e \times \mathbf{B}^{\text{ext}}) - \mathbf{B}_3^{\text{ext}} (\mathbf{r}_e \times \mathbf{j} (\mathbf{r}_e)) \right].$$

Consider the case where $\mathbf{B}^{\text{ext}}$ is uniform:

$$\mathbf{B}_i^{\text{ext}} \int d^3r_e j_3 (\mathbf{r}_e) \mathbf{r}_e^* - \mathbf{B}_3^{\text{ext}} \int d^3r_e j_i (\mathbf{r}_e) \mathbf{r}_e^*.$$

The second term is zero because the magnetic dipole moment tensor is antisymmetric. The first term can be written in terms of the moment by using Equation 17.6:

$$= \mathbf{B}_i^{\text{ext}} \frac{1}{2} \varepsilon_{3ik} \int d^3r_e (\mathbf{j} \times \mathbf{r}_e)_k.$$

More generally the full torque is

$$\mathbf{\tau} = -\mathbf{B}^{\text{ext}} \times \mathbf{D}_M, \quad \text{uniform external field} \quad (17.14)$$

If the external field is nonuniform, then Equation 17.14 will have further terms involving the quadrupole and higher moments.

We conclude that a free magnetic dipole of fixed strength in an external field experiences a torque that tries to align its moment with the external field. Once it is aligned, Equation 17.13 shows that it also feels a force driving it to a region of higher $||\mathbf{B}^{\text{ext}}||$. You’ll explore a practical application of these results to manipulation of micrometer objects in Problem 17.3.

Note that a quantum-mechanical spin cannot freely “reorient,” due to spatial quantization. Thus, a single neutron, which has an intrinsic magnetic dipole moment, will migrate along or against the gradient of $||\mathbf{B}^{\text{ext}}||$ depending on its spin state: The Stern–Gerlach effect (1922). Even particles currently thought to be fundamental, like the electron and muon, have intrinsic magnetic dipole moments.

---

8 Instead of free neutrons, Stern and Gerlach used neutral silver atoms, which also have net magnetic dipole moment. Spin physics was born when they found an even number of discrete spin states, not the odd number predicted from the theory of orbital angular momentum. Section 34.11 (page 545) will discuss how intrinsic spin fits into the tensor analysis methods developed in this book.
Your Turn 17H

Comment on how Equation 17.14 behaves if we change our convention for which hand is deemed to be “right.”

17.5.2 Diamagnetism, paramagnetism, ferromagnetism

Just as atoms and molecules can “polarize” (develop an electric dipole moment) under the influence of an external electric field, so others are magnetically polarizable: They develop persistent internal currents under the influence of an external magnetic field, which in turn give rise to an induced magnetic dipole moment. Bulk materials containing such molecules can develop a density of magnetic dipole moment throughout their volume. Also analogously to the electric case, a material can polarize simply by the alignment of preexisting, but initially disordered, intrinsic dipole moments.

The induced moment can be parallel to the applied field (paramagnetism), or antiparallel to it (diamagnetism). If a material has a state that can retain a net permanent magnetic dipole moment density even in zero applied field, we call it ferromagnetic. An object made of such material, in its ferromagnetic state, is a “magnet”; the point on its surface with maximal field strength and with $\mathbf{B}$ pointing outward is its “north pole.”

Section 17.5.2’ (page 261) discusses higher-multipole polarization.

17.5.3 Purification of oxygen via magnetic forces

Most gases are diamagnetic; that is, they tend to be repelled from a magnetic field. Only a very few gases are paramagnetic and tend to be attracted into or toward a magnetic field, and of these, oxygen is the only common gas. It is a very important circumstance that the magnitude of the magnetic susceptibility of oxygen is many times greater than that of any other common gas. – https://patents.google.com/patent/US7771509B1/en

17.5.4 Magnetic levitation of macroscopic objects at room temperature

See Media 9 = https://www.youtube.com/watch?v=a8sCtLY-vZY and https://www.ru.nl/hfml/research/levitation/diamagnetic-levitation/.

FURTHER READING

General: Zangwill, 2013, chap. 11.
Force on a dipole: Goedecke et al., 1999
Diamagnetic levitation: Berry & Geim, 1997.

See Problems 17.1 and 17.3. Chapter 49 will also develop this idea.
Chapter 17  Magnetostatic Multipole Expansion

Rybak et al., 2011; Cai et al., 2007; Madaeni et al., 2011; Nakano & Shiraishi, 2004; Hajduk et al., 2013.
Magnetic tweezers: Lyubchenko, 2018, chap. 4; Lionnet et al., 2012b; Lionnet et al., 2012a.
17.5.2’ Higher multipole polarization
The main text introduced a density of magnetic dipole moment induced in matter by an applied field. Similarly to the case with electric polarization (Section 6.6’, page 92), individual molecules also have higher multipole moments, which could potentially change upon the imposition of an external field. Again a more complete analysis shows that the densities of higher induced moments enter the expressions for bound current and charge density via their higher derivatives (Jackson, 1999a, §6.6). Dimensional analysis again implies that these corrections may be neglected for imposed fields that vary spatially on length scales much longer than the molecular scale.
17.1  **Cell sorting**
Magnetic cell sorting is a way to isolate cells of one particular type. Small particles (about 50 nm diameter spheres) are bound to an antibody that attaches specifically to the cell type of interest (for example, a cancer cell). Cells are then mechanically separated by the difference in force applied to the target cells versus normal cells.

The magnetic particles are “superparamagnetic”; you may simply assume that they respond to an external magnetic field $\vec{B}$ by developing their own magnetic dipole moment $\vec{D}_M = v\vec{B} / \mu_0$, where $v$ is the volume of the particle.\(^\text{10}\)

The cells are then placed in a magnetic field gradient. What is the force if 100 of these particles are attached to a cell that is in a magnetic field of $10$ T, with gradient $10$ T/m?\(^\text{11}\)

17.2  **Magnetic monopole potential**
We found that a localized current distribution will not create any magnetic monopole field. Nevertheless, we can imagine a stationary magnetic field configuration for which $\vec{B}$ points everywhere radially outward from some point in space, much like the electric field from a point charge. We hit an interesting problem when we seek a vector potential for this field.

Let $r, \theta, \varphi$ be the usual spherical polar coordinates. In the following steps, express all vectors via their cartesian components, regarding $\vec{r}, \vec{\theta}, \vec{\varphi}$ as just three functions on space.

a. Find an expression for the gradient $\vec{\nabla} \varphi$. Find an expression for $\vec{\nabla} \theta$. Find an expression for the cross product $\vec{\nabla} \varphi \times \vec{\nabla} \theta$.

Consider the time-independent magnetic vector potential given by

$$\vec{A} = g\phi \frac{\cos \theta}{r \sin \theta}. \quad (17.15)$$

Here $g$ is an overall constant and $\phi$ is the unit vector in the azimuthal direction. Compute the curl of this vector potential as follows.

b. First re-express $\vec{A}$ as a scalar function times $\vec{\nabla} \varphi$ using your result in (a).

c. Review why $\vec{\nabla} \times (f \cdot \vec{V}) = (\vec{\nabla} f) \times \vec{V} + f \cdot \vec{\nabla} \times \vec{V}$ for any scalar function $f$ and vector field $\vec{V}$\(^\text{11}\).

d. Use (a–c) to compute the curl of $\vec{A}$ and interpret the result. Also find $\vec{\nabla} \times (\vec{\nabla} \times \vec{A})$ and explain why that result is significant.

e. Show that the two modified expressions

$$\vec{A}^{(\pm)} = g\phi \frac{\pm 1 + \cos \theta}{r \sin \theta} \quad (17.16)$$

 differ from the one in (b) only by gauge transformations, and hence describe the same magnetic field.

---

\(^{10}\) In more detail, generally $\vec{B} = \mu_0(\vec{H} + \vec{M})$ where $\vec{H} = \vec{M} / \chi_m$ and $\vec{M} = \vec{D}_M / v$ (see Chapter 49). Superparamagnetic means the susceptibility $\chi_m \gg 1$, so $\vec{D}_M = v\vec{B} / \mu_0$.

\(^{11}\) See Your Turn 0C (page 10).
Problems 263

magnetization per mass [emu/g]

magnetic field [mT]

Figure 17.1: [Experimental data.] See Problem 17.3. (a) Dots: Mass density of induced magnetic dipole moment for beads of diameter 1 μm and mass about 10^{-12} g. The authors stated that “1 emu = 10^{-3} SI.” Curve: Fit to the function \( M = a(\coth(B/B_0) - B_0/B) \) with \( B_0 = 20 \text{ mT} \) and \( a = 22 \text{ emu/g} \). (b) Nearly exponential dependence of magnetic field strength with distance. [Data courtesy Timothée Lionnet and Vincent Croquette.]

f. Not surprisingly, the expression Equation 17.15 is singular at \( r = 0 \). It’s also bad all along the polar axis! Show, however, that one of the vector potentials in (e) is nonsingular all along the half-line \( \theta = 0 \), whereas the other one is nonsingular all along \( \theta = \pi \). Thus, those formulas offer us vector potentials whose nonsingular domains jointly cover all of space except the origin (where there is a genuine singularity), and whose \( \vec{B} \) fields agree on the overlaps of their domains (most of space).

17.3 Magnetic tweezers

Figure 17.1 shows some information about a magnetic tweezer setup. The first graph gives the magnetic moment per gram of their bead, as a function of applied magnetic field. The second graph shows the measured magnetic field as a function of the vertical distance \( z \) from the magnet pole.

a. Apparently emu is some gaussian unit for magnetic dipole moment. Figure out the appropriate unit and explain the notation “1 emu = 10^{-3} SI.”

b. Look at the central part of the first graph, where it’s approximately linear, and estimate the slope. Use this linear approximation from now on.

c. Look at the part of the second graph for \( z \) between 2 and 4 mm. Approximate the curve in this semilog graph as a straight line. That is, set \( B \approx B_{\text{max}} \exp(-z/z_0) \) and find the constants \( B_{\text{max}} \) and \( z_0 \).

d. Now derive an approximate formula for the force on the bead as a function of \( z \), using Equation 17.12 and your results from (a–b). Sketch the expected force-versus-\( z \) curve for \( 0 < z < 6 \text{ mm} \).

e. For comparison, estimate the weight of this bead in air and comment. (It will effectively be a bit reduced in water due to buoyancy.)

12The dipole moment is not constant, so don’t use Equation 17.13.
17.4 Levitation of single cells
[[Not ready]]

17.5 Symmetry of magnetic quadrupole moment
Our formula for $\bar{Q}_m$ yields a tensor that is not necessarily symmetric. But its antisymmetric part, if any, may as usual be expressed as $\varepsilon_{imn} \bar{V}_m$ for some vector $\bar{V}$. Substitute into Equation 17.11 (page 255), then show that such a term’s contribution to the vector potential may be eliminated by a gauge transformation.

17.6 Magnetic quadrupole
Derive Equation 17.11. Then work out the curl to find the corresponding $\bar{B}$ field to quadrupole order (that is, $r^{-4}$), and confirm the claim in the chapter that only the symmetric part of $\bar{Q}_m$ enters your expression.

17.7 Basepoint dependence
Suppose that we have evaluated the magnetic dipole and quadrupole moments of a particular current distribution. Now we rigidly shift the distribution by a displacement $\Delta$. The main text showed that the dipole moment is unchanged. What happens to the quadrupole moment?

17.8 Static toroidal moment?
Evaluate the symmetric part of the magnetic quadrupole tensor of the steady current distribution shown in Figure 17.2. [Hint: First consider a ring of charge centered on the origin. Now displace that ring perpendicular to its magnetic dipole moment and use Problem 17.7. Sum up a ring of many such current rings.]

17.9 Planar loop
a. For current confined to a thin wire, the magnetic dipole moment becomes a contour integral over a curve in space. If a segment of that wire is a straight line, $\tilde{c}(u) = \tilde{w} + u\tilde{v}$, show that we just need to integrate $I\tilde{w} \times \tilde{v} du$.

b. Suppose that current $I$ is confined to a thin wire in the form of an equilateral triangle with edge length $2a$ in the $xy$ plane, and find the magnetic dipole moment. [Hint: If
you place one vertex on the origin, then only one leg of the triangle will contribute to your answer.]

17.10 Cube loop
a. Suppose that one segment of a wire loop is straight, so that \( \vec{\ell}(s) = \vec{r}_0 + s\vec{n} \) for some constant vector \( \vec{r}_0 \) and constant unit vector \( \vec{n} \). Derive an expression for the contribution to magnetic moment (Equation 17.6, page 254) from this segment. What’s special if the segment, or an extension of it, passes through the origin of coordinates?
b. The closed loop of thin wire shown in Figure 17.3 carries current \( I \). Each segment of the loop follows an edge of a cube of length \( a \). Use (a) to find the magnetic dipole moment vector of this arrangement. Why is your answer qualitatively reasonable? [Hint: It may be helpful to translate the cube so that one corner is at the origin of coordinates.]
It is impossible to study this wonderful theory without feeling as if the mathematical equations had an independent life and an intelligence of their own, as if they were wiser than ourselves, indeed wiser than their discoverer, as if they gave forth more than he had put into them. And this is not altogether impossible; it may happen when the equations prove to be more correct than their discoverer could have known with certainty. It is true that such comprehensive and accurate equations only reveal themselves to those who with keen insight pick out every indication of the truth that is only faintly visible in nature. — Heinrich Hertz, 1896
Beyond Statics

The skilled angler does not drop the fly at random on the surface of a stream. Mind penetrates the troubled water, it sees the fish below and knows the habits of its kind. Faraday sometimes compares himself with a fisherman.

— Robert Park

18.1 FRAMING: SELF-CONSISTENCY

The equations of static electricity and magnetism have a lot of practical implications. We can understand nerve impulses, photocopiers, lightning rods, molecular recognition, and much more with static, stationary, or quasistationary fields. But charges and currents are not always static, even approximately. Finding the right equations requires both experiments and the polestar of mathematical self-consistency.

Electromagnetic phenomenon: Light can be created in helicity states, that is, states with electric field vector that rotates instead of oscillating as the wave advances.

Physical idea: Linearity of the Maxwell equations implies that each polarization component may be independently phase shifted relative to the other one.

18.2 REVIEW

18.2.1 Field equations

We have explored some equations whose solutions seem to describe the electric and magnetic fields set up by stationary charges and currents:

\[ \nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0} \quad \text{Gauss} \quad (18.1) \]

\[ \nabla \cdot \mathbf{B} = 0 \quad \text{Gauss} \quad (18.2) \]

\[ \nabla \times \mathbf{E} = \mathbf{0} \quad \text{(stationary case)} \quad (18.3) \]

\[ \nabla \times \mathbf{B} = \mu_0 \mathbf{j} \quad \text{Ampère (stationary case)} \quad (18.4) \]

18.2.2 A coil carrying constant current

To anchor all the abstractions that are to come, here is an old result from magnetostatics that you probably recall from first-year physics. Figure 18.1 represents a coil of wire wound in a helix of radius \( a \) around a long cylinder of length \( w \). Such a coil is often called a
Figure 18.1: **Solenoid.** The dashed rectangle, with boundary orientation shown, can be decomposed into elements \( d\vec{S} \) all pointing into the page, which appear in the integral in Equation 18.5. The wire repeatedly pierces this surface with current always passing into the page if \( I > 0 \). The text uses the convention that \( B > 0 \) refers to \( \vec{B} \) pointing leftward.

**solenoid.** It consists of \( N \) loops. Steady current \( I \) is sent through the wire. Work through the next paragraphs to exercise those Stokes-theorem muscles.

Each loop contributes to the magnetic field. The helicity of the coil shown is such that if \( I > 0 \), then in front of the page, current is moving upward; behind the page, current is moving downward. Deep inside the cylinder (far from its ends), symmetry suggests that \( \vec{B} \) will point axially as shown, though we still need to confirm the direction. To find its magnitude, consider the path shown as a dashed line. We can traverse that path in either direction; a specific choice is shown. That choice determines a vector perpendicular to the rectangular surface bounded by that path via the right-hand rule: For the arrangement shown, \( \vec{\Sigma} \) points into the page.

Integrating Ampère’s law over the surface bounded by the path gives

\[
\int d^2\vec{\Sigma} \cdot (\vec{\nabla} \times \vec{B}) = \oint d\vec{l} \cdot \vec{B} = (\mu_0) \cdot j.
\] (18.5)

Stokes’s theorem gives the left side as \( \oint d\vec{l} \cdot \vec{B} \) where the line integral is over the closed dashed path in the figure. The part of the path lying inside the cylinder contributes \( Bw \), because \( \vec{B} \) is uniform along the coil and points axially (\( B \) is its component in the leftward direction). The short sides of the rectangular path are perpendicular to \( \vec{B} \), so here \( d\vec{l} \cdot \vec{B} = 0 \). And \( \vec{B} \approx 0 \) outside the cylinder, because the field lines fan out once they exit the ends.

Show that Equation 18.5 becomes

\[
Bw = \mu_0 NI, \quad \text{or} \quad B = \frac{\mu_0 NI}{w},
\] (18.6)

a familiar result.

**Solution:** Substitute Equation 15.19 (page 230) into the right side of Equation 18.5. Each point where the wire pierces the surface contributes \( \mu_0 I \) to the integral, because \( \vec{\Sigma} \) and \( j \) both point into the page at each such point.

Inside the solenoid, \( \vec{B} \) is uniform; it does not depend on how far we are from the coil’s axis (nor on the coil radius \( a \)). If \( I > 0 \) then \( B > 0 \), which in our convention means that \( \vec{B} \) points to the left.

---

1 Your Turn 0B (page 10).
2 We are neglecting end effects.
**Your Turn 18A**

Repeat the argument, but traverse the dashed path in the opposite direction; make sure the physical result doesn’t change.

As an aside, Equation 18.6 is sometimes expressed in terms of the quantity\(^3 \Phi_B = N \pi a^2 B\) as

\[
\Phi_B = LI, \tag{18.7}
\]

where we packaged all the constants into a single quantity to describe the coil geometry: the **self-inductance** \(L\). For the solenoid that we are considering, \(L = \pi a^2 \mu_0 (N/\omega)^2(\omega)\). That expression emphasizes that the self-inductance is an **extensive quantity**: If we double the length of the coil, holding fixed its radius and density of loops, then \(L\) doubles.

### 18.3 SLOWLY TIME-VARYING CURRENTS

#### 18.3.1 Faraday observed an \( \vec{B} \) field associated to a time-varying magnetic field

In electrostatics, Equation 18.3 says that the electric field gives rise to a **conservative** force on charges, similarly to the newtonian gravitational force. Before Michael Faraday, many scientists assumed that this situation would persist in non-static situations. After all, the newtonian gravitational equations retain the same form even for time-dependent situations, for example, even with all those planets whizzing around.\(^4\) Perhaps that prejudice was what prevented the Continental scientists from seeing what Faraday saw.\(^5\)

In the gravitational case, a roller-coaster that traverses a closed loop returns to its starting point with the same kinetic energy as it began (minus frictional losses), because the gravitational force on it is the gradient of a potential energy function. Faraday observed, however, that plunging a magnet into a loop of wire generates an effect that pushes on the electrons all around the loop. Rather than suppose that this effect is something entirely new, we will regard it as a contribution to the electric field that is **not** conservative: This contribution does not obey \( \vec{\nabla} \times \vec{E} = 0 \). Faraday found that its effects were proportional to the time rate of change of the magnetic field, which suggests the following modification to Equation 18.3:

\[
\vec{\nabla}_i \vec{E}_j - \vec{\nabla}_j \vec{E}_i = -2 \frac{\partial}{\partial t} \bar{\omega}_{ij}. \quad \text{Faraday} \tag{18.8}
\]

The left side is an antisymmetric tensor field, which matches the object \( \bar{\omega} \) that naturally

---

\(^3\)Many authors use the phrase “magnetic flux” for this quantity. That traditional terminology violates our convention that a flux is the rate of transport of some conserved quantity (such as charge) per unit transverse area, so we will not give \( \Phi_B \) any particular name.

\(^4\)Today we know that the newtonian equations also require modification; in Einstein’s theory those moving planets actually do generate tiny “gravitomagnetic” effects.

\(^5\)Chapter 36 will outline more of Faraday’s thinking. In the USA, Joseph Henry independently discovered Faraday’s law, but did not publish it promptly.
Figure 18.2: Solenoid II. The green object is a disk-shaped surface of radius \(a\), viewed edge-on. If we make the choice shown for its boundary orientation, then it can be decomposed into elements \(d^2S\) all pointing to the left, integrated over in Equation 18.10.

Describes magnetism.\(^6\) Equation 18.8 equates tensors of the same type, and it contains no Levi-Civita tensor nor any “axial vector” quantities. Also, the units match on each side.

It’s more conventional, however, to contract both sides of the preceding formula with a Levi-Civita tensor, which yields\(^7\)

\[
\vec{\nabla} \times \vec{E} = -\frac{\partial}{\partial t} \vec{B}. \quad \text{Faraday}
\]

**Your Turn 18B**

Suppose that we have a circular loop of wire. Integrate both sides of Equation 18.9 over a surface bounded by the loop, and show that the current induced by a time-dependent applied \(\vec{B}\) field flows in the direction that generates an opposing \(\vec{B}\) (Lenz’s law). Does the result depend on which perpendicular you chose for the area integration?

18.3.2 Work must be done to increase current through a solenoid

We now return to the concrete situation considered Section 18.2.2, but this time suppose that we force a current \(I(t)\) through the coil that varies slowly in time. Here “slowly” means too slowly for us to need to account for the time-derivative term in Ampère’s law (which is multiplied by a very small constant). Faraday’s law says that an electric field will result. To find it, we integrate both sides of Faraday’s law over a surface, though not the same surface as in Figure 18.1. Instead, our surface will be a disk transverse to the axis, bounded by the cylinder on which we wrapped the wire (Figure 18.2). Again we can choose either direction for the rim of that disk; to keep things simple, in the figure we chose the same direction as that of current flow. So

\[
\int d^2S \cdot (\vec{\nabla} \times \vec{E}) = -\frac{d}{dt} \int d^2S \cdot \vec{B}. \quad \text{(18.10)}
\]

This time, Stokes’s theorem gives the left side as \(\oint d\vec{r} \cdot \vec{E}\). By axial symmetry, the integrand is constant, so we get \(2\pi a E_{||}\), where \(E_{||}\) is the component in the direction of current flow.

---

\(^6\)See Equation 15.3 (page 222). Chapters 32–34 will argue that the factor of 2 is ultimately dictated by Lorentz invariance. The minus sign comes from Faraday’s experimental observations.

\(^7\)Recall Equation 15.2 (page 222).
The right side of Equation 18.10 involves the perpendicular vector to our surface that points leftward. Equation 18.6 gives the magnitude of $\vec{B}$. Thus, the right side of Equation 18.10 is

$$-\pi a^2 \frac{dB}{dt} = -\frac{\pi a^2}{w} \mu_0 N \frac{dI}{dt}. $$

Setting this expression equal to the left side of Equation 18.10 gives

$$E_\parallel = -\frac{\mu_0 Na}{2w} \frac{dI}{dt}. $$

(18.11)

The minus sign says that the induced electric field opposes changes in current.$^8$ Thus, to increase $I$ we must do work against an opposing electric field. Let’s see how much work is needed.

Our solenoid consists of a wire that contains mobile charge carriers with some linear charge density $\rho^{[1D]}_q$. Imagine the charge as being subdivided into packets $\Delta q$. Thus, there are $(2\pi a N \rho^{[1D]}_q)/\Delta q$ such packets in the wire. Each feels the same electric field, for a total force of

$$f_\parallel = \frac{2\pi a N \rho^{[1D]}_q}{\Delta q} (E_\parallel \Delta q). $$

To understand this formula, think of a pipe full of water, acted on by a body force like gravity. Pushing a volume $\delta V$ of water into the bottom of the pipe requires that we push every volume element upward against gravity, with an energy cost proportional to the weight of all the water inside the pipe, and hence to the pipe’s length. Similarly, here too every charge carrier is pushed on by the tangential electric field.

If the current is increasing in time, the minus sign in Equation 18.11 says that the induced electric field opposes that change. To overcome the electric force, some external agency must actively push charge $q$ into the solenoid with an equal and opposite force. The work required to do this is force times the distance $\Delta x = \Delta q / \rho^{[1D]}_q$. We can write the work per unit charge in terms of the self-inductance (Equation 18.7) as$^9$

$$\frac{\Delta W}{\Delta q} = -f_\parallel \Delta x = \frac{\mu_0 N^2 a^2 \pi}{w} \frac{dI}{dt} = L \frac{dI}{dt}. $$

(18.12)

This work does not arise from changing any true potential energy, as we see from the fact that Equation 18.12 contains a time derivative.

Although there is no true electrostatic potential in problems like this one, because the electric force is not conservative, nevertheless in electrical circuit theory we may treat Equation 18.12 the same way we treat a true potential drop (for example, the one across a capacitor): The total net work needed to push charge around a circuit must equal that supplied by a battery or other external source; otherwise, charge won’t flow in that direction. Luckily Equation 18.12 is still linear in the current, so the analysis of circuits with inductors is mathematically just as straightforward as that involving resistors and capacitors.

$^8$Lenz’s law again.

$^9$The self-inductance was defined in Equation 18.7. Some books use the abbreviation “back-EMF” to describe this quantity, but this book doesn’t use that term. The “F” in that traditional term stands for “force,” but force is a vector with units N; in contrast, the quantity under discussion is a scalar with units V.
18.3.3 Self-inductance also affects signal propagation along a cable

Chapter 11 introduced a model for the propagation of an electrical disturbance along a cable. As cables grew to transatlantic length, it became clear that a sharp step function introduced at one end emerged at the other end not only weakened but also blurred, limiting the speed of transmission. The problem was not just resistive loss. As mentioned earlier, Thomson made a big advance by introducing the capacitance of an undersea cable into his mathematical model. However, with the transition from Morse code to audio signals, the bandwidth requirement grew and the inadequacies of even Thomson’s model became evident. Eventually Heaviside and others realized that the problem was the neglect of self-inductance in Thomson’s model. Incorporating self-inductance creates the possibility of true traveling wave solutions, but those solutions again suffer from dispersion (Problems 18.4 and 18.5).

Remarkably, Heaviside discovered that dispersion could be eliminated, even if the leakage and capacitance of a cable were nonzero. Problems 18.5–18.9 have the details.

18.3.4 Magnetic field energy is proportional to volume

To push charge through our solenoid at rate \( I \), an external agency must therefore do work at rate given by \( I \) times Equation 18.12:

\[
\mathcal{P} = LI \frac{dI}{dt} = \frac{L}{2} \frac{d}{dt} (I^2),
\]

where again \( L \) is the self-inductance introduced in Equation 18.7. To find the total energy cost to bring current up from zero to \( I \), integrate the above formula over time. We can do that by just dropping the time derivative:

\[
\mathcal{E} = \frac{\mu_0 \pi a^2 N^2 I^2}{2} = \frac{1}{2} LI^2. \tag{18.13}
\]

It’s important not to confuse the work \( \mathcal{E} \) with “frictional” loss (ohmic heating due to resistance). Resistive losses occur even when current is held steady, and only in ohmic media (not superconductors). In contrast, the energy cost \( \mathcal{E} \) just computed applies even to superconductors, but only when current is changing. Moreover, the magnetic energy that we invest in increasing the current through the solenoid can be recovered if later we let the current decrease—the induced electric field also opposes that change, and can even be used to extract the same amount of useful work that we expended when we set up \( I \), analogously to a flywheel in mechanics. In short, a superconducting coil stores energy; it doesn’t dissipate it.

Using Equation 18.6 to re-express our answer in terms of the magnetic field yields a suggestive result:

\[
\mathcal{E} = \frac{\mu_0 \pi a^2 w}{2 \mu_0} \| \vec{B} \|^2. \tag{18.14}
\]

---

10. Heaviside also introduced the terms “inductance,” “impedance,” and “attenuation.”

11. Not so astonishingly, given his personality, Heaviside neglected to patent his very practical discovery, so others made a fortune from it.
Although we derived this formula for a specific situation, it seems to have forgotten everything about the original geometry except the volume of the region with significant fields. It suggests that there’s energy inside the cylinder, with volume density equal to $\frac{1}{2}||\mathbf{B}||^2/\mu_0$.

Equation 18.14 is reminiscent of a result from Chapter 6: We found that the energy needed to charge up a capacitor is $\frac{1}{2}C_0||\mathbf{E}||^2$ times the volume of the region with nonzero electric field. That result suggested that there’s stored energy between the capacitor plates, again with volume density given by a constant times field strength squared. That is:

The equations of electrodynamics appear to be compatible with energy conservation if we attribute energy density to empty space.

(18.15)

How are these energies stored? A more precise version of this question is, “Can we prove a general statement of the conservation of energy, in which electric and magnetic fields themselves can carry it?” We’ll pursue this inquiry in Chapter 35. Right now, we have just circumstantial evidence in special cases (a parallel-plate capacitor and a solenoid).

In Maxwell’s time, the answer seemed obvious. Paraphrasing what many believed:

“So-called vacuum, which you get by removing all the air from a vessel, is still full of stuff, the ‘ether.’ An electric field stretches that stuff, storing elastic energy. A magnetic field sets it in motion, storing kinetic energy.”

We’ll soon see that after Einstein, eventually nobody believed that proposition. Then the question got more urgent: What, then, carries the energy? We’ll return to that story after we understand Einstein.

18.4 MAXWELL’S MODIFICATION TO AMPÈRE’S LAW

We modified the static equation $\nabla \times \mathbf{E} = 0$ in order to accommodate experimental reality (Faraday’s induction). Next, we’ll see that we must also modify Ampère’s law, but for a different reason.

18.4.1 Mathematical consistency hinges on the continuity equation for charge

Hanging Question #D (page 14) raised the issue that we must solve eight Maxwell equations with just six fields $\mathbf{E}$ and $\mathbf{B}$. In statics, Section 15.5.5 (page 230) argued that the field equations are secretly just six independent equations, by taking the divergences of the two curl equations.

Moving beyond statics, we now take the divergence of both sides of the Faraday law, Equation 18.9, and find that it’s still automatically satisfied (it doesn’t constrain the fields). But taking the divergence of Equation 18.4 (page 267) and using the continuity equation

---

12See Equation 6.2 (page 78).
13This statement parallels Idea 6.3 (page 78).
14Like many overturned ideas, however, this one had a long half-life. Lenard, Lorentz, and Michelson reportedly never gave up on it.
now gives

\[ 0 = \mu_0 \nabla \cdot \vec{j} = -\mu_0 \frac{\partial}{\partial t} \rho_q. \] (18.16)

That’s just false in nonstatic situations, so we have a problem.

To make progress, notice that the Gauss law implies that the bad right-hand side of Equation 18.16 equals \(-\mu_0 \varepsilon_0 \nabla \cdot \frac{\partial \vec{E}}{\partial t}\), that is, a divergence. Next, consider a modified form of Equation 18.4 (page 267):

\[ \nabla \times \vec{B} = \mu_0 \varepsilon_0 \frac{\partial \vec{E}}{\partial t} + \mu_0 \vec{j}. \] (18.17)

This time, taking the divergence of both sides and using Equation 18.16 leads to the identity \(0 = 0\), not a contradiction.

We have arrived at Maxwell’s proposed generalization of the Ampère law to nonstationary situations, albeit not by following Maxwell’s original train of thought,\(^{16}\) and we’re all done tinkering with the equations of electrodynamics. *Equations 18.1–18.2, 18.9, and 18.17 are the equations of classical electrodynamics as they are understood today.*\(^{17}\) Later, we’ll find a clearer re-expression of those same equations, but we won’t modify their content. Later still, we’ll build a useful alternate version of these equations to describe electromagnetism in media without having to handle every electron explicitly. That version is an approximation to the equations written here, which are more fundamental and universal.

\(^{15}\)See Section 18.4.1b (page 287).

\(^{16}\)And as they appear in the Prologue.

Section 18.4.4’ (page 287) reconciles the equations as presented here with some older ideas, and meditates on Where Theories Come From.

### 18.4.2 Boundary conditions

We can now revisit some conclusions we got in electro- and magnetostatics, concerning fields at interfaces. The results that rested on integrating Gauss laws are unmodified in dynamics, because the Gauss laws themselves are unmodified:

\[ \Delta \vec{B}_\perp = 0. \] always

\[ \mathbf{n} \cdot (\vec{E}_\text{vac} - \vec{E}_\text{(d)}) = \frac{-\mathbf{n} \cdot \vec{F}_\text{(d)}}{\varepsilon_0}, \text{ dielectric/vacuum} \] with a similar formula for a dielectric/dielectric interface.

Turning now to the results that rested on integrating the Faraday and Ampère laws, we find that they, too are unchanged! That’s because the time derivative terms are to be multiplied over an area that goes to zero in the limit of a narrow rectangle in Figure 6.6b or Figure 15.2b. Thus,

\[ \Delta \vec{E}_\parallel = 0 \] and

\[ \Delta \vec{E}_\perp = \vec{0} \] always

\(^{15}\)Note that the left hand side of Equation 18.17 can be expressed without any Levi-Civita tensors, if we use the antisymmetric tensor representation of the magnetic field. And the right side certainly doesn’t have them, so the whole thing is invariant under spatial inversions.

\(^{16}\)See Section 18.4.1'b (page 287).
\[ \Delta \vec{B} = \mu_0 \vec{j}^{[2D]} \times \hat{n}, \quad [15.23, \text{page 233}] \]

where \( \vec{j}^{[2D]} \) is the net 2D charge flux at the surface.

Sometimes it is reasonable to approximate a conductor as perfectly conducting. Then there can be no electric field inside it, and the boundary condition becomes \( \vec{E} = \Delta \vec{E} = \vec{0} \). Moreover, Faraday’s law then says that \( \vec{\partial} \vec{B} / \partial t = \vec{0} \) inside. Supposing that the interior magnetic field is zero at some initial time then gives that it is always zero, so \( \vec{B}_\perp = \Delta \vec{B}_\perp = 0 \).

### 18.5 WAVE SOLUTIONS

#### 18.5.1 About traveling plane waves

In one spatial dimension, we call a function of the form

\[ \phi(t, r) = f(r - uv) \]

a **traveling wave**. Figure 11.2b shows a representation of a function of this sort as a surface. If we take a snapshot at one particular time \( t \), the result is a function of \( x \). Now take another snapshot at \( t + \Delta t \). The two snapshots are related, because \( x - uv = (x + \nabla \Delta t) - u(t + \Delta t) \) for any \( x \) and \( t \). Hence, the second snapshot is the same function of \( x \) as the first, just shifted in space by \( \Delta x = uv \). (You can visualize slicing the surface along two lines of constant \( t \); the result in each case is a bump function, just shifted in space.)

Equivalently, we could stand in one place and record the time series as the wave passes (heavy line in Figure 11.2b). If another observer stands at a different place \( x + \Delta x \), she’ll observe the same time series, just shifted in time by \( \Delta t = (\Delta x) / v \).

In two or more dimensions, we can upgrade these considerations: Any function of the form \( f(\vec{k} \cdot \vec{r} - vt) \) has the properties discussed above, where \( \vec{k} \) is any unit vector. Such a function is called a traveling **plane wave**, because there is a stack of planes (each perpendicular to \( \vec{k} \)), on each of which it is constant. Suppose that we take snapshots at \( t \) and \( t + \Delta t \). The snapshots will differ by a shift of \( \Delta \vec{r} = u(\Delta t) \vec{k} \); that is, the wave moves at speed \( u \).

We will often specialize to periodic functions, for example, taking \( f(u) = \cos(2\pi \omega u / v) \). Then our function becomes

\[ \phi(t, \vec{r}) = \cos(2\pi(\vec{k} \cdot \vec{r} - \omega t)). \]

Here \( \omega \) is any constant and \( \vec{k} = \vec{k} \omega / v \). The temporal period of this function is that it repeats when time advances by \( 2\pi / \omega \). The spatial period is that it repeats when we move along \( \vec{k} \) a distance \( 2\pi v / \omega \).

As in the mechanics of a rotating rigid body, we will call \( \omega \) the **angular frequency** (dimensions \( \text{T}^{-1}, \text{SI unit rad/s} \)) and \( k \) the **wavenumber** (dimensions \( \text{L}^{-1}, \text{SI unit rad/m} \)). Note that “radian” is a dimensionless unit of angle (because it equals circumference divided by radius), and many authors omit it when stating numerical values of \( \omega \) and \( k \). But that risks confusion with the related quantities:
Chapter 18: Beyond Statics

- **circular frequency** \( \nu = \omega / (2\pi) \) (dimensions \( T^{-1} \), SI unit \( s^{-1} \), also called Hz);
- **spectroscopic wavenumber** \( k / (2\pi) \) (dimensions \( L^{-1} \), SI unit \( m^{-1} \)). Some books call our \( k \) the “angular wavenumber” to avoid confusion with this quantity.

Additional descriptors include:

- **period** \( T = 1 / \nu \) (dimensions \( T \), SI unit \( s \));
- **wavelength** \( \lambda = 2\pi / k \) (dimensions \( L \), SI unit \( m \)).

The period is how long you have to wait at a fixed position for the wave to repeat. The wavelength is how far you have to travel at a fixed instant of time for the wave to repeat.

For a vector quantity, such as the \( \vec{E} \) or \( \vec{B} \) field, the corresponding construction is a constant vector times the same \( f \) as before.

### 18.5.2 The final form of the vacuum Maxwell equations have plane wave solutions

Section 18.4 suggested that Maxwell’s modification to Ampère’s law might not be quantitatively important in experiments. But let’s keep an open mind, and look for solutions to the modified equations. They look a bit complex—lots of equations in lots of unknowns. Let’s try to eliminate \( \vec{B} \), arriving at a smaller set of equations just involving \( \vec{E} \). Also, we will temporarily simplify by looking at empty space, a region with no charges nor currents. Already we know lots of static solutions applicable to that situation.

To do the elimination, consider taking the curl of both sides of the curl equations. In vacuum, the Faraday law gives

\[
\vec{\nabla} \times (\vec{\nabla} \times \vec{E}) = -\frac{\partial}{\partial t} \vec{\nabla} \times \vec{B}
\]

or (by using the electric Gauss law)\(^{18}\)

\[
-\nabla^2 \vec{E} = -\mu_0 \varepsilon_0 \frac{\partial^2}{\partial t^2} \vec{E}.
\]

(18.19)

Maxwell noticed that this is an example of a wave equation. Consider the trial solution of plane-wave form:

\[
\vec{E} (t, \vec{r}) = \vec{E} \cos(kz - \omega t),
\]

(18.20)

where \( \vec{\theta} \) is any real, constant vector, \( k \) is a real constant, and \( \omega \) is a real positive constant. The wave moves at speed \( \omega / k \).

Substituting Equation 18.20 into Equations 18.19 and 18.1 (page 267) gives the conditions for the trial solution to work:

\[
k^2 = \mu_0 \varepsilon_0 \omega^2 \quad \text{and} \quad \hat{z} \cdot \vec{E} = 0.
\]

(18.21)

\(^{18}\)The derivation of this formula depends on our default choice of cartesian coordinates. The left-hand side would look more complicated in curvilinear coordinates.
Your Turn 18C

a. Confirm that Equations 18.20–18.21 really do yield a solution to all of the Maxwell equations, not just the one combination Equation 18.19. You’ll need to find the appropriate $\vec{B}(t, \vec{r})$ first.
b. Try generalizing Equation 18.20 to arbitrary waveforms, that is, trial solution

$$\vec{E}(t, \vec{r}) = \vec{E}_0 f(kz - \omega t), \quad \vec{B}(t, \vec{r}) = \vec{B}_0 g(kz - \omega t).$$

What conditions, if any, must the functions $f$ and $g$ meet to yield a solution?

The last result you just found is perhaps not new: Equations 18.20–18.21 show that cosine plane waves all travel at the same speed $(\varepsilon_0 \mu_0)^{-1/2}$, independent of the amplitude $||\vec{E}||$ or frequency. So if we decompose any waveform into Fourier components, after time $t$ they will reassemble into the original waveform, just shifted in space.

Moreover, the wave speed is also independent of the polarization (direction of $\vec{E}$), or the direction of travel (sign of $k$). It’s a constant of Nature, which we’ll call $c$. Substituting the known measured values of $\mu_0$ and $\varepsilon_0$ shows that Maxwell’s modification leads to wave solutions that travel at about three hundred million meters per second. That rang a bell for Maxwell.

18.6 EM WAVES RESEMBLE LIGHT

We have anchored each ingredient in the Maxwell equations, including the sign of each term, by using an observable Electromagnetic Phenomenon. The only exception is Maxwell’s new term, but its form was dictated by the need to salvage mathematical consistency. And now the equations have yielded a testable prediction: Solutions that resemble the behavior of light.

Despite many clues that light was connected to electricity and magnetism, it still took considerable courage for Maxwell to propose that light is itself an electromagnetic phenomenon. He knew that substituting numerical values for $\varepsilon_0$ and $\mu_0$ does lead to the observed value$^{19}$ for $c$. But there are many other of aspects to light, which must all be checked to see if the equations correctly predict them. So we need to work on those, after introducing some helpful machinery in the following sections. First, however, a few remarks:

- There are many other interesting solutions besides plane waves, for example, spherical waves that spread from a point (Chapter 38).
- Because the Maxwell equations are linear, we can get more solutions by superposing (adding) the fields of two solutions at each point of spacetime. So the rich world of interference phenomena observed with light and other EM radiation is all contained in the electromagnetic-wave theory of light.

$^{19}$Maxwell was not the first to observe this: See Section 18.6' (page 288).
• All kinds of wave phenomena display interference, for example sound, ripples on water, and so on. But here we get the more specific prediction that there are polarizations of light corresponding to the directions $\hat{E}$ transverse to the direction of propagation (Equation 18.21). Indeed, transverse polarization effects are prominent in a classroom demonstration using microwave radiation generated via electric currents. And visible light in vacuum was already well known in Maxwell’s time to display two independent polarizations, a detailed agreement with the electromagnetic theory. In contrast, there is only one kind of sound wave in air or water (one “polarization”). Sound in a rigid solid like steel has a three-dimensional space of polarizations, because steel can elastically resist both compression (longitudinal) and shear (transverse) deformation. In short, light differs from all kinds of sound by having no longitudinal polarization, and Maxwell’s theory correctly predicted that.

• Notice from Your Turn 18C that the solutions we have found so far have $\vec{E}$ and $\vec{B}$ perpendicular to each other, and that each is perpendicular to the direction of motion $\hat{z}$. Also notice that if each varies sinusoidally with time and space, then they are in phase with each other. So at any instant, there are periodically-spaced planes where both equal zero! Normally we don’t notice that, because light waves rush around so fast that we can only perceive the time-averaged fields. But we can use superposition to create a standing wave, and it really does have “nodes” with zero field.

18.7 COMPLEX EXPONENTIAL NOTATION FOR WAVES

Although you showed in Your Turn 18C that there is nothing special about cosines, nevertheless, sines and cosines are a convenient basis, from which any waveform can be constructed by Fourier synthesis. An even more convenient basis is the complex exponentials; we will usually write waves in terms of the basis functions

$$\Phi_{k,\omega}(t, \vec{r}) = e^{i(k\cdot\vec{r} - \omega t)}. \tag{18.22}$$

Of course, $\vec{E}$ and $\vec{B}$ must still be real-valued vector fields, so in any formula involving $\Phi_{k,\omega}$ we will eventually need to take the real part to get the physical fields. But in intermediate steps, the complex notation is often helpful, because sine and cosine exchange roles under differentiation, whereas the derivative of exponential is always still exponential:

$$\frac{\partial \Phi_{k,\omega}}{\partial t} = -i\omega \Phi_{k,\omega}, \quad \hat{\nabla} \Phi_{k,\omega} = i\vec{k} \Phi_{k,\omega}.$$

Let’s use complex notation to redo what was done in the preceding section, and extend it in two ways. We’ll write a trial solution of the form

$$\vec{E}(t, \vec{r}) = \frac{1}{2} \vec{E}_0 \Phi_{k,\omega}(t, \vec{r}) + \text{c.c.} \tag{18.23}$$

---

20See Media 1.
The notation “c.c.” denotes the complex conjugate of whatever precedes it, and guarantees that the overall expression is real.\textsuperscript{21} The factor of one half says that specifically we are taking the real part of the first term. The notation $\vec{E}$ refers to a constant vector, called the complex amplitude (or “Jones vector”) of the real vector field $\vec{E}(t, \vec{r})$.

Compared to Equation 18.20, the two extensions we are considering are that:

- The wavevector $\vec{k}$ need not point along $\hat{z}$.
- The amplitude vector $\vec{E}$ need not be real. Write it as $\vec{E} = \frac{1}{2}(\vec{E}(R) + i\vec{E}(I)) + \text{c.c.}$, so that

$$\vec{E} = \frac{1}{2}(\vec{E}(R) + i\vec{E}(I))(\cos(\cdot) + i\sin(\cdot)) + \text{c.c.} = \vec{E}(R)\cos(\cdot) - \vec{E}(I)\sin(\cdot).$$

Now impose the Maxwell equations one by one.

### 18.7.1 Electric Gauss law

In vacuum, the electric Gauss law says $\vec{\nabla} \cdot \vec{E} = 0$. Spatial gradients are easy to compute by the rule $\vec{\nabla} \Phi_{k,\omega} \rightarrow i\vec{k}\Phi_{k,\omega}$, so Equations 18.22 and 18.23 give

$$\vec{\nabla} \cdot \vec{E} = 0 = \frac{1}{2}[\vec{\nabla} \cdot (i\vec{k}\Phi) + \vec{E}^* \cdot (-i\vec{k}\Phi^*)] = \frac{1}{2}\vec{k} \cdot [(\vec{E}(R) + i\vec{E}(I))(\cos(\cdot) + i\sin(\cdot)) + \text{c.c.}]$$

$$= \vec{k} \cdot (-\vec{E}(R)\sin(\cdot) - \vec{E}(I)\cos(\cdot)) \tag{18.24}$$

The ellipses denote $\vec{k} \cdot \vec{\nabla} - \omega \hat{z}$. In the last step, half of the terms get clobbered by taking the real part.

Equation 18.24 must hold at every point of space, at every time. The only way this could happen is if the coefficients of $\sin(\cdot)$ and $\cos(\cdot)$ separately vanish. So each of $\vec{k} \cdot \vec{E}(R) = 0$ and $\vec{k} \cdot \vec{E}(I) = 0$ must hold, or

$$\vec{k} \cdot \vec{E} = 0. \tag{18.25}$$

In short, when dealing with linear expressions in the fields, we don’t need to think explicitly about the complex conjugate terms. From now on, we’ll abbreviate logic like the foregoing by passing directly from an equation of the form $\frac{1}{2}B\Phi_{k,\omega} + \text{c.c.} = 0$, where $B$ is some complex constant, to the conclusion\textsuperscript{22} that $B = 0$.

For the special case where $\vec{k} = k\hat{z}$, Equation 18.25 is the same transversality condition that we found earlier (Equation 18.21).

### 18.7.2 Faraday law

If $\vec{B}$ is a plane wave, it seems a reasonable guess that $\vec{B}$ will be too, so extend the trial solution:

$$\vec{B}(t, \vec{r}) = \frac{1}{2}\vec{\Phi}_{k,\omega}(t, \vec{r}) + \text{c.c.}$$

\textsuperscript{21} Beware that many authors abbreviate by dropping the 1/2 and the +c.c.; you are supposed to understand that in any complex expression, the real part is meant. This convention can lead to confusion when dealing with nonlinear expressions like energy density. We will always write the fields in full, as in Equation 18.23.

\textsuperscript{22} Nonlinear expressions require more care; see Section 18.10.
where $\vec{B}$ are three more unknown complex constants. Note that we allow for the possibility that the magnetic field’s variation may be shifted in phase relative to that of the electric field: One advantage of the complex exponential notation is that such a shift can be represented as a complex multiplicative factor in $\vec{B}$.

Again, every $\vec{V}$ becomes a factor of $\pm ik\vec{r}$, and also $\partial / \partial t$ becomes $\mp i\omega$. Thus, the Faraday law becomes

$$\frac{1}{2} ik \times \vec{E} \Phi_{k,\omega} + \text{c.c.} = -(-i\omega)\vec{B} \Phi_{k,\omega} + \text{c.c.}$$

Solving gives

$$\vec{B} = (\vec{k}/\omega) \times \vec{E}.$$ 

We conclude that $\vec{B}$ must be perpendicular to $\vec{k}$, and also to $\vec{E}$. Moreover, the spatial and temporal variation of $\vec{B}$ match that of $\vec{E}$ (no relative phase shift). We see this from the fact that $\vec{B}$ is a real constant vector crossed with $\vec{E}$. These results generalize what you found in Your Turn 18C.

18.7.3 Magnetic Gauss law

Similar logic as before reduces this equation to $\vec{k} \cdot \vec{B} = 0$, but we already knew that from the Faraday law. Thus, we get no additional restriction on our trial solution.

18.7.4 Ampère law

$$i\vec{k} \times \vec{B} \Phi_{k,\omega} + \text{c.c.} = c^{-2}(-i\omega)\vec{E} \Phi_{k,\omega} + \text{c.c.}$$

$$\vec{k} \times \left( \frac{\vec{k}}{\omega} \times \vec{E} \right) = -c^{-2}\omega \vec{E}.$$ 

**Your Turn 18D**

Simplify the triple cross product to show that $ck = \omega$ as before (Equation 18.21).

18.7.5 Traveling wave with attenuation

In the preceding sections, $\vec{k}$ and $\omega$ were real constants. But Equation 18.18 is also interesting if $\vec{k} = \vec{k}\,(R) + i\vec{k}\,(I)$ is not real. If we sit at one position $\vec{r}$ and record the wave as it goes by, then repeat at a position $\vec{r} + \Delta\vec{r}$, the second time series will be shifted in time (by $\vec{k}\,(R) \cdot \Delta\vec{r}/\omega$) but also decreased in amplitude by a factor of $\exp(-\vec{k}\,(I) \cdot \Delta\vec{r})$. Such waves do not arise in vacuum, but later we’ll see they could describe signals traveling through a cable with a current leak that gradually saps its strength.\(^{23}\)

\(^{23}\)You’ll explore an application without resistance in Problem 18.4, and Section 49.4.1 (page 695) will discuss one involving media (total internal reflection).
18.7.6 Summary

There are plane-wave solutions in vacuum that move in any direction, with any frequency, and any polarization as long as it’s perpendicular to the direction of propagation. All such solutions move at the same speed $c$. All have $\vec{B}$ perpendicular to, but in phase with, $\vec{E}$. Each satisfies the dispersion relation $ck = \omega$. The name is a bit misleading, because in general dispersion refers to the possibility that the phase velocity $\omega/k$ may depend on frequency. What we have found is that in vacuum, the dispersion relation predicts constant phase velocity, and hence no dispersion. Light in transparent media, and signals in coaxial and other cables, can all have more interesting dispersion relations.\(^{24}\)

18.8 POTENTIALS BEYOND STATICS

18.8.1 $\vec{E}$ and $\vec{B}$ can still be represented by using potentials

We found simplified reformulations of electrostatics and magnetostatics by introducing potentials $\psi$ and $\vec{A}$. Can we do something similar for the full Maxwell equations?

We still have $\vec{\nabla} \cdot \vec{B} = 0$, so we can still write $\vec{B} = \vec{\nabla} \times \vec{A}$ for some vector potential $\vec{A}$.\(^{25}\) However, we no longer have $\vec{\nabla} \times \vec{E} = \vec{0}$, so electrons feel a nonconservative force,\(^{26}\) unlike in statics. That is, there is no function whose gradient is minus the electric field. Nevertheless, we can find a different vector quantity whose curl equals zero: Expressing Faraday in terms of $\vec{A}$ gives that $\vec{E} + \partial \vec{A}/\partial t$ is curl-free. Accordingly, we can construct a function whose negative gradient equals that quantity. We’ll continue to call it the “scalar potential” $\psi$, but keep in mind that $\psi$ can no longer be interpreted as potential energy of a test body per unit charge. In short, we can always find potential functions such that\(^{27}\)

$$\vec{E} = -\vec{\nabla} \psi - \frac{\partial \vec{A}}{\partial t} \quad \text{and} \quad \vec{B} = \vec{\nabla} \times \vec{A}. \quad (18.26)$$

Equations 18.26 let us express six unknown fields ($\vec{E}$ and $\vec{B}$) in terms of just four unknown potentials ($\vec{A}$ and $\psi$), a significant simplification. We will soon see that further simplifications arise when we substitute this representation into the Maxwell equations.

18.8.2 Gauge invariance and Coulomb gauge also extend beyond statics

One key idea about potentials in the static case was gauge invariance.\(^{28}\) Does it still hold good?

---

\(^{24}\)See Section 11.2.3 (page 172), Problems 18.5–18.9, Section 21.2 (page 308), Problem 18.8 (page 294), Section 49.3.5 (page 694), and Section 54.3 (page 754).

\(^{25}\)Section 15.3.5 (page 226).

\(^{26}\)Section 18.3.2 (page 270).

\(^{27}\)This result addresses Hanging Question #G (page 22).

\(^{28}\)Section 15.4 (page 227).
Your Turn 18E

Show that the substitutions
\[ \vec{A} \to \vec{A} + \vec{\nabla} \Xi, \quad \psi \to \psi - \frac{\partial \Xi}{\partial t} \]

gauge transformation \hspace{1cm} (18.27)

don’t change the electric or magnetic fields in Equation 18.26. Here \( \Xi \) is any scalar function of space and time.

Thus again, the potentials are not uniquely determined by the fields, and we can use that fact to insist on a subsidiary condition if doing so simplifies our equations. For the moment, we will again impose Coulomb gauge:
\[ \vec{\nabla} \cdot \vec{A} = 0. \]

[15.14, page 227]

The proof that this is always possible locally is the same as it was in statics (Section 15.3.5), because we did not modify the gauge transformation formula for \( \vec{A} \) (the first of Equations 18.27 is the same as the formula in Section 15.4, page 227).

We can now substitute Equation 18.26 into the Maxwell equations and simplify by using Coulomb gauge. As in statics, \[ \vec{\nabla} \cdot (\vec{\nabla} \times \vec{A}) = 0 \] is an identity, so we can forget the magnetic Gauss law. The Faraday law is also automatically satisfied, so forget it, too. We are left with four equations in the four unknowns \( \vec{A} \) and \( \psi \):\( ^{29} \)

\[ \nabla^2 \psi = -\frac{\rho}{\varepsilon_0} \]  
(electric Gauss in Coulomb gauge), and \hspace{1cm} (18.28)

\[ \nabla^2 \vec{A} = -\mu_0 \vec{j} + \mu_0 \varepsilon_0 \left( \nabla \frac{\partial \psi}{\partial t} + \frac{\partial^2 \vec{A}}{\partial t^2} \right). \]  
(Ampère in Coulomb gauge) \hspace{1cm} (18.29)

It’s tempting to say that we have just found another resolution of Hanging Question #D (page 14) (“eight equations in six unknowns” paradox), but we must be a bit careful. The four equations just given are only correct if \( \vec{\nabla} \cdot \vec{A} = 0 \), which looks like a fifth equation constraining the four potential functions. However, when we take the divergence of both sides of Equation 18.28, then Equation 18.29 shows that this combination is automatically satisfied; it does not constrain the potentials. So effectively, we do have just four independent equations in four unknowns.

18.8.3 Coulomb gauge can be augmented if the charge density is zero

We can simplify still more if we’re studying a region with zero net charge density.\( ^{30} \) (There may still be currents, however, as in a neutral wire.)

Even if we restrict to Coulomb gauge, there still is some further freedom to apply certain gauge transformations, because transforming with any function \( \Xi \) that obeys \( \nabla^2 \Xi = 0 \)

---

\( ^{29} \)The derivation of these formulas depends on our default choice of cartesian coordinates. The left-hand sides would look more complicated in curvilinear coordinates.

\( ^{30} \)Actually, Chapter 38 will achieve a similar simplification even with charges present, but we don’t need that much power yet.
will not spoil the Coulomb gauge condition. Let’s try

\[ \Xi(t, \vec{r}) = \int_{t_0}^{t} dt' \psi(t', \vec{r}). \]  

(18.30)

**Your Turn 18F**

Show that, if \( \rho_q = 0 \) everywhere, then:

a. Gauge transformation by the function in Equation 18.30 preserves Coulomb gauge, and
b. This transformation eliminates the scalar potential altogether (transforms it to zero).

There can still be electric fields—they are just being represented by the time derivative terms in Equation 18.26. In short, we have now found that in vacuum we can reduce still further from four unknown potential functions to just three.

**Your Turn 18G**

a. Show that the electric Gauss law is now automatically satisfied (an identity).

b. Show that what remains is actually three decoupled equations in three unknowns:

\[ \nabla^2 \vec{A} = -\mu_0 \vec{j} + \mu_0 \varepsilon_0 \frac{\partial^2 \vec{A}}{\partial t^2} \quad \text{in Coulomb gauge extended by } \psi = 0. \]  

(18.31)

We have thus found yet another resolution to Hanging Question #D, for the special case where net charge density is zero. As before, the additional condition \( \vec{\nabla} \cdot \vec{A} = 0 \) is balanced by the fact that the divergence of Equation 18.31 is automatically satisfied. Moreover, we get the simplification that in this gauge Equations 18.31 decouple, much as they did in magnetostatics.31

**18.9 WAVES VIA POTENTIALS**

We can use the representation of fields by potentials to explore plane wave solutions in vacuum more systematically. For example, we can quickly recover the results in Section 18.5, and other results too. The plane wave solutions of Equation 18.31 moving along \( \vec{z} \) take the form

\[ A(t, \vec{r}) = \frac{1}{2} \Phi_{\vec{k}, \omega}(t, \vec{r}) + \text{c.c.}, \]  

(18.32)

where the polarization vector \( \vec{\zeta} \) is a constant lying in the xy plane, \( \vec{k} = k \vec{z} \), and \( \Phi_{\vec{k}, \omega} \) is one of the family of complex traveling waves in Equation 18.22 (page 278).

31See Your Turn 15F (page 229).
Your Turn 18H

Show that more generally, Equation 18.32 yields plane wave solutions for $A$ that move in any direction, as long as $\mathbf{k}$ and $\omega$ obey the dispersion relation

$$||\mathbf{k}|| = \frac{\omega}{c} \quad \text{where} \quad c = 1/\sqrt{\mu_0 \varepsilon_0} \quad (18.33)$$

and $\mathbf{\zeta}$ is perpendicular to $\mathbf{k}$.

We have simplified the Maxwell equations, and streamlined the derivation of plane waves, but it may seem that we have been too successful: For any choice of $\mathbf{k}$, Equation 18.32 seems to give three linearly independent solutions, whereas the analysis in either Section 18.5 or Section 18.7 gave only two (for the two directions perpendicular to $\mathbf{k}$)! The resolution to this puzzle is that our trial solution only works if $\mathbf{\zeta} \perp \mathbf{k}$, because Equation 18.31 is only equivalent to the Maxwell equations in Coulomb gauge. Thus, in vacuum the longitudinal polarization is not physical; it does not correspond to a solution of the Maxwell equations.

Your Turn 18I

Work out the electric and magnetic fields arising from the solution Equation 18.32, and hence the relation between the polarization vector $\mathbf{\zeta}$ and the vector $\mathbf{E}$ appearing in Equation 18.20. Show that as before, $\mathbf{E}$, $\mathbf{B}$, and $\mathbf{k}$ are mutually perpendicular.

18.10 COMPLEX POLARIZATIONS

18.10.1 Linear, circular, elliptical

If $\mathbf{\zeta}$ is a vector with real components, then $\mathbf{E}$ oscillates about the $\pm \mathbf{\zeta}$ direction; we say the plane wave is linearly polarized, because the tip of its $\mathbf{E}$ vector oscillates back and forth on a line in the plane perpendicular to $\mathbf{k}$.

But there are other options. There’s nothing mathematically wrong with a complex polarization vector, just as in our earlier derivation (Section 18.7). Indeed, this is a new and physically interesting plane wave.

Your Turn 18J

If you assumed that $\mathbf{\zeta}$ was real when you worked Your Turn 18I, work through it again without this assumption. Specifically, work out $\mathbf{E} \cdot \mathbf{k}$, $\mathbf{B} \cdot \mathbf{k}$, and $\mathbf{E} \cdot \mathbf{B}$. 

Light can be created in circularly polarized states.
Your Turn 18K

a. Consider the wave with \( \hat{k} = k \hat{z} \) and \( \hat{\zeta} = \hat{x} + i \hat{y} \) (times a real constant). If we sit at a fixed location in space, say the origin of coordinates, and watch \( \vec{E}(t, \vec{0}) \) as time goes by, what figure does its tip trace out? Explain why this wave is said to be **circularly polarized**.

b. Repeat with \( \hat{\zeta} \propto \hat{x} + 2i \hat{y} \) and interpret such **elliptically polarized** solutions.

Another name for circular polarization is **helicity**.

18.10.2 Helicity basis for circular polarization

Starting from a particular \( \hat{k} \), choose a pair of real unit vectors \( \hat{\zeta}^{(1)}, \hat{\zeta}^{(2)} \) perpendicular to \( \hat{k} \) and forming a right-handed triad with it. That is, \( \hat{\zeta}^{(1)} \times \hat{\zeta}^{(2)} = \hat{k} \). Any polarization for the given \( \hat{k} \) can be written as a linear combination of these two basis vectors.

Alternatively, we can define complex basis vectors:

\[
\hat{\xi}^{(\pm)} = (\hat{\zeta}^{(1)} \pm i \hat{\zeta}^{(2)})/\sqrt{2}, \quad \text{helicity basis} \tag{18.34}
\]

Any polarization vector \( \hat{\zeta} \) can be written as a (possibly complex) linear combination either of \( \hat{\zeta}^{(1,2)} \), or of \( \hat{\zeta}^{(3)} \). If the polarization vector is purely along \( \hat{\zeta}^{(+)} \), then the wave is said to be circularly polarized with **positive helicity**, and similarly for a pure \( \hat{\zeta}^{(-)} \) wave (which is **negative helicity**).\(^{32}\)

18.11 PLUS ULTRA

18.11.1 Spherical waves foreshadowed

You may ask, “What was the point of redoing everything with potentials? Section 18.5 already found plane waves directly in terms of \( \vec{E} \) and \( \vec{B} \), and it wasn’t much easier in Section 18.9.” One answer is that the calculations will get harder, and the benefit of the potential formulation will therefore become more important, when we study spherical waves (Chapter 38) and beams (Chapter 41).

18.11.2 Dielectric media

Section 6.5.1 (page 79) argued that electric fields in a dielectric medium obey the same Gauss law as in vacuum but with a modified permittivity \( \epsilon > \epsilon_0 \). Applying that change to the analysis of this chapter leads us to expect **transverse EM waves with reduced speed** \( (\epsilon \mu_0)^{-1/2} \).

\(^{32}\)Beware that different authors disagree about the convention for which is positive and which negative.
18.11.3 Necessity of the field concept

Let’s pause to underscore the character of Maxwell’s advance: Ampère’s law for magneto-statics had to be modified for dynamics, not because of any electromagnetic phenomenon known at the time, but for mathematical consistency. That modification led to a prediction of new phenomena. One class of those phenomena resembled light, which was not known at the time to have any relation to electricity nor to magnetism.

In electrostatics, the electric field could be regarded as a mathematical convenience—introducing it into the formulas was optional. We could, after all, just say that all charges exert forces on each other directly, following Coulomb’s law. Although we found a useful concept of electrostatic energy density in the space between capacitor plates, this interpretation, too, was physically optional—we could just say that the energy of a capacitor was the total potential energy of all the separated charges in each others’ force fields.

Waves change everything. We’ll see that shaking (accelerating) a charge generates these waves, and they in turn can shake other distant charges. Suppose that we shake a charge for a while, then stop. Suppose too that the nearest other charges are far away. Then there will be a period after the original charge has lost some energy, but before any other charge has gained energy. Hanging Question #H (page 32) already asked: Where is the energy during that time?

As mentioned in Section 18.3.4, Maxwell and his contemporaries believed that the so-called vacuum was actually filled with some substance, the stuff that jiggles when a wave goes by. The fields were just the state of motion and deformation of that stuff, and their stored energy was just its kinetic and deformation energy, just as when sound passes through steel. Einstein realized, however, that this stuff (the “æther”) had to have contradictory physical properties. Eventually he concluded that it didn’t exist, or at least not as any ordinary substance. Then the question comes back to us: If vacuum is truly empty, what carries that energy? It’s easy to say, “It’s in the fields themselves,” but we’ll need to make sure this is a meaningful statement.33

Section 18.11’ (page 289) discusses some 20th century developments with a similar flavor.

FURTHER READING

Semipopular:
Zeeman effect: Media 10.

Intermediate:

33See Chapter 35.
18.4.1’a Connection to ohmic materials
The Maxwell equations and the Lorentz force law may look very clean, and they may be clearly applicable to, say, one charged particle flying through vacuum between two charged plates. But the connection to more familiar situations—for example, resistors—may not be so clear.

For a mechanical analogy, consider the equally humble matter of sedimentation. We take a beaker with a suspension of particles, mix well, then wait. If the particles are heavy, then over time they settle to the bottom of the beaker; if they are microscopic, they may instead arrive at an equilibrium concentration profile enriched at the bottom and depleted at the top; but in no case do they appear to be obeying Newton’s $z = z_0 - \frac{1}{2}gt^2$! The answer to this puzzle is that there is more in the beaker than the particles of interest, and more acting on them than gravitation. Indeed, surrounding water molecules are constantly making random collisions with the suspended particles, impeding their progress and diverting some of their kinetic energy into heat. If we don’t wish to account for each collision in detail, a phenomenological model may be accurate enough; in the colloidal setting, a suitable model says that a net “viscous friction” force proportional to velocity is added to gravitation:

- The gravitational force on a particle is certainly still present, as in vacuum, but unbalanced collisional forces cancel it and the particle rapidly comes to constant “terminal” speed.
- The gravitational potential energy drop as a particle falls is also still present, but each particle’s kinetic energy also rapidly saturates to a constant; after that, the lost potential energy ends up as heat.

Similarly, an ohmic material (for example salt water) impedes the flow of charge carriers.

- The electric force on a carrier from an external source is certainly still present, but unbalanced collisional forces cancel it and the carrier rapidly comes to constant “drift” speed.
- The electrostatic potential energy drop as a carrier advances is also still present as in vacuum, but each carrier’s kinetic energy as it arrives at the low-potential end is the same as when it began; the lost total electrostatic potential energy ends up as heat.

18.4.1’b Stumbling yet pulled forward
We are all in the gutter, but some of us are looking at the stars.
— Oscar Wilde

The argument from mathematical consistency in Section 18.4.1 looks nearly trivial to us because we have the clean notation of vector calculus, and clean conceptions of quantities like charge density. What makes us call Maxwell a genius was his ability to see through the fog of the unclear notation and conceptual framework of his day.

Maxwell never said he was motivated by any symmetry of the equations upon exchange of $\vec{E}$ and $\vec{B}$, which in any case was obscured by his presentation. The actual reasoning that he used to motivate his change to Ampère’s law is hard to express in modern language, although it does seem reasonable to suppose that the current associated with bound charge in a real dielectric medium

---

34 Although effectively reduced by buoyancy.
35 Heaviside uncovered this symmetry, but only much later.
36 See Section 49.2.1.
might be accompanied by a similar current from the æther that Maxwell and others believed filled empty space.\textsuperscript{37} In fact, in his first publication introducing a displacement current\textsuperscript{38} Maxwell does attribute it to distorted ‘æther cells.’

Nor was Maxwell’s original form for the displacement term quite correct. He quietly changed it to the present form in a later work,\textsuperscript{39} and only then found a satisfactory derivation of the wave equation. Even with that change, his equations were still inconsistent due to a faulty notion of charge; fixing this flaw required yet another quiet revision.\textsuperscript{40}

Nor was Maxwell explaining some existing, definitive experimental result: The constant of proportionality $\mu_0 \varepsilon_0 \approx 1.1 \cdot 10^{-17} \text{m}^{-2} \text{s}^2$ on the new term is extremely small, so no experiment envisioned in his day could directly confirm or refute it. One would need fields with extremely fast time dependence (large time derivative) to start seeing the effects of this hypothetical term on laboratory length scales.\textsuperscript{41}

So how did Maxwell manage to keep incrementally approaching the true equations, despite all the stumbles? He left no real record outside his publications, but any physicist can imagine a possibility:\textsuperscript{42} Maxwell’s eyes may have been fixed on the distant mountains, not on his feet. Once the long-sought goal appears to be nearly within reach, we are seized with an overmastering urge to steamroll the obstacles. A genius steamrolls bigger obstacles than the rest of us, but nearly every discovery great or small goes through this phase. A genius has the exquisite extra sense to say, “Eventually somebody will figure out that detail that’s eluding me right now,” and be right about that.\textsuperscript{43} But each of us can develop a smaller version of that sense by studying the thinking of others.

In the 1860s, many scientists shared the goal to relate electromagnetism and optics. Naturally, lesser minds got hung up on Maxwell’s inconsistencies, ad hoc changes, and missing details. For many scientists, the conclusive proof came only with Hertz’s detailed experimental confirmation that a purely electrical circuit (generating high frequencies via a spark gap) created the predicted propagating waves, which were detected by their purely electrical effects and shared all the key phenomena of light.

\subsection*{18.6’a On the speed of light}

Aristotle held that the speed of light was infinite. But already in the eleventh century, Ibn Sina and al-Haytham broke with Aristotle’s authority, believing that the speed of light, although large, was finite. Centuries later, Galileo proposed to measure the speed with a terrestrial experiment similar to one that could measure the speed of sound. The experiment was attempted after Galileo’s death, but the

\textsuperscript{37}How then could Einstein retain this term \textit{even after denying the reality of the æther}? In the intervening decades, other scientists had developed the more abstract viewpoint used today. Ironically, however, the name “displacement current” inspired by the analogy has stuck.

\textsuperscript{38}Part 3 of “On physical lines,” 1862.

\textsuperscript{39}“Dynamical theory,” 1865.

\textsuperscript{40}The “Treatise,” 1873.

\textsuperscript{41}Maxwell wrote in 1868: “This part of the theory… has not been verified by direct experiment. The experiment would be a very delicate and difficult one.” Actually, Joseph Henry had speculated in 1842 that an electric spark from a Leyden jar was a high-frequency alternating current. B. Fedderson confirmed this photographically in 1859, but scientists did not immediately see the implications for confirming Maxwell’s theory.

\textsuperscript{42}Chalmers, 1975.

\textsuperscript{43}In Maxwell’s case, that “somebody” was himself, but later. Many others contributed later still, notably H. Lorentz.
apparatus wasn’t able to discriminate between infinite speed and the actual value. However, a few years later Ole Römer succeeded with a clever astronomical measurement. Terrestrial measurement had to await another clever idea from Fizeau (Chapter 29).

W. Weber and R. Kohlrausch demonstrated in 1856 that independently measured values of \( c_0 \) and \( \mu_0 \) led to a quantity with no charge units that agreed with \( c \) to order of magnitude. But lacking Maxwell’s theory, the quantity they reported had an extra factor of \( \sqrt{2} \), which obscured just how good the agreement was, and so they did not emphasize their result. Such a skeptical attitude to a numerical coincidence was permissible—maybe even required—in the absence of any, independently grounded theory predicting it. Maxwell supplied that theory, adding “We can scarcely avoid the conclusion that light consists in the transverse undulations of the same medium which is the cause of electric and magnetic phenomena.”

Specifically, G. Kirchhoff noted that Weber and Kohlrausch’s measurements implied waves with speed \( c = 3.107 \times 10^8 \) m/s, close to Fizeau’s measured 3.148 m/s. Later Maxwell, too was impressed by this close agreement, asserting in 1862 the identity of the two phenomena. Later (1868), Maxwell and C. Hockin made an improved measurement of \( (c_0\mu_0)^{-1/2} \approx 2.88 \times 10^8 \) m/s, and compared it to Foucault’s improved measurement of light speed 2.9836.\(^{44}\)

18.11’ On the guidance of mathematical consistency

Maxwell’s first great article\(^{45}\) omitted the vacuum displacement charge flux term but explicitly pointed out that the resulting equations imply \( \nabla \cdot j = 0 \); he added “we know little of the magnetic effects of any currents which [have \( \nabla \cdot j \neq 0 \)].” So one motivation for his introduction of the new term in later articles was the need for mathematical consistency in that general situation. Such a big win makes us wonder if this sort of thing happens a lot.

- The discovery of the tau lepton in 1975 led directly to the prediction of top and bottom quarks and tau neutrino, via an argument of mathematical consistency (“gauge anomaly” cancellation—Bouchet, Iliopoulos, Meyer).
- When superstring theorists tell us there must be extra hidden dimensions, again this prediction stems from a mathematical inconsistency of all other cases. This prediction is still awaiting confirmation, however.

Why study the structure of the theory so much? Why not just do real-world problems? You need to develop a sense of what makes a theory great. This sixth sense can be helpful in your real world. Later, when you create something, you’ll get that tingling sense of recognition, this feels right, some subliminal echo of great theories you have met, the click of links falling into place automatically.

\(^{44}\)Later still, (1892) Abraham obtained a still more precise measurement of \( c_0\mu_0 \).

\(^{45}\)On Faraday’s lines of force,” 1855–6.
18.1 Parity II
Work Problem 15.4a (page 238) again, this time for the full Maxwell equations (Equations 0.1–0.4, page 2).

18.2 Faraday
A thin ring of copper spins freely in zero gravity, about an axis that includes one of its diameters. The ring’s radius is 0.1 m. Its initial angular velocity is \( \omega_0 \), a certain number of radians per second.

At time zero, we turn on a magnetic field \( \mathbf{B}_0 \), with magnitude 0.02 T and directed perpendicular to the axis of rotation. The ring’s initial kinetic energy gets dissipated in resistive heating of the ring. Calculate the time needed for the angular frequency to decrease to \( \omega_0 / \exp(1) \) (the “e-folding time”).

The electrical resistivity of cold-drawn copper is \( 46 \cdot 10^{-8} \Omega \cdot m \), and its mass density is \( 9.0 \cdot 10^3 \) kg/m\(^3\). You may assume that the slowdown is gradual, or \[ \frac{d}{dt} \ln \omega \ll \omega_0. \]

18.3 Feeling the heat
In this problem, you will develop a simple model for estimating radio-frequency (RF) energy absorption in a patient undergoing an MRI scan.

a. The wavelength of an RF wave is bigger than a person, so suppose that a spatially uniform, but time-varying magnetic field \( \mathbf{B}(t) = B(t) \cos \omega t \) is applied. Apply Faraday’s law to a circular path in a plane perpendicular to \( \mathbf{B} \) to find the amplitude of the resulting electric field. Your answer depends on the radius \( R \) of the circular path; later we will set \( R \) to a value comparable to a human radius.

b. Model the patient as a uniform conductor with electrical conductivity \( \kappa \). Use the ohmic relation (Equation 8.9, page 119) to find the average power dissipated in the conductor per volume. Actually, the RF signal is not continuous; it consists of pulses of duration \( \Delta t \) which come once every repetition period \( T_R \), so make the appropriate correction.

c. It’s customary to report the “specific absorbed rate,” which is power per unit body mass.

Find the SAR in terms of body mass density \( \rho_m \) and \( \kappa, R, \omega, \Delta t, T_R \).

d. The pulse duration, field strength, and angular frequency are related by the requirement that the pulse rotate proton spins by an angle \( \pi/2 \). You can take as given that this requirement amounts to \( \Delta B = 2\pi / (2\gamma \Delta t) \) and \( \omega = \gamma B(0) \), where the “gyromagnetic ratio” \( \gamma \) of a proton is some constant and \( B(0) \) is the background magnetic field, a known number. Use this information to eliminate \( \Delta B \) and \( \omega \) from your formula for SAR.

e. Now substitute typical human values: \( R \leq 0.17 \) m, \( \kappa \approx 0.3 \Omega^{-1} \) m\(^{-1}\). And use typical instrument values \( B(0) \approx 0.5 \) T and \( T_R \approx 1 \) s. Also, \( \gamma \approx 2.7 \cdot 10^8 \) Hz/T.

---

\(^{46}\)Conductivity is the reciprocal of resistivity.
f. Safety requires that we not heat the patient too much! So demand that SAR < 0.4 W/kg.
Find the corresponding requirements on \( \Delta t \) and also on \( \delta B \).

18.4 Lumped-element transmission line

This problem explores a circuit that is sometimes useful for signal conditioning, for example, removing noise known to have a specific frequency. A filter like this was added to the MicroBooNE experiment’s electronics at Fermilab.

The main text introduced a solenoid. More generally, any circuit element that obeys the linear relation Equation 18.12 (page 271) for some constant \( C \) is called an inductor. You can purchase devices that approach this idealized behavior (approximately, over some frequency range).

Consider a chain of discrete modules each with circuit diagram like the ones shown in Figure 18.3. The symbol on the left denotes an alternating, fixed-voltage power supply. Each module contains an inductor with inductance \( L \) and a capacitor with capacitance \( C \).

Write expressions analogous to the ones in Section 11.2.2 for the cable equation but appropriate to this situation (inductors, no resistors). Unlike in the cable equation, however, we will not take any continuum limit.

a. Show that the quantity \( LC \) has the dimensions \( T^2 \).

b. Following the analysis in Chapter 11, eliminate the currents \( I_j \) to get an infinite set of coupled, linear, ordinary differential equations in the remaining variables \( \{\psi_j\} \). The equations have constant coefficients, so we expect single-frequency solutions:

\[
\psi_j(t) = \frac{1}{2}\psi_0 e^{-i\omega t} + \text{c.c.} \tag{18.35}
\]

c. Substitute that trial solution to get an infinite set of coupled algebraic equations.

d. It still looks hard, but the equations are unchanged upon shifting everything one step in space. So our experience with related systems suggests that we make the trial solution

\[
\hat{\psi}_j = \psi_0 e^{i k}, \tag{18.36}
\]

where \( k \) is some constant. Substitute this into your algebraic equations for a given angular frequency \( \omega \) and see what \( k \) must be in order to get a solution.

e. If \( \omega \) lies in a certain range, there will be a real solution \( k \) to your condition. Then Equations 18.35–18.36 describe a wave traveling along the chain to infinity. Outside that frequency range, however, there will be no real solution; the transmission line has a cutoff. Find the allowed range of frequencies.

---

\[47\] In particular, an idealized inductor has negligible electrical resistance and capacitance.
18.5 \textit{Realistic transmission line}
This problem extends Problem 18.4. Figure 18.4 shows another transmission line, but made more realistic by the addition of resistance \( R \) along the segment shown and leak conductance \( G \). The figure shows the line as a series of lumped-element modules, but this time we will suppose that all four material properties \( R, L, C, \) and \( G \) are continuously distributed with linear densities \( r, \ell, \rho_c, \) and \( g \) respectively.\(^{48}\) Thus, you should initially consider a segment of length \( \Delta x \), with \( R = r \Delta x \) and so on; at an appropriate moment, take the limit \( \Delta x \to 0 \).

The line is infinitely long. We suppose that at some point an external agency imposes a harmonic potential \( \psi(0,t) = \frac{1}{2} \tilde{\psi} e^{-i \omega t} + \text{c.c.} \). We would like to find the solution everywhere else. The problem is time-translation invariant, so again a reasonable trial solution is harmonic: \( \psi(x,t) = \frac{1}{2} \tilde{\psi}(x) e^{-i \omega t} + \text{c.c.} \).

\begin{enumerate}
\item Follow the strategy in Chapter 11 to write a second-order differential equation for \( \tilde{\psi}(x) \).
\item The problem is also spatially translation invariant, apart from the imposed boundary condition, so seek a solution of the form \( \psi(x) = e^{ikx} \) where \( k \) is a function of \( \omega \) that you are to find.
\item The dependence of the wavenumber \( k \) on the angular frequency is called the cable’s dispersion relation. Why would it be desirable for \( k = \pm \left( \omega/v_{\text{cable}} + i \lambda \right) \), where \( v_{\text{cable}} \) and \( \lambda \) are independent of \( \omega \)?
\item The desirable condition does not generally hold, but Heaviside found that it does hold if the material parameters \( r, \ell, \rho_c, \) and \( g \) obey a certain relation. Find that condition.
\item Some resistance is unavoidable in any long cable. But it had previously seemed that any nonzero value of \( g \) would be a bad thing, to be avoided at all costs. Why did Heaviside disagree?
\end{enumerate}

18.6 \textit{Helicity basis}
The helicity basis is defined in Equation 18.34, starting from a choice of two vectors \( \tilde{\zeta}^{(1)}, \tilde{\zeta}^{(2)} \) perpendicular to each other and to \( \tilde{k} \) and forming a right-handed triad with it. This construction may sound too arbitrary.

\begin{enumerate}
\item Show that if we choose a different pair of unit vectors \( \tilde{\zeta}^{(1)}, \tilde{\zeta}^{(2)} \), which also make a right-handed, orthonormal triad with \( \tilde{k} \), then we get essentially the same basis. That is, \( \tilde{\zeta}^{(1)} \) is a scalar constant times \( \tilde{\zeta}^{(+)\ast} \) and similarly for \( \tilde{\zeta}^{(-)\ast} \).
\end{enumerate}

Now establish two properties that will be useful later:

\(^{48}\)Note that axial resistance \( R \) is proportional to length, but leak resistance is proportional to the inverse of the area of the cylindrical surface (Equation 8.8, page 119), and so the leak conductance \( G \) is proportional to \( \Delta x \).
b. Show that $\hat{\gamma}_{(\pm)} \cdot \hat{\gamma}_{(\pm)} = 1$ and $\hat{\gamma}_{(\pm)} \cdot \hat{\gamma}_{(\mp)} = 0$.

b. Show that $\hat{\gamma}_{(\pm)} \cdot \hat{\gamma}_{(\pm)} = 1$ and $\hat{\gamma}_{(\pm)} \cdot \hat{\gamma}_{(\mp)} = 0$.

c. Compute $\hat{\epsilon} \times \hat{\gamma}_{(\pm)}$ and express it in the helicity basis; show that the helicity basis vectors are eigenvectors of the operator “$\hat{\epsilon} \times$.” (This operator generates infinitesimal rotation about $\hat{\epsilon}$.)

18.7 Zeeman effect

Background: The Zeeman effect refers to the effect on atomic spectra of an applied magnetic field. Remarkably we can understand it (partially) without using quantum mechanics.

Problem: Consider a charged particle of mass $m$ and charge $q$ in an isotropic, 3D harmonic oscillator potential: $U(\vec{r}) = \frac{1}{2} k \|\vec{r}\|^2$. The particle has three independent normal modes of oscillation,\(^{49}\) all with the same angular frequency $\omega_0 = \sqrt{k/m}$.

a. Now we place this system in a static external magnetic field $\vec{B}$ directed along the $+z$ axis. Find the new frequencies of the resulting oscillation modes. You can suppose that $\vec{B}$ is “small” in any relevant sense, and work to leading nontrivial order in it.

[Hint: Treat oscillations in the $xy$ plane together, but separately from those along $z$. Try to guess two trial solutions for $xy$ motions that will still give solutions to Newton’s $\vec{F} = m\vec{a}$, even when $\vec{B}$ is turned on.]

b. The frequencies you found in (a) correspond to three kinds of radiation the system can emit. We have not yet systematically worked out the radiation by a moving point charge. However, from the symmetries of the problem and what you do know about light, you should be able to make an educated guess about what kinds of polarizations will be emitted. So find the frequencies and corresponding polarizations of radiation seen by an observer located far away on the $z$ axis.

c. Explain how observation of this radiation can be used to determine the charge/mass ratio of the electron, including its sign, even if the value of the spring constant $k$ is unknown.

d. Evaluate your answer for the frequency shift numerically, assuming $\|\vec{B}\| = 2$ T. Com-

---

\(^{49}\)You may assume that it moves much more slowly than the speed of light, and that its oscillation is affected very little by the radiation it gives off.
pare to the frequency of visible light. Is it a big effect?

Comments: P. Zeeman did this experiment in 1896. Following a suggestion by H. Lorentz, he looked for, and found, the polarization effect discussed in the problem. Lorentz then analyzed the data and obtained the charge to mass ratio that they implied. Crucially, that value and sign agreed with the one for cathode rays in free space, supporting the theory that ordinary atoms contained bound constituents—“electrons”—identical to the constituents of cathode rays. Zeeman and Lorentz shared a Nobel Prize for this work. Some highly magnetized stars (magnetars) have much bigger $B$ than what is attainable in the lab; the Zeeman effect gives a useful way to estimate the value of $B$ on a distant object.

### 18.8 Waves in conductive medium

An electromagnetic plane wave propagates a medium. The medium is not polarizable ($\varepsilon = \varepsilon_0$, $\mu = \mu_0$). However, it is electrically conductive, obeying an ohmic relation with conductivity $\kappa$:

$$j = \kappa E.$$

Assume the medium is everywhere electrically neutral. Find the dispersion relation for plane waves of angular frequency $\omega$ traveling through such a medium, and interpret it physically. How far into the medium can such a wave penetrate?

### 18.9 T2 Realistic transmission line II

This is a continuation of Problem 18.5. There you studied a class of problems described by four parameters $r$, $\ell$, $\rho_C$, and $g$. Four parameters is a lot—it may seem hard to catalog all the behaviors in such a high-dimensional space. But as often happens, things get much simpler after we nondimensionalize everything. You’ll now show that really, there is just a one-parameter family of distinct behaviors.

Specifically, we seek a combination of the four parameters with dimensions $L$ and then let $\tilde{x}$ be position divided by that scale. Then we define dimensionless $\tilde{k}$ as $k$ multiplied by that same scale, so that $\tilde{k} \tilde{x} = kx$. We also find another combination of the parameters with dimensions $T$ and then let $\tilde{t}$ be time divided by that scale. Then we define dimensionless $\tilde{\omega}$ as angular frequency multiplied by that scale, so that $\tilde{\omega} \tilde{t} = \omega t$.

a. Find expressions for length and time scales with the property that the dispersion relation from Problem 18.5 becomes

$$\tilde{k} = \pm \sqrt{\left(\tilde{\omega} + i\tilde{\omega} + ig\ell/(r\rho_C)\right)}.$$  (18.37)

We are interested in a problem where a signal generator fixes a definite potential $\psi_0(\tilde{t})$ at the point $\tilde{x} = 0$ in a semiinfinite wire. So we make the sign choice above that gives signals decaying as $\tilde{x} \to \infty$, and use the input signal as a boundary condition at $\tilde{x} = 0$.

In Problem 18.5 you found some solutions of the form $\frac{1}{2} e^{-i\omega t + i\tilde{k}(\omega) \tilde{x}} + c.c.$ Each such solution is a sinusoidal in time, with amplitude that decays exponentially with distance. But a sine wave of infinite duration does not communicate information! Now we wish to

---

50 Later experiments showed that the effect is sometimes more complicated than the simple classical picture discussed here (“anomalous Zeeman effect”). However, the qualitative conclusion about the sign of $q/m$ is valid.
assemble those solutions into something that looks more like a pulse. The pulse could represent the binary digit ‘1’ in a digital signal.

b. Use a computer to plot the function

$$\psi_0(\bar{t}) = 2a + \sum_{m=1}^{n_{\text{max}}} \frac{T}{\pi m} \sin(2\pi am/T)(e^{2\pi im/T} + e^{-2\pi im/T})$$

over the range $-0.1 < \bar{t} < 3.1$. Use illustrative parameter values $a = 0.05, T = 3, n_{\text{max}} = 1000$. How could we have predicted, without making the plot, that this particular function would generate a train of square pulses?

You now know how each term of the above sum will propagate along the cable, so you can use superposition to find what happens to the entire square pulse train. First, you’ll need to choose a value of the one relevant parameter characterizing the cable, as follows:

Write Equation 18.37 as

$$\bar{k} = \pm (\bar{\omega} + i)\sqrt{1 - ib/(\bar{\omega} + i)}.$$ 

Here $b$ is a dimensionless combination of $r, \ell, \rho_C,$ and $g$ that you are to find.

c. Try the cases $b = -2, 0,$ and $1$. Specifically, set $b = 0$ and find and plot the time course of electric potential as measured at the points $\bar{x} = 0, 1,$ and $2$. You can put all three resulting curves on a single set of axes. Then make two other plots with the other values of $b$ mentioned.

d. One value of $b$ has a special, nice property. Which one, and why?
19.1 Microwave waveguide
We have been studying wave solutions in infinite empty space, but there are interesting solutions in confined regions as well.

A waveguide consists of an infinitely long, hollow, rectangular prism along the z axis, with walls made of perfect conductor. It encloses the region \(0 < x < a, 0 < y < h\). Inside is vacuum, where the electric and magnetic fields must obey the wave equation as usual. All we need are solutions obeying boundary conditions appropriate to conductors.

Consider the following trial solution for the electric field in this region:

\[
\vec{E}(t, \vec{r}) = \frac{1}{2} \tilde{E}(y) \hat{x} e^{-i(\omega t - k z)} + \text{c.c.}
\]

Here \(\tilde{E}(y)\) is a function of \(y\) only, which you are to find.

a. Find a condition on the function \(\tilde{E}(y)\) so that the electric field obeys the wave equation, as it must inside the waveguide. Find boundary conditions on \(\tilde{E}(y)\) that ensure that the electric field obeys \(\tilde{E}_z = 0\) on the walls of the cavity. This is a familiar math problem with a series of solutions; find the simplest one (other than \(\tilde{E} = 0\)).

b. Use Faraday’s law to find the magnetic field corresponding to your solution in (a). It must satisfy \(\tilde{B}_z = 0\) on the walls of the cavity, so impose that condition.

c. Once you have found your solution, examine its dispersion relation (relation between \(\omega\) and \(k\)). In empty space, there are always solutions for any value of \(\omega\); is that the case...
here? Show that all solutions of this form have **group velocity** smaller than $c$. (The
group velocity is defined as $d\omega/dk$, in contrast to **phase velocity**, which is just $\omega/k$.)
[[Vista: Johnson noise]]
CHAPTER 20
First Look at Energy and Momentum
Transport by Waves

“The so-called ‘electromagnetic theory of light’… is rather a backward step…. The one thing about it that seems intelligible to me, I do not think is admissible… that there should be an electric displacement perpendicular to the line of propagation.”

— Kelvin, who never did accept it, in 1904

20.1 FRAMING: PRESSURE

Sound and water waves transport energy: Sound can actuate those tiny bones in your inner ear; the tsunami brings the earthquake to your shores. Also, we have seen that

- EM fields store energy, and
- The field equations have traveling wave solutions.

So it’s not surprising that EM waves can also transport energy, though the details are significantly different from the fluid-mechanics cases. Eventually Chapter 35 will make a general framework for studying this claim, but first let’s do some simple calculations in a concrete situation. Along the way, we’ll see that light also does some completely new things: It also transports linear momentum (and angular momentum).

Electromagnetic phenomenon: The expansion of the early Universe was faster than predicted from gas pressure alone.

Physical idea: The electric fields in a wave induce transverse, oscillatory motion on charges, which in turn gives rise to a longitudinal magnetic force that does not average to zero.

20.2 LINEAR POLARIZATION

20.2.1 Electromagnetic waves transport energy

As in Chapter 18, make the useful abbreviation

\[ \Phi_{k,\omega}(t, \vec{r}) = e^{i(\vec{k} \cdot \vec{r} - \omega t)}, \]

and consider a solution to the Maxwell equations that propagates along the +\( \hat{z} \) direction and is linearly polarized along \( \hat{\epsilon} \):\)

\[ \vec{E}_x = \frac{1}{2} i \omega \Phi_{k,\omega} \hat{\epsilon} + \text{c.c.,} \quad \vec{B}_y = \frac{1}{2} ik \Phi_{k,\omega} \hat{\epsilon} + \text{c.c.} \quad (20.1) \]

\(^1\)See Section 18.7 (page 278).
Here $\vec{E}$ is a real scalar constant and the other six components are all zero.

Suppose that this wave travels through empty space, then impinges on a test particle with charge $q$ and mass $m$. The particle is constrained to move only in the $xy$ plane, that is, the plane $z = 0$; we will denote its trajectory by $\vec{r}_\perp(t)$. We assume that within that plane, its motion is damped by viscous friction with coefficient $\eta$. That is, it feels a friction force $-\eta(\vec{d}\vec{r}_\perp/\vec{dt})$.

In the limit of strong friction, we may neglect inertia in Newton’s law of motion and the value of $m$ becomes irrelevant:

$$\vec{0} = -\eta \frac{d\vec{r}_\perp}{dt} + q\left(\vec{E} + \frac{d\vec{r}_\perp}{dt} \times \vec{B}\right)_\perp.\quad(20.2)$$

The last term on the right equals zero, because $d\vec{r}_\perp/\vec{dt}$ and $\vec{B}$ both lie in the $xy$ plane, so their cross product has no component in that plane. Thus,

$$\frac{d\vec{r}_\perp}{dt} = \frac{q\vec{E}}{\eta}.\quad(20.3)$$

We can now find the rate at which the field does work on the particle. Because the particle is constrained to move only in the $xy$ plane, and we assumed $\vec{E}$ is real,

$$\mathcal{P} = \vec{f}_\perp \cdot \frac{d\vec{r}_\perp}{dt} = q^2||\vec{E}||^2/\eta\quad(20.2)$$

$$= \frac{q^2\omega^2E^2}{4\eta} \left|\left|i\vec{e}^{-i\omega t} - i\vec{e}^{+i\omega t}\right|\right|^2 = \frac{q^2\omega^2E^2}{\eta} \left(\text{Im} e^{-i\omega t}\right)^2.\quad(20.3)$$

This quantity is always greater than or equal to zero. Its time average involves the average of $\sin^2(\omega t)$, which is $1/2$, so

$$\langle\mathcal{P}\rangle = \frac{q^2\omega^2E^2}{2\eta}.\quad(20.4)$$

**Your Turn 20A**

a. Check that the units in this formula (and every formula) make sense.

b. Also, redo this derivation for the more general case in which $\eta$ is not so huge, so that we must also account for the inertial term $m(d^2\vec{r}_\perp/d\vec{t}^2)$ in Newton’s law. Check that the limits $m \to 0$ and $\eta \to \infty$ holding frequency fixed work the way you expect.

So far, our result is not very surprising: Like any wave, an EM wave carries energy proportional to its amplitude squared. The charged particle can extract some of that energy, roughly as a cork floating on water extracts kinetic energy from passing waves.

**20.2.2 Although momentum is a vector, its transport in a wave does not time-average to zero**

Even though we assumed our particle was constrained to move only in the $xy$ plane, still it can feel forces in every direction. You might expect that because force is a vector,
unlike energy, such forces would average out to zero. Indeed the electric force, which is directed along $\pm \hat{x}$, does follow that expectation. But a moving particle will also experience a magnetic force directed along $\hat{k}$:

$$\vec{f}_\parallel = q \left( \frac{d\vec{r}}{dt} \times \vec{B} \right)_\parallel = q \left( \frac{q \vec{E}}{\eta} \times \vec{B} \right)_\parallel. \quad (20.5)$$

Substitute Equation 20.1:

$$\vec{f}_\parallel = \frac{q^2 \omega k}{\eta} \frac{1}{4} \left( (\chi \hat{x} e^{-i\omega t} + \text{c.c.}) \times (\gamma \hat{y} e^{-i\omega t} + \text{c.c.}) \right)_\parallel$$

$$= -\frac{q^2 \omega k E^2}{4\eta} (\text{c.c.})^2 = \frac{q^2 \omega k E^2}{\eta} (\text{Im} e^{-i\omega t})^2. \quad (20.6)$$

The time average is then

$$\langle \vec{f}_\parallel \rangle = \frac{q^2 \omega k E^2}{2\eta}. \quad (20.7)$$

Recall that force is the rate of momentum transfer. So the wave continually transfers momentum to the particle, or in other words the particle continually extracts momentum from the wave.

**Your Turn 20B**

As before, generalize the calculation to include the inertia term, and check the limits $m \to 0$ and $\eta \to \infty$ for reasonableness.

Our result has no counterpart with, say, sound waves: Sound in air involves pressure variation. It can *shake* things along its direction of propagation, but gives no net push. Even sound in, say, steel, which can have transverse polarizations, only shakes things. In contrast, Equation 20.7 is net momentum transport.

### 20.2.3 Radiation pressure underpins many electromagnetic phenomena

J. Poynting predicted the phenomenon of radiation pressure in 1884, and independently O. Heaviside a bit later. P Lebedev, and independently E. Nichols and G. Hull, detected its effect on macroscopic objects and absorbing gases in 1901.

Our derivation still suffers from the same critique as in the preceding section: We see that the wave carries momentum, but we don’t yet know how much. All we found was how much momentum one particular system can extract.\(^2\)

But just knowing that light can transport momentum, and that the delivered momentum is in the direction of its propagation, already gives us a lot of physics payoff:

- This “radiation pressure” phenomenon underlies the observation that a comet’s dust tail always streams away from the comet in the direction away from the Sun.
- At the earliest times after the Big Bang, radiation pressure dominated over the gas pressure of ordinary matter, so it is crucial for cosmology.\(^3\)

\(^2\)Chapter 35 will do a more systematic job, at the expense of more abstraction.

\(^3\)Section 37.4 will give a quantitative formula.
• It also supplements ordinary gas pressure in stars, opposing gravitational collapse (until the nuclear fuel is exhausted).
• It allows exquisitely fine manipulation of micrometer-size objects via optical tweezers.  
• One day it may even provide a tiny but inexhaustible source of impulse for “solar sail” spacecraft.

20.3 LIGHT CANNOT BE INTERPRETED AS A STREAM OF NEWTONIAN PARTICLES

Although we haven’t found the absolute energy or momentum content of a wave, something interesting comes up if we divide the results of the two preceding sections:

\[
\frac{\text{rate of energy extraction}}{\text{rate of momentum extraction}} = c.\] \hspace{1cm} (20.8)

Everything specific to our silly little imagined system (amplitude \(E\), charge \(q\), friction coefficient) cancels out of this universal ratio.

**Your Turn 20C**

Confirm that the particle mass \(m\), which you added in Your Turns 20A–20B, also drops out.

So it’s plausible that this result will have far greater generality, and will continue to apply to all the energy and momentum carried by a plane wave.

This result gains further significance in the quantum theory of light. That is a dual picture of light as a stream of particles, each with energy \(E = h\omega\). Our charged particle intercepts and absorbs some of them at a rate \(r\). That rate cancels from Equation 20.8, which then implies that each particle of light must also carry linear momentum \(p = h\omega/c\), or

\[
E = pc.\] \hspace{1cm} (20.9)

That result sounds paradoxical: Newtonian mechanics instead says that \(E = p^2/(2m) = pv/2\), but Equation 20.9 is missing the factor of 1/2. Chapter 31 will give Einstein’s resolution to this apparent contradiction.

20.4 CIRCULAR AND ELLIPTICAL POLARIZATIONS

Section 18.10.1 (page 284) showed that there are plane waves in which the electric and magnetic fields twirl around the axis of propagation, instead of shaking along a fixed direction. We can study them by dropping the assumption that our wave is linearly polarized along \(\hat{x}\). That is, let \(\vec{E}\) be any complex vector satisfying \(\vec{E} \cdot \vec{k} = 0\).
Your Turn 20D

a. Start from Equation 20.2 and find the analog of Equation 20.4 in this situation. 
   [Hint: This time, the charged particle will execute uniform circular motion in the xy plane.]
b. Start from Equation 20.5 and find the analog of Equation 20.6.
c. Is Equation 20.8 still true in this more general situation?

For the case of real polarization vector, Equations 20.3 and 20.6 showed that the power and force transmitted to a particle by a linearly polarized wave fluctuate (though they don’t change sign). Now, in contrast:

Your Turn 20E

a. Show that on the contrary, if the wave is circularly polarized then the power and axial force are both constant in time.
b. Show that elliptical polarization gives something in between those extremes.

20.5 ELECTROMAGNETIC WAVES CAN ALSO TRANSPORT ANGULAR MOMENTUM

You found in Your Turn 20D that for circular polarization, a charged particle confined to the transverse plane will execute uniform circular motion, in a direction determined by the wave’s helicity. That motion implies a torque to overcome the friction, or in other words the transfer of angular momentum from the wave to the particle (which in turn is coupled by friction to the surrounding fluid that we imagined). You’ll work out details in Problem 20.2, along the way learning something more about photons.

FURTHER READING

Intermediate:
It was important that we chose to include damping in Section 20.2: See Rothman & Boughn, 2009 for why an isolated, free electron won’t extract energy or momentum from a beam of light.

Technical:
Historical: Poynting also predicted angular momentum of EM fields. Experimental discovery: Beth, 1935; Beth, 1936.
20.2.2’ [Ponderomotive force and the Paul trap]
[Not ready]
20.1 *Radiation pressure*

“Yuri Milner, a Russian physicist and billionaire investor, announced a plan to develop the technologies that interstellar flight would need. Mr. Milner is devoting himself to the challenges of deep space… He is going to spend $100m on a “Breakthrough Starshot” research programme.” – *The Economist*

Sounds crazy, but for $100m maybe we should investigate.

Milner’s idea is to power a tiny spacecraft—with mass just five grams—by radiation pressure from a huge laser based on Earth. The *Economist* makes it all clear by stating that “A gigawatt laser beam—roughly the power output of a large nuclear plant—provides a force equivalent to that required to lift a glass of beer.”

a. Estimate the attainable force and see if the *Economist* got it right. If that last quote is not precisely phrased (for example, if it’s missing some other parameter describing the spacecraft or laser), choose some parameter value(s) that seem reasonable to you and that allow a precise statement.

b. Milner’s plan involves illuminating a reflector on the tiny spacecraft for ten minutes. The spacecraft is launched from outside Earth’s atmosphere (no air resistance). With the acceleration corresponding to the force you found in (a), how fast would the spacecraft be flying at the end of ten minutes?

20.2 *Angular momentum transport*

Suppose that a plane, circularly polarized electromagnetic wave of angular frequency $\omega$ travels along the $+\hat{z}$ direction.

a. Write the electric and magnetic fields analogous to Equations 20.1, again parameterized by a single real constant $E$ with appropriate dimensions.

The wave encounters a point charge $q$. Again, the charge is free to move in the $xy$ plane. There is friction slowing it down; assume that its equation of motion is

$$m(d^2\vec{r}_1/dt^2) = -\eta(d\vec{r}_1/dt) + \text{(Lorentz force)}.$$ 

As in Section 20.2.1, neglect any radiation by the charge, and also neglect the left-hand side of the above formula (suppose that it’s negligible compared to either term on the right).

b. Find the late-time solution to the equation of motion for the charge; that is, the motion after any initial transient has died out. Your formula will involve the amplitude, $\eta$, $q$, and possibly other constants.

c. The Lorentz force does work against friction. Let $\mathcal{P}$ be the rate at which it does this work, averaged over a cycle. Find $\mathcal{P}$.

d. The wave also pushes the charge in the $xy$ plane, exerting a *torque* $\tau_z$. Find the average of this torque over a cycle.

e. The ratio $(\tau_z)/\mathcal{P}$ has a remarkably simple form: Find it in terms of the parameters in the problem.
f. Following Section 20.3, momentarily unlock the quantum part of your brain and reinterpret your answer (e) in terms of a stream of little packets, each carrying a lump of energy \( \mathcal{E}_s \) and a lump of angular momentum \( L_s \). That is, interpret your answers to (c,d) as saying that the charge absorbs some of these lumps; then make a statement about the relation between \( \mathcal{E}_s \) and \( L_s \) using your result in (e). Draw a conclusion about the intrinsic angular momentum carried by one packet.
CHAPTER 21

Geometrical Optics and the Eikonal

If you love nature, you respond to her phenomena as naturally as you breathe…. The habit of observing refines our sense of beauty and adds a brighter hue to the richly colored background against which each particular fact is outlined.

— Marcel Minnaert

21.1 FRAMING: LOCALLY-PLANE WAVES

Real-world problems are mathematically harder than the idealized problems we encounter in our first textbooks. We may need some sort of unfair advantage before we can make a dent in such a problem. Often such an advantage comes in the form of a limiting case; for example, some quantity may be numerically small in cases of interest.

This chapter will study the propagation of light in media that, while not uniform, at least vary in limiting ways:

• We’ll begin with piecewise-uniform media that meet at sharp, planar boundaries (or nonplanar boundaries that are nearly flat on the length scale of wavelength). In this situation, we can join together simple solutions in each medium to get an overall solution.

Figure 21.1: [Saul Steinberg.]
We’ll also study the opposite limit of media whose properties vary continuously but gradually (again on length scales much bigger than the wavelength of the light under consideration). Here a class of approximate solutions, which we might call locally-plane waves, will be very useful.

Such situations arise in many practical problems; the approximation scheme we will develop will let us approach otherwise forbiddingly complex situations.

For example, in the situations just mentioned, light seems to travel along “rays” that are generally straight lines—except when the light gets reflected or refracted. No concept of rays appears explicitly in the Maxwell equations, however. What, then, is a “ray?” This chapter will explore that question. Along the way, we will encounter a form of focusing called “caustics.”

Electromagnetic phenomenon: Light generically focuses into networks of bright lines against a diffuse background when it encounters curved interfaces.

Physical idea: Irregular media create caustic surfaces. (Special arrangements, including lenses, can create approximately point focus.)

### 21.2 LIGHT HAS PLANE WAVE SOLUTIONS IN A UNIFORM MEDIUM

Consider a uniform, isotropic dielectric medium. Following Chapter 6, we will assume that the medium can be summarized simply by using an effective permittivity $\varepsilon \neq \varepsilon_0$. The assumption of isotropy means that $\varepsilon$ is a scalar. Section 18.11.2 (page 285) argued that there will be transverse wave solutions with dispersion relation $\omega = \left( \frac{c}{n} \right) |k|$, where the refractive index $n = \sqrt{\varepsilon / \varepsilon_0}$. For dielectric materials, it is larger than 1. However, the index may not be a constant: If it depends on frequency, we will say that the medium is dispersive.

### 21.3 PIECEWISE-UNIFORM MEDIUM

#### 21.3.1 The law of reflection

Before considering a dielectric medium, however, first suppose that an electromagnetic plane wave in vacuum encounters the flat surface of a stationary conductor, for example, a block of polished aluminum (Figure 21.2).

Throughout this chapter we will neglect the possibility of dissipation; for example, here the conductivity is supposed to be infinite. Hence, there can be no electric field inside the conductor. To fulfill that conditions, its charges must move in a way that creates a new field that cancels the incoming one inside the conductor. That sounds complicated, but outside the conductor we may guess that the new field will also be a plane wave, and simply apply the boundary condition.

---

1Chapter 49 will justify this prescription in more detail. For simplicity, we will also assume that $\mu = \mu_0$, but similar formulas ensue if that’s not the case.

2Chapter 49 will describe this phenomenon further.
Figure 21.2: Reflection of a plane wave. Crests (solid lines) of the incoming wave must coincide with troughs (dashed lines) of the reflected wave at a conducting surface (heavy line). Hence, their spacing at the surface must be equal. But also the perpendicular spacing between crests, or wavelength, is set by the frequency and so must also agree. The only way to satisfy both of those requirements is for angle of incidence $\theta$ to equal angle of reflection $\theta'$. In particular, the peaks of the incoming wave’s electric field must align with the troughs of the new one at the surface. Figure 21.2 shows that therefore, the new wave’s angle with the perpendicular (angle of reflection) must equal that of the old one (angle of incidence)—the familiar law of reflection.\footnote{A moving mirror, however, does not obey the usual reflection “law” (Section 28.2’a, page 432).} The amplitude of the outgoing wave must also match that of the incoming. Hence, the energy fluxes of both waves also match.\footnote{See Section 20.2.1 (page 299).} By energy conservation, we conclude that no net energy is deposited in the conductor.

A similar graphical analysis also applies to the reflected part of a plane wave that encounters a dielectric medium, such as glass or water. This time, some of the wave penetrates the medium. More detailed analysis of the boundary condition is therefore needed to find the fraction of energy flux that is reflected; that fraction can depend on the incoming wave’s polarization, the nature of the dielectric, and the angle of incidence.\footnote{Section 49.4.1 (page 695) and Problem 49.4 (page 703) will pick up this story, introducing a “reflection factor” $\mathcal{R}$ and expressing the reflected fraction as $|\mathcal{R}|^2$.} Apart from the amplitude, however, the form of the reflected wave is still determined by matching, and hence still follows the reflection law.

21.3.2 The law of refraction arises from matching fields across a planar boundary

Next, again suppose that a uniform dielectric medium, called $\mathcal{D}$, has a sharp, planar boundary with vacuum on the other side. (A planar junction between two different media can be handled similarly.) We’ll restrict to the simple case where the medium and its boundary are not changing in time (think about a stationary chunk of glass). Then the Maxwell equations are still linear partial differential equations with coefficients that are constant in time, so they will still have solutions with overall time dependence everywhere $e^{-i\omega t}$. The coefficients are not spatially constant, due to the boundary, so we can’t expect solutions with a single overall $e^{i(k\cdot x)}$. Separately on each side of the boundary, however, there are solutions of this form. So consider a trial solution with transverse plane waves on either side of the boundary, and with locally constant wavevectors $\vec{k}$ and $\vec{k}'$.

Figure 21.3a illustrates the situation when $\vec{k}$ is perpendicular to the interface. The horizontal lines represent planes of constant phase. For example, these lines could represent$\ldots$
Figure 21.3: Refraction. A plane wave passes through a planar interface. Lines represent the planes of constant phase for the incident- and transmitted-wave parts of the solution. (A reflected-wave part is also present but not shown.)

(a) Perpendicular incidence \( \theta = \theta' = 0 \). The spacing between the wavefronts (wavelength) changes upon exit from the medium. (b) The angle of incidence \( \theta \) is nonzero but less than the critical value. The light’s direction bends farther from the perpendicular as it exits the medium. (c) No solution is possible when the angle \( \theta \) is too large. A wave that is coming from within the medium (traveling upward) as shown cannot escape as a plane wave; instead it will be totally internally reflected.

Figure 21.3b shows a more general situation. Faraday’s law shows that the parallel components of electric field must be continuous across the interface.\(^6\) Consider in particular the loci where \( \vec{E}_\parallel = 0 \). These are parallel planes on either side of the boundary, and at the boundary they must match up. The only way for that condition to be consistent with different wavefront spacing is for the wavevector to change direction, as shown. This phenomenon is called refraction. Define the angle of incidence \( \theta \), and the angle that \( \vec{k}' \) makes with the perpendicular (angle of refraction) as \( \theta' \).

Your Turn 21A

a. Explain geometrically why the perpendiculars to the wavefronts must change at a vacuum–medium interface according to the law of refraction:\(^7\)

\[
\sin \theta' = n \sin \theta \quad \text{isotropic medium/vacuum} \tag{21.1}
\]

b. How does this formula change for an interface between two dielectric media?

As in Section 21.3.1, a caveat is needed: Although we have constrained the angle of the transmitted wave, its amplitude and polarization will depend on the refractive index, angle of incidence, and incoming polarization. Again we postpone discussion of such effects to Chapters 49 and 52.

Figure 21.4a shows one familiar consequence of refraction. It also illustrates a less

---

\(^6\) Section 18.4.2 (page 274). It is true that the component perpendicular to the interface may change discontinuously as we cross it, but the continuity of parallel components is enough for present purposes.

\(^7\) Chapter 53 will point out a subtlety in the interpretation of this result for anisotropic media.
21.3 Piecewise-uniform Medium

Figure 21.4: **Refraction versus total internal reflection.** (a) Light reflected from the fish’s head travels in every direction along rays (red). The observer on the bridge accurately concludes that the fish’s head is at 1. The observer on the bank, unconsciously assuming that light travels on unbent lines, gets an inaccurate impression (the head seems to be at 2). (b) Viewed from below, part of the air-water interface appears to be a mirror (circled)—an instance of total internal reflection.

cluttered visual representation: Instead of drawing all the wavefronts, the lines in the figure show “rays,” which we provisionally define as piecewise-straight lines that are everywhere parallel to \( \mathbf{k} \).

21.3.3 Geometrical optics extends these exact results to approximate rules for non-planar boundaries

We can now summarize Sections 21.3.1–21.3.2 by saying that

*In a piecewise-homogeneous medium, initially parallel rays remain parallel, but they change direction at a planar boundary following the laws of reflection and refraction.* 

Problems involving light in such situations therefore turn into geometry problems, as long as we do not need to predict amplitudes nor polarization effects.

Non-flat interfaces are more complicated, but in many practical situations a simplification is possible: Typical lenses are nearly flat on the scale of visible light’s wavelength. So we can imagine a narrow part of a wavefront striking a nearly-planar surface element and following nearly the usual laws of reflection and refraction appropriate for that element’s orientation, an approximate approach we’ll provisionally call “geometrical optics”.

For a curved interface, initially parallel rays will encounter differently-oriented boundary surface elements, and hence may emerge nonparallel to each other. Section 21.5 will generalize this idea, but first we will look at some concrete examples.

---

8See Section 21.5.3 for a more general definition. Some authors use the synonym “ray optics.”
21.3.4 Optical tweezers exploit the momentum transfer implied by refraction

Experience with water waves suggests that even if the wavefronts are all parallel, they need not be infinite in extent. For example, waves arriving at a gap in a breakwater pass beyond it in a way that stays localized, at least for some distance. Chapter 41 will pursue this analogy mathematically, but it should be reasonable that light, too, can travel approximately as a “beam” of finite width. We can represent the situation with rays if we make the convention that their transverse density represents the intensity of light; thus, there are no rays outside the region occupied by the beam.

Figure 21.5 uses such imagery to sketch how a spherical object with refractive index different from its surroundings will feel a net sideways force when it encounters a beam of light. If the refractive index of the sphere is larger than that of the surroundings, as shown, then the effect is to pull it toward the center of the beam. This phenomenon is useful for manipulation of micrometer-scale objects (and of nanometer-scale objects that we may tether to them): the optical tweezers effect.

21.3.5 Spherical aberration limits the focusing power of simple lenses

The law of refraction also explains the focusing of light by a lens. Figure 21.6a shows an incoming plane wave, represented by a bundle of parallel rays, that impinge on a glass sphere. If the sphere’s diameter is much bigger than the wavelength of the light, then we may apply the law of refraction separately to each of the rays shown. We start with a plane wave on the left (a collection of parallel rays), convert each ray to a new direction upon entering the medium via the law of refraction, extend the resulting rays until they again hit the interface on the right-hand side of the figure, and again apply the law of refraction there. As shown in the figure, the ray passing through the center of the sphere is undeflected, but flanking rays are bent more and more, which tends to bring them to a common point, or focus. The incoming plane wave carries uniform energy flux

---

9 Chapter 53 will consider the momentum transported by light in more detail.
10 Although the static dielectric constant of water is enormous (Section 6.8, page 87), its value at optical frequencies is much closer to ε₀, and typically is exceeded by that of glass or plastic beads (Appendix C).
21.3.6 Dispersion from a triangular prism

Section 21.2 introduced the concept of dispersion: Because different vacuum wavelengths travel at slightly different speeds in a transparent medium, the bending predicted by the law of refraction is wavelength-dependent. Figure 21.7 shows the resulting behavior in a macroscopic setting. A collection of parallel rays with different frequencies gets spread in angle upon entering a prism, then spreads further upon exiting. Rays with different spectral colors exit at different characteristic angles relative to the incoming rays.
21.3.7 Chromatic aberration

Although the effect of dispersion may be pretty to look at, it’s not always desirable. We have seen how the curved surface of a lens bends each ray in a parallel bundle by a different amount, in such a way that many of them arrive at a common focus. But the location in space of the focus depends on the index of refraction as well as the shape of the lens, so dispersion implies that each wavelength will have a different focal point. This chromatic aberration effect degrades the ability of a microscope to make images with white-light illumination. If monochromatic illumination is not adequate, then the lens material must be chosen to be minimally dispersive in the visible range. Alternatively, carefully designed compound lenses can combine elements with partially cancelling dispersion.

21.3.8 Total internal reflection arises when the law of refraction cannot be satisfied

Figure 21.3c shows geometrically that there will be no solution to Equation 21.1 (page 310) if the angle of incidence exceeds a critical value. In terms of your result from Your Turn 21A, sin \( \theta \) must be smaller than \( 1/n \) because \( \sin \theta' \) cannot exceed 1. If a plane wave originates in the medium (heading toward the vacuum side) but this condition is violated, then the Maxwell equations and their boundary conditions do not allow any transmitted plane wave. Instead, all incoming energy reflects back into the medium, a phenomenon called total internal reflection (TIR, Figure 21.4b).

---

11The eyelenses of humans and other animals also suffer from chromatic aberration. Our eyes acknowledge the impossibility of simultaneously focusing all colors of visible light by deploying a much lower-resolution pixel grid for blue than for red and green photoreceptors.
Your Turn 21B

a. Imagine yourself submerged in a swimming pool. Looking straight upward, you see sky. But beyond a certain angle, the surface above you looks like a mirror (try it!). Why?

b. Is there also a critical angle if a plane wave instead originates on the vacuum side (k directed toward the medium)?

TIR is the basis for guiding light through glass fibers. As long as the fiber does not bend too sharply, a beam that initially propagates axially will internally reflect off the boundaries, remaining trapped inside the fiber. Such a fiber can have much greater data bandwidth than a coaxial cable because the frequency of visible light is so much higher than the radio frequencies that the coax can carry. Also, a flexible bundle of such fibers can carry each pixel of a complete input image faithfully to the same relative position at its output end, regardless of overall bends along the way. Such fiber-optic endoscopes are indispensable for minimally invasive medical diagnosis.

When we go beyond geometrical optics, we’ll find that TIR is not quite total after all: Chapter 49 will discuss how field disturbances penetrate a short distance into the second medium, even at high angle of incidence. This “evanescent field” phenomenon is the basis for an important microscopy technique.

21.3.9 Interfering-ray optics: Michelson interferometer

Section 21.3.1 mentioned that light impinging on an interface may partly enter it while also partly reflecting. In ray language, we therefore say that each incoming ray can split. Instead of a simple interface, a thin, uniform layer of metal can be evaporated onto glass. By varying the thickness of the layer, one can create a surface that reflects any desired fraction of incoming light. If that fraction is 50%, we call the surface a half-silvered mirror. Later, we will study an instrument that exploits such an optical element, called the Michelson interferometer.

The two resulting rays can be arranged to recombine elsewhere. As we mechanically change the path length in one arm, the illumination from the combined beams periodically transitions between a maximum (constructive interference) and zero (destructive interference) with each half wavelength. Thus, even though optical frequencies are too high for instruments to detect wave motion directly, we can nevertheless observe wave behavior.

A hybrid form of ray optics can explain such phenomena by ascribing to each ray a phase, proportional to (geometrical path length) × (nω/c), where the refractive index n equals 1 in vacuum. Each of the two split rays maintains phase information as it travels

---

12 This primitive description is appropriate for thick fibers. Modern fiber-optic lines are thin and function more like waveguides (Problem 19.1); their composition is also modulated across their cross-section. Some even transmit light in the form of nonlinear traveling waves (solitons).

13 See Section 49.4, page 695.

14 See Figure 37.1 (page 578) and Section 28.2.a (page 432).
through space or dielectric media and reflects off mirrors and other interfaces. If their transit times along their respective paths are not too different, then the waves represented by the rays can interfere with each other, like water or sound waves.\textsuperscript{15} Their complex amplitudes are simply added at the recombination point. The illumination intensity is proportional to the absolute square of the complex amplitude, so it can periodically drop to zero, as is observed. We will call this extension of geometrical optics \textit{interfering-ray optics}, to emphasize that it still idealizes light as rays.\textsuperscript{16}

\section*{21.4 RAINBOWS AND OTHER CAUSTICS}

\subsection*{21.4.1 Millennia of confusion}

If any natural phenomena can be called “intrinsically beautiful,” surely the rainbow is an example (Figure 21.8). Throughout human history it has inspired awe, playing key roles in art, literature, and even religious scriptures.\textsuperscript{17} And yet, for millennia the rainbow was also a goad, a rebuke to our attempts to understand Nature. Practically every natural philosopher of any note had something to say about rainbows, yet very little was understood until René Descartes published an essay titled \textit{La Dioptrique} in 1637.\textsuperscript{18} Today we see the rainbow as an exemplar of a much broader class of optical phenomena called “caustics.”

\subsection*{21.4.2 Natural versus contrived focusing}

Section 21.3.5 discussed the focusing of light by a lens. Focusing to a point requires the lens to have the right shape and distance to a focusing screen. In contrast, another form of focusing arises naturally and is ubiquitous. For example, sunlight that arrives at the bottom of a swimming pool gets concentrated into a network of bright \textit{lines} (not points). The constantly changing shape of the water surface due to wind does not destroy these lines, but merely moves them. A partial focus of this sort is generically called a \textit{caustic}; point focus is just one special case. Generically, the caustic itself occupies a 2D surface in 3D space; when intercepted by a viewing screen, it therefore creates a bright line of illumination.

The next section explores how a natural rainbow involves a particular kind of caustic.
21.4 Rainbows and Other Caustics

Figure 21.8: Natural rainbow. Note (1) the sky appears darker between the two bows; (2) the secondary bow is less bright than the primary; (3) the order of colors is opposite in the two bows (see Problem 21.3); (4) the main bow’s color sequence has some repetition (here green and violet are repeated). The faint, extra last bands are called “supernumerary” bows; see Section 21.4.5. The photo was taken with a polarizing filter; this does not affect the rainbow, which is already polarized, but it does reduce background light from the sky. In the inset, color saturation was digitally enhanced to bring out the structure. However, all features in this photo are visible with the naked eye under favorable atmospheric conditions. [Photo courtesy Steve Nelson (Fayfoto, Boston MA).]

21.4.3 Rainbows as caustics

Rainbows are always seen with the Sun behind the observer. Eventually, early scientists realized that the light we see must be sunlight scattered backward to us from water droplets in the sky. Figure 21.9 shows how this is possible: Light rays are shown entering a droplet, then partially reflecting inside it. Much of the incoming light exits the droplet at point \( Q \), continuing away from the observer, but some will be internally reflected, exit at \( R \), and be directed backward. The overall change in direction is called the scattering angle \( \gamma \). As with the prism, two rounds of refractive bending are involved, and each can spread

---

13 If the light source is monochromatic and coherent (for example, a laser), then interference will be observable even if the transit times are not similar. Even partially coherent light, however, can display interference if the transit times in the two arms of an interferometer are equalized by inserting a “compensator plate” into one of them.

14 More generally, the term “interference” refers to the effect of combining two or some small number of rays; continuous combinations of infinitely many contributions will lead in Chapter 22 to new phenomena.

17 Awe and fear: Many folk traditions attribute grim consequences to pointing at a rainbow.

18 Thomas Harriot performed similar calculations 30 years earlier, but apparently never published them.
Incoming white light into its constituent wavelengths. Theodoric of Freiberg, and independently his Arab contemporaries Kamal al-Din al-Farisi and Qutb al-Din al-Shirazi, understood the main points of the diagram around 1304. Theodoric also confirmed his idea with an experiment in which a spherical flask of water was illuminated by a shaft of sunlight in an otherwise dark room.

Closer inspection of Figure 21.9 reveals a problem with this explanation, however. A triangular prism will indeed send incoming, monochromatic, parallel rays to outgoing parallel rays (Figure 21.7). The dispersion of white light then creates its rainbow. But for the case of a spherical water droplet in air, outgoing ray directions depend both on wavelength and also on where the ray enters the sphere, that is, on which ray we are following (Figure 21.9). Hence, each individual wavelength arrives at a distant projection screen spread over a broad region and different wavelengths are all mixed. How, then, can we get a rainbow? Won’t all the spectral colors overlap, and hence appear white? Here is where Descartes made his decisive contribution, more than three hundred years after

---

19Even earlier, Roger Bacon had measured the rainbow’s caustic angle in 1266.
21.4 Rainbows and Other Caustics

Figure 21.10: Rainbow caustic. (a) Incoming parallel, monochromatic rays (dashed lines) refract as light enters a rain droplet. Some of the light reflects from the rear surface and proceeds backward (solid lines), refracting again as some of it exits the droplet. As we move from the more central rays out to those that entered near the periphery, the scattering angle decreases, arrives at a minimum (heavy blue line), and then begins to increase again. (b) Total scattering angle $\gamma$ versus entry point $\theta$, showing the minimum for the case of water (Problem 21.2). The minimum depends only on the refractive index; droplet size is immaterial.

Figure 21.11: Laboratory rainbow. White light was shone onto a single glass sphere of diameter 3 cm (lower left) through an aperture. Some of the light entering the sphere was scattered backward (into the page), where it was observed on a screen 1 m away. The scattered light diffusely fills a disk-shaped region, with bright, colored rings at its edge and darkness outside. See also Media 14. [Photo by Jax Gottschalch.]

Theodoric.

Descartes showed that although incoming parallel rays will emerge in many directions, nevertheless there is a minimum scattering angle $\gamma_s$ (Figure 21.10). That is, in the geometrical optics approximation no light will be returned at angles less than this minimum.\(^\text{20}\)

\(^\text{20}\)Larger angles, even beyond 180 deg, are however possible (Problem 21.3).
Figure 21.12: [Cross-sectional diagram.] Where to look. An observer standing with the Sun at their back sees bright caustics from many rain droplets (including 1, 2, 3) located on a cone (dashed lines). Typically, however, there are many more droplets on the upper part of this cone because the Earth stands in the way of the lower lines of sight (3). There is also some illumination from droplets located inside the cone (4, dotted line), leading to a diffuse wash of light inside the ring, but not from those outside it (5). Superimposed on the rainbow, the observer also sees light reflected from the front surfaces of the droplets, as well as ordinary sky light from behind them. See also Media 15.

In Problem 21.2, you’ll see that

- The value of $\gamma_*$ does not depend on droplet size, but it does depend on the wavelength of light if the medium is dispersive (as indeed water is).
- Even if the incoming light is of uniform intensity, there will be a pileup at angular position $\gamma_*$, leading to the prediction of infinite light intensity in that one direction.

The first point explains why Theodoric could see a realistic rainbow with a macroscopic sphere (Figure 21.11), and why rain droplets of varying sizes all contribute to a single apparent bow in the sky (Figure 21.12). The second point stands in contrast to focusing by a lens, which only works at one special distance, and even then is only approximate (Figure 21.6).

Although Chapter 23 will argue that the predicted singularity is a failure of geometrical-optics approximation, nevertheless we do conclude that a projection screen placed behind the droplet will catch light covering a disk-shaped region, leaving the exterior of that disk dark, and moreover the edge of the disk will be especially bright. In fact, the edge is a caustic. We will refer to its minimal scattering angle $\gamma_*$ as the rainbow’s caustic angle.21 The dark outer region is often called Alexander’s dark band, in honor of Alexander of Aphrodisias, who pointed it out around 200 CE (Figures 21.8 and 21.11).

21Some authors call $\gamma_*$ the “rainbow angle,” or the “cartesian angle.”
Each droplet reflects light back in a caustic shaped like a cone, with the droplet at its apex. Even if the droplets are not all the same size, each of their cones has the same opening angle, because that angle does not depend on the droplet size.²²

We can catch part of that droplet's cone of light if our angle of view relative to the Sun's rays equals the cone's opening angle.

The locus of all droplet locations that meet this condition is also a cone, but with our eye at the apex: When we look in a direction lying on that cone, we see glints from every droplet lying along that line of sight.

Thus, droplets at many distances from the observer all contribute, so that many faint reflections become one bright ring (Figure 21.12).

The overall effect is an apparent glowing arch in the sky.

Descartes did not have the framework needed to complete his understanding of the rainbow, because he did not understand that white light is a mixture of different spectral colors. But shortly after his work, Newton pointed out that each spectral component of light will have a slightly different angular radius for its bright ring, due to dispersion in water: The ring is spatially spread out by wavelength, a bit like refraction from a prism. Unlike the prism, the weaker light sent to larger scattering angles creates a diffuse, white illumination inside the bow. Indeed, real rainbows are noticeably brighter inside the primary ring than outside it (Figures 21.8 and 21.11).

The secondary bow arises from double internal reflection

In good conditions a second bow is often visible (Figure 21.8). Remarkably, Theodoric also explained it in the 14th century as a consequence of two internal reflections. Because each reflection is only partial, the secondary bow is less bright than the primary. Because the total scattering angle is now greater than 180°, the extra bending of blue relative to red implies that the order of colors is reversed relative to the primary bow, as is also seen in Figure 21.8.²³

Additional bows are possible, from light that reflects more than twice. They are too faint and diffuse to be observable in natural rainbows, but up to 15 orders have been observed in lab experiments.

21.4.5 Thomas Young’s critique of geometrical optics results

Geometrical optics has given us a good account of several of the properties of rainbows. Conspicuously absent from the list, however, is any explanation of the supernumerary bows clearly visible in Figure 21.8. Although this phenomenon had been recorded in the 13th century, it is easy to miss if you are not looking for it. Even if you look, it is not always visible, and artistic depictions almost never include it. Perhaps due to the overwhelming authority of Isaac Newton, who never mentioned them, the supernumerary bows were

²³You'll pursue this in Problem 21.3.
dismissed\textsuperscript{24} until Thomas Young offered them as evidence for the wavelike aspect of light in 1803.

Young knew that for scattering angle $\gamma$ greater than the caustic value $\gamma_c$ (that is, inside the primary bow) there are two different incoming rays (two stationary-phase paths), with unequal lengths, both obeying the laws of reflection and refraction and emerging at the chosen angle $\gamma$.\textsuperscript{25} The difference of their path lengths depends on $\gamma$, leading to an interference pattern reminiscent of the one in the Michelson interferometer (Section 21.3.9). The pattern is slightly different for each wavelength, due both to dispersion and to the different number of wavelengths that fit in a given path length. Young proposed that such considerations, which Section 21.3.9 called interfering-ray optics, could explain the supernumerary bows.

Although Young’s idea is correct in essence, it is not a quantitative theory: It attempts to join together the geometrical optics approach, which neglects the wavelike character of light, with interference, which is inherently a wave phenomenon. Chapter 23 will outline a self-consistent calculation. In particular, interfering-ray optics still predicts infinite light intensity right at the caustic angle, contrary to observation.\textsuperscript{26} This pathology of the geometrical optics approximation is similar to the conclusion from geometrical optics that a lens can focus light down to a mathematical point. Really, diffraction limits the focus of even a perfect lens; similarly, Chapter 23 will find finite light intensity at the rainbow caustic.

Young was also unaware of a subtle relative phase shift between the two interfering rays. Luckily, this will come out automatically, without requiring special argument, in Chapter 23. Before pursuing that, the rest of this chapter will look at other phenomena that can be understood via geometrical optics.

21.5 GRADIENT-INDEX MEDIUM

21.5.1 Rays of light can be regarded as streamlines of energy flux

We can now return to a question posed in Section 21.1: What is a ray? Extending our provisional definition in Section 21.3.2,

\begin{equation}
\text{Rays of light are the streamlines of the energy flux of almost-plane waves.}^\text{27}
\end{equation}

An analogy may be useful: When water flows steadily around an obstacle, its mass flux is a vector field. The streamlines of that vector field tell us where the water is going.

Consider the streamlines of the energy flux. In a uniform, isotropic medium, Section 21.2 found plane wave solutions, for which $\vec{k}$ is a constant. Because the energy flux

\textsuperscript{24}No doubt the word “supernumerary,” meaning “unneeded,” was chosen by scientists who could not explain this undeniably real phenomenon. Some even used the phrase “spurious bow” to denigrate it.

\textsuperscript{25}You’ll find this result in Problem 21.2. More precisely, it holds for scattering $\gamma$ angles greater than the caustic value, but less than about 160 deg.

\textsuperscript{26}See Problem 21.2.

\textsuperscript{27}See Section 0.3.1 (page 7).
is always parallel to \( \hat{k} \), the streamlines of a plane wave are a family of straight, parallel lines. But this idea has wider usefulness than that one example.

Just as a given chamber can have water flowing in various ways, so too the rays in a given optical system can have various configurations, depending on what sort of light we send in. The definition makes sense in other ways as well: For example, when we focus light these streamlines converge, leading to high energy density (sometimes enough to ignite paper under a magnifying lens).

In a piecewise-uniform medium, Section 21.3.3 interpreted the law of refraction as saying that “rays bend as they pass a boundary,” a phenomenon that indeed corresponds to the behavior of a laser pointer’s beam when crossing from air to water (or the other way). The following sections will extend this geometrical-optics viewpoint for the case of continuously varying refractive index.

21.5.2 Locally-plane waves are a useful idealization when there is a separation of length scales

In a medium whose refractive index changes, but only gradually compared to the wavelength of light, it seems reasonable to look for solutions to the Maxwell equations that locally resemble plane waves, but for which \( \hat{\text{r}}_{\text{local}} \) varies slowly over space. In a moment we’ll make that notion precise, and verify our expectation that such solutions exist.

Here are some Electromagnetic Phenomena we’d like to understand:

- Radio waves that originally were sent away from Earth’s surface encounter a gradually changing medium as they enter the ionosphere. Section 21.5.6 will discuss the resulting refraction phenomenon.
- The air close to a hot road surface has nonuniform temperature and density, and hence also refractive index, leading to mirage phenomena (Section 21.6.1).
- Animal eyepieces have nonuniform refractive index to better focus light: \( n \) varies continuously from a maximum at the center to a minimum at the surface (Section 21.6.2).
- Perhaps most exotic, Einstein’s gravity theory predicts that strong gravitational fields can bend light (Section 21.6.3).

We might expect some continuous version of the law of refraction to hold in situations like these. Let’s find it.

21.5.3 The eikonal equation controls propagation of a locally-plane wave

Solving vector partial differential equations is in general difficult. But at least the refractive index profiles just mentioned are all nearly stationary, that is, unchanged under time translation. As far as time dependence is concerned, then, the Maxwell equations are linear differential equations with constant (in time) coefficients. Such equations have solutions with overall dependence on time that is exponential, so we can still assume harmonic time dependence for our solutions, as in Section 21.3.2 (page 309). Moreover,

---

28For an anisotropic medium such as calcite, we must reconsider this statement (Chapter 53).
all of the situations in the preceding list share two additional convenient features: The medium is locally isotropic, and the length scale $L_0$ over which the index varies is much greater than the wavelength of the light we wish to study, or in other words, $c/(nL_0\omega) \ll 1$. In this regime, it’s reasonable to look for approximate solutions to the Maxwell equations for which the vector potential takes the almost-plane wave form

$$\vec{A} \approx \frac{1}{2} e^{-i\omega t} \vec{\zeta}(\vec{r}) e^{i\omega S(\vec{r})/c} + \text{c.c.}$$  \hspace{1cm} (21.5)

In this expression, $S(\vec{r})$ is called the \textbf{eikonal function}, or simply “the eikonal.” For a plane wave, it would be the linear function $\text{const} + \vec{k} \cdot \vec{r}$. The other unknown function, $\vec{\zeta}(\vec{r})$, allows for the possibility that the polarization is not constant throughout space, unlike a plane wave in a uniform isotropic medium.

Mathematically, we can let $\omega = \omega_0\sigma$ and consider the limit $\sigma \to \infty$, holding $n(\vec{r})$, $\omega_0$, $\vec{\zeta}(\vec{r})$, and $\beta(\vec{r})$ fixed. We now ask whether the trial solution Equation 21.5 can be made to work, to leading order in the small parameter $1/\sigma$, by making appropriate choices for the unknown functions $\vec{\zeta}$ and $\beta$. The resulting framework is called \textbf{geometrical optics}. Geometrical optics studies light in vacuum or in dielectrics that either:

- are piecewise homogeneous with sharp boundaries between regions, which in turn have characteristic size scale much larger than the wavelength of light (as in Section 21.3), or else
- have refractive index function that varies gradually compared to the wavelength of light (as in the present section).

We will begin by working in vacuum, where we have some expectations; later, we will generalize to nonuniform (“gradient-index”) media.

Close to any point $\vec{r}_e$, we may Taylor expand $\omega\beta(\vec{r})/c \approx \text{const} + \vec{r}_e \cdot \vec{k}_{\text{local}} + \cdots$, where the “local wavevector” $\vec{k}_{\text{local}} = \frac{\omega}{c} \vec{\nabla} S|_{\vec{r}_e}$. Hence our trial solution Equation 21.5 locally resembles a plane wave. In particular, the energy flux everywhere points along $\vec{\nabla} S$. So once we establish that a particular eikonal function $S(\vec{r})$ gives a solution to the Maxwell equations, the streamlines of $\vec{\nabla} S$ will be the rays that we seek.

It’s convenient to impose Coulomb gauge:

$$0 = \frac{1}{2} e^{-i\omega t} (\vec{\nabla} \cdot \vec{\zeta} + \frac{\vec{\alpha}}{c} \vec{\nabla} \cdot \vec{S}) e^{i\omega S/c} + \text{c.c.}$$

We may drop the first term, because it is $O(\sigma^0)$ whereas the second is $O(\sigma^1)$. Thus, $0 = \vec{\zeta} \cdot \vec{k}_{\text{local}}$, just as Chapter 18 found for plane waves in vacuum.

The Maxwell equations then take the form in Equation 18.31 (page 283):

$$\frac{1}{2} e^{-i\omega t} \left( \vec{\nabla} \cdot \left( \left( \vec{\nabla} \cdot \vec{\zeta} + \frac{\vec{\alpha}}{c} (\vec{\nabla} \cdot \vec{S}) \vec{\zeta} \right) e^{i\omega S/c} \right) \right) + \text{c.c.} = -\left( \frac{\omega}{c} \right)^2 \frac{1}{2} \vec{\zeta} e^{-i\omega t + i\omega S/c} + \text{c.c.}$$

---

29 One way to describe this short-wavelength limit is to say that we are neglecting diffraction effects (Chapter 22); that is the regime where we may hope that a “ray” concept will be useful.

30 Section 18.8.2 (page 281).
Again drop the first term in parentheses on the left, because it is lower order in $\sigma$ than the other term. So

$$\frac{i\omega}{c} \left( \nabla_j S \right) \cdot \left( \nabla_j S \right) + \frac{i\omega}{c} \nabla^2 S + \left( \frac{i\omega}{c} \right)^2 \xi (\nabla S)^2 = -\left( \frac{\omega}{c} \right)^2 \xi.$$ 

The last term on the left dominates the others, so the condition for Equation 21.5 to approximately solve the Maxwell equations is

$$||\nabla S||^2 = 1.$$ 

Some simple solutions to the eikonal equation include $S(\vec{r}) = \hat{k} \cdot \vec{r}$ (plane wave) or $= ||\vec{r}||$ (spherical wave). In the former case, the streamlines of $\nabla S$ are parallel, straight lines; in the latter case, they are straight radial lines.

In principle, we’re now done with the vacuum case, but it may not be clear that we have made progress: We have approximated the Maxwell equations, which are linear, with the new PDE Equation 21.6, which is nonlinear. But we do not always need all the information in the phase function $S$. Let’s now convert our equation into a direct characterization of the rays (streamlines of $\nabla S$) themselves.\(^{31}\)

### 21.5.4 Rays in vacuum

The rays are a family of curves, each everywhere tangent to $\hat{k}_{\text{local}}$. We can write any curve in parametric form as $\bar{\xi}(s)$, where $s$ is arc length. That is, $d\bar{\xi}/ds$ is the field of unit tangent vectors along the curve (recall Equation 21.6). The tangent must be parallel to $\nabla S$, which itself is everywhere a unit vector, so

$$\frac{d\bar{\xi}}{ds} = \nabla S|_{\bar{\xi}}(s) \quad \text{for all } s.$$ 

One way to characterize a curve is to state its curvature, that is, how its tangent vector deviates from being a constant.\(^{32}\) More precisely, we define the curvature vector as the

---

\(^{31}\)Note that the polarization vector drops out of Equation 21.6, so we learn nothing about $\vec{\xi}$ from this approach other than that it must everywhere be perpendicular to $\nabla S$. To learn more, we would have to retain some of the subleading terms dropped earlier; instead we will concentrate on just the rays, and not their polarization. The transport of polarization along rays can be important for light in a liquid crystal; see Problem 52.3.

\(^{32}\)Chapter 7 introduced an equivalent notion of curvature as measuring deviation of a curve from its tangent at a point.
derivative of the unit tangent to the curve with respect to arc length, finding:

\[ \text{curvature} = \frac{d^2 \ell}{ds^2} = \frac{d}{ds} \left( \frac{\nabla S}{|\ell|} \right). \]

The right side of this formula is the derivative of a vector field as we walk along the curve. To evaluate it, we can take the dot product of the gradient operator (that is, all partial derivatives) with the unit tangent and apply to \( \frac{\nabla S}{|\ell|} \):

\[ \frac{d^2 \ell}{ds^2} = \left( \frac{d \ell}{ds} \cdot \nabla \right) \left( \frac{\nabla S}{|\ell|} \right). \]

Then Equation 21.7 gives

\[ = \left( \nabla_j S \right) \left( \nabla_j \left( \frac{\nabla_i S}{|\ell|} \right) \right) = \left( \nabla_j S \right) \left( \frac{\nabla_i (\nabla_j S)}{|\ell|} \right) \]

\[ = \frac{1}{2} \frac{\nabla_i ||\nabla S||^2}{|\ell|} = 0, \text{ by Equation 21.6.} \quad (21.8) \]

We conclude that the curvature is zero:

**Light rays in vacuum are straight lines.**

That makes sense: Geometrical optics neglects diffraction, and when that approximation holds, objects indeed cast sharp shadows. The two illustrative families of solutions found earlier (straight parallel rays and straight radial rays) both obey this rule.

Note that the eikonal function \( S \) has disappeared from Idea 21.9; we don’t need to solve the eikonal equation after all in order to find the rays.

### 21.5.5 Rays bend continuously in a gradient-index medium

Next, generalize to the case in which the local speed of light, \( c/n(\vec{r}) \), is not constant in space. (The symbol \( c \) always refers to the speed of light in vacuum.) Chapter 6 suggested that we can just replace \( \epsilon_0 \) by a permittivity function \( \epsilon(\vec{r}) \), so Ampère’s law now says

\[ \nabla^2 \vec{B} = (\mu_0 \epsilon) \frac{d^2 \vec{B}}{dt^2}. \]

**Your Turn 21C**

a. Similarly to the derivation that led to Equation 21.6, show that this time, there is an eikonal solution for \( \vec{B} \) with

\[ ||\nabla S||^2 = n^2. \quad \text{eikonal equation in medium} \quad (21.10) \]

b. Show that therefore the analog to Equation 21.7 gives the unit tangent to a ray as

\[ \frac{d \ell}{ds} = \frac{\nabla S}{|\ell|} \quad (21.11) \]
Your Turn 21D

a. Next adapt the derivation leading to Equation 21.8 to show

\[ \frac{d}{ds}\left(n(\hat{r}) \frac{d\hat{r}}{ds}\right) = \nu_n \left| \frac{d\hat{r}}{ds} \right|, \quad \text{ray equation in medium} \]  

(21.12)

at every position \( s \) along a ray.

b. Check that your result from (a) is compatible with arc length parameterization. That is, show that if \( \|d\hat{r}/ds\| \) starts out equal to one, then it retains that property as we step forward in \( s \).

As in the vacuum case, the ray equation makes no explicit mention of the eikonal function \( S \). It tells us how light rays bend as they pass through a medium—a generalization of the law of refraction. When the geometrical-optics approximation is justified, this equation reduces Maxwell’s partial differential equations to the ordinary vector differential equation (21.12). This simplification is offset by the fact that we must solve the ray equation for every desired starting point, instead of finding one global solution \( S \) and then tracing its rays.

A bit like Newton’s \( \frac{d^2r}{dt^2} = \frac{f}{m} \), we can start a ray trajectory at any point, with any initial direction of motion, and then step through the ray equation to find the subsequent path of that ray.\(^{34}\) Solving systems of ODEs numerically is a routine task.\(^{35}\)

Section 21.5.4’ (page 333) connects the ray equation to Fermat’s principle.

21.5.6 Shortwave radio skip (skywave transmission)

After G. Marconi and C. Franklin established the practicality of using radio waves to communicate with ships at sea, it was natural to want to cover greater distances. Marconi set out to transmit across the Atlantic ocean. Others scoffed: Electromagnetic rays moved on straight lines, and so even if launched parallel to the surface, they would propagate out into space as the curved Earth bent away from them. Without any scientific justification, Marconi nevertheless invested vast sums constructing huge transmitting and receiving stations, and was eventually rewarded with success in 1902. How was this possible?

Later, Heaviside deduced that there must be an ionized atmospheric layer at high altitude—a thin plasma. The dielectric constant of air is close to that of vacuum, but perhaps such variation at high altitude could bounce (“skip”) radio signals at high enough angle of incidence, similarly to the total internal reflection that can trap light in a curved optical fiber (Section 21.3.8). This hypothesis also explained why the effect was more pronounced at night\(^{36}\) and at short wavelength (see Chapter 54). Together with other improve-

---

\(^{34}\)In Newtonian mechanics, however, the initial speed of a particle is another arbitrary initial condition; in optics, each ray is parameterized by arc length (unit “speed”).

\(^{35}\)See Problems 21.4 and 21.5.

\(^{36}\)Some atmospheric ionization from solar irradiation is necessary, but too much leads to absorption during the day.
ments in receivers, these insights brought “short-wave” radio reception into the reach of thousands of nocturnal amateurs, who routinely picked up stations halfway around the Earth from their homes. Let’s investigate.

Suppose that the refractive index at radio wave frequencies depends only on one variable, the altitude. This could be the case when radio waves travel upward and encounter the Earth’s ionosphere (over distances short enough to neglect Earth’s curvature).

Initially, a ray makes an angle $\theta_0$ with respect to the vertical. Farther along on the ray, $\theta = \cos^{-1}(\hat{z} \cdot d\hat{r}/ds)$ may change. Taking the dot product of Equation 21.12 with $\hat{z}$ gives

$$\frac{d}{ds}(n \cos \theta) = \frac{dn}{dz} \big|_{\theta(0)}.$$

Multiply both sides by $n$:

$$n \frac{d}{ds}(n \cos \theta) = \frac{1}{2} \frac{dn^2}{dz}.$$

Next, note that when we move by arc length $ds$, altitude changes by $dz = ds \cos \theta$, so

$$n \cos \theta \frac{d}{dz}(n \cos \theta) = \frac{1}{2} \frac{dn^2}{dz} (n^2 \cos^2 \theta) = \frac{1}{2} \frac{d}{dz} (n^2).$$

Thus, $n^2 \cos^2 \theta - n^2$ is a constant along the ray, a generalized law of refraction:

$$n \sin \theta = \text{const} \quad \text{if } n \text{ depends only on } z. \quad (21.13)$$

This result is similar to the usual law of refraction (Your Turn 21A), which also assumed planar geometry.

More generally, the refractive index for a plasma can be smaller than one. Hence, as a ray ascends to the ionosphere, $n$ decreases from $\approx 1$ at Earth’s surface. Equation 21.13 then implies that $\theta$ will increase; if $\theta$ ever increases to $\pi/2$, then the ray can bend (“skip”) back down to Earth.

### 21.6 MORE PHENOMENA

Here are several more situations in which light travels through a medium whose index varies slowly on the length scale of the light’s wavelength.

#### 21.6.1 Mirages rely on our brains’ assumptions about light propagation

On a long, flat stretch of highway, solar heating creates a layer of air near height $z = 0$ that is hotter than elsewhere. That air is less dense than the cooler upper layers. Thus, it can happen that, when we direct our gaze downward (toward the road) we’ll see light originating from the sky that has traveled on a curved path. It is easy to misinterpret that light as a reflection from (nonexistent) water on the road, particularly because it tends to shimmer, due to air convection currents. You probably know from experience that this...
21.6 More Phenomena

Figure 21.14: Superior mirage. (a) An observer (left) sees rays arriving from various distant points. Scanning from the top (5) to the bottom (−5) of the visual scene, the observer sees objects whose true altitudes are not strictly decreasing. The rearrangement is due in part to a thermal inversion layer, in this case called a “duct” (gray), where atmospheric temperature varies nonmonotonically. [Earth’s surface does not appear circular (lower curve) because different scales have been used on the horizontal and vertical axes.] (b) Superior mirage of a ship and a smaller boat at the entrance of the harbor at Victoria, British Columbia, Canada. Both upright and inverted images of the two vessels appear. [(a) Data courtesy Andrew Young. (b) Photo by Craig Clements.]

illusion only appears in the distance, not up close. You’ll work out this and other details in Problem 21.4.

Figure 21.14 shows a more elaborate phenomenon that can occur with a different kind of temperature profile.

21.6.2 A gradient-index lens can minimize spherical aberration

Section 21.3.5 pointed out that a spherical lens will not focus light perfectly (Figure 21.6b). In fact, any homogeneous lens whose front and back surfaces are sections of spheres suffers from spherical aberration.

Animal eyes address this problem in part by implementing gradient-index eyelenses. For example, fish eyelenses are constructed from concentric spherical shells of proteins with slightly different composition. Figure 21.15 illustrates rays traversing a properly designed lens of this sort.38

21.6.3 Gravitational fields can bend light rays even in vacuum

Einstein’s theory of gravitation proposes that space and time can deviate from the euclidean (flat) geometry assumed throughout this book, and that this deviation is responsible for the effects of gravitation. Moreover, because light (and everything else) inhabits

38You’ll work out details in Problem 21.5.
Correction for spherical aberration by a continuously graded refractive index. (a) A set of parallel incoming rays is shown, with their computed trajectories upon entering the medium. In this case, the rays curve inside the lens, because its refractive index is greater in the center than at the periphery. The extra bending has been arranged to make all the rays nearly meet at a common focus. (Problem 21.5 describes the index function that was used to make this diagram.) (b) Actual light rays traversing the eyepiece of an octopus, illustrating how evolution has implemented to solution in (a). [From Jagger & Sands, 1999.]

Einstein rings. The arcs at the center of this image from the Hubble Space Telescope are actually the distorted light of distant galaxies, stretched by the gravitational influence of the closer galaxy cluster SDSS J0146-0929. [ESA/Hubble and NASA; Acknowledgment: Judy Schmidt. www.nasa.gov/image-feature/goddard/2018/hubble-finds-an-einstein-ring].

A gravitational field generates an effective refractive index.

Einstein realized that, although the mathematics of curved spacetime gets complicated, his final expression for the bending of a light ray was mathematically identical to that of a ray passing through ordinary spacetime with a refracting medium having effective index given by

\[ n_{\text{eff}} \approx 1 - 2\phi N/c^2 + \cdots. \]  

(21.14)
Here \( \phi_N \) is the newtonian gravitational potential far from the mass,\(^{39}\) and the ellipsis represents terms of higher order in \( \phi_N/c^2 \).

**Your Turn 21E**

a. Equation 21.14 may be unfamiliar to you, so check that the units make sense.

b. In the neighborhood of a point mass \( M \), the formula becomes \( n_{\text{eff}} \approx 1 + r_e/r \). Use Equation 21.14 to find a formula for \( r_e \) in terms of \( M \).

c. Our Sun isn’t really a point object; a ray originating from behind it will be blocked unless its distance of closest approach to the Sun’s center exceeds the Sun’s radius. Look up that radius, and the mass of the Sun, and hence find the maximum value of the effective refractive index along the ray.

In Problem 21.6 you’ll obtain a complete trajectory of this type.

**21.7 PLUS ULTRA: FROM GEOMETRICAL OPTICS TO WAVE MECHANICS**

For hundreds of years, newtonian mechanics seemed to rule the macroscopic world. However, by the early 1920s new physical ideas were swirling around. In retrospect, three stand out:

- Einstein had proposed that light had a particle aspect.
- Conversely, in the atomic world, Bohr’s idea that bound electrons have a wave aspect could explain the observed quantization of energy, analogously to the discreteness of harmonics in an organ pipe.
- So L. de Broglie proposed that even free electrons also have a wave aspect.

In fact, in 1925 C. Davisson found some odd results about the reflection of electrons by crystals, which he sent to M. Born for advice. Born passed the data along to his student W. Elsasser, along with a hint from Einstein that interference might be expected in such a situation. Elsasser quickly found a quantitative interpretation of Davisson’s data via de Broglie’s hypothesis. But it was hardly obvious what to do with that result.

Erwin Schrödinger was well trained in mechanics, optics and acoustics, so he was well aware that, remarkably, the trajectories of newtonian mechanics could be reformulated as rays solving an equation of eikonal type. So it was natural for him to ask, “Is there any wave equation whose geometrical-‘optics’ limit is the eikonal equation of mechanics?”

After finding such an underlying wave equation, “all” Schrödinger then had to do was to take it seriously outside of the domain where eikonal approximation holds. This crazy idea needed some interpretation, to be sure, but Schrödinger published it in 1926 and it worked out pretty well. In fact, it was one of the biggest lateral-thinking jumps in scientific history.\(^{40}\)

---

\(^{39}\)See Chapter 1.

\(^{40}\)Recall Section 0.4.1 (page 12).
Section 21.7' (page 334) gives some details about the correspondence just sketched.

FURTHER READING

Semipopular:
Uncorrected spherical aberration led to an expensive retrofit of the Hubble space telescope:  
www.nasa.gov/content/hubbles-mirror-flaw. The original report of the defects from NASA: ntrs.nasa.gov/citations/19910003124.
Optical tweezers: In the early 1970s, Arthur Ashkin showed that laser-induced forces could be used to alter the motion of microscopic particles and neutral atoms:  

Intermediate:
General: Garg, 2012; Peatross & Ware, 2015; Zangwill, 2013.
Mirage: Richey et al., 2006.
Gravitational lensing: Schutz, 2022, chap. 11.
Optical tweezers: Jones et al., 2015; Perkins, 2014; van Mameren et al., 2011; Bechhoefer & Wilson, 2002.
Atmospheric optics: Bohren et al., 2015; Bohren & Clothiaux, 2006.


Technical:
Sometimes an apparent mirage may actually be specular reflection: Zhou et al., 2011.
21.5.4 Fermat’s principle

We can attribute a “transit time” \( T \) to any geometrical curve \( \vec{r} \) in space by adding up differentials for each element of that curve. The speed of light near any point \( \vec{r} \) in space is \( c/n(\vec{r}) \), so

\[
T[\vec{r}] = \int ds \frac{n(\vec{r}(s))}{c},
\]

where \( s \) is arc length.

We can now ask, “What curve minimizes the transit time between two given points?” To set this up as a problem in the calculus of variations, we must first generalize the preceding formula to any parameterization (not necessarily arc length) having fixed values at the ends:

\[
T[\vec{r}] = \int_0^{\xi_i} d\xi \|d\vec{r}/d\xi\| n(\vec{r}(\xi))/c, \text{ where } \vec{r}(0) = \vec{r}_i \text{ and } \vec{r}(\xi_i) = \vec{r}_f.
\]

We now ask, “For what curve is this expression stationary?”

Let \( \delta\vec{r}(\xi) \) be a small variation of the curve, equal to zero at each end. Then the first-order variation of transit time is

\[
c^{-1} \int_0^{\xi_i} d\xi \left[ (n + \delta\vec{r} \cdot \nabla) \|d\vec{r}/d\xi\| + n \left( \|\frac{d\vec{r}}{d\xi}\| + \frac{d\delta\vec{r}}{d\xi}\| \right)^{1/2} \right] - T[\vec{r}].
\]

Now simplify the square root via a Taylor expansion:

\[
(\|\frac{d\vec{r}}{d\xi}\| + \frac{d\delta\vec{r}}{d\xi}\|)^{1/2} = \|\frac{d\vec{r}}{d\xi}\| \left( 1 + \frac{d\vec{r}}{d\xi} \cdot \frac{d\delta\vec{r}}{d\xi} \right).
\]

Substituting into the preceding formula and integrating by parts gives

\[
\delta T = c^{-1} \int_0^{\xi_i} d\xi \delta\vec{r} \cdot \left[ \|\frac{d\vec{r}}{d\xi}\| \nabla n - \frac{d}{d\xi} \left( n \left( \|\frac{d\vec{r}}{d\xi}\|^{-1} \frac{d\vec{r}}{d\xi} \right) \right) \right].
\]

(21.15)

Because \( \delta\vec{r} \) is an arbitrary function away from its endpoints, we find that in order for transit time to be stationary, the quantity in square brackets must be zero everywhere.

Now that we have finished the variation step, it is safe to specialize our result to arc length parameterization, in which \( \|d\vec{r}/ds\| = 1 \). Then Equation 21.15 says that every point of the curve must obey

\[
0 = \nabla n - \frac{d}{ds} \left( n \frac{d\vec{r}}{ds} \right).
\]

But this is just the ray equation, so we have derived a form of Fermat’s principle:

* Rays can also be characterized as curves in space that minimize transit time between fixed points.*

This viewpoint is complementary to the one in the main text, which regarded rays as solutions to a differential equation with given starting position and direction.

Fermat’s principle may sound familiar: Newtonian particle trajectories are curves in spacetime that minimize the action functional subject to fixed starting and ending points. That viewpoint is complementary to the one in first-year physics, that regarded trajectories as solutions to a differential equation with given starting position and velocity. In fact, steps similar to the ones just given will also prove to be useful when we find a variational formulation of relativistic particle mechanics in Section 40.3.3.
21.7’ Eikonal approach to classical mechanics, and to wave mechanics

Suppose that $n(\vec{r})$ is the index of refraction in an inhomogeneous, time-independent medium. Section 21.5.5 (page 326) stated that if $S(\vec{r})$ solves the eikonal equation $|\nabla S|^2 = n^2$, then the streamlines of $\nabla S$, traversed at variable speed $c/|\nabla S|$, are possible rays of light. That is, if $L_0 \gg c/(2\pi \omega)$ where $L_0$ is the length scale over which $n$ varies, then $\exp(i\omega(-t + S(\vec{r})/c))$ is an approximate solution of the wave equation. For a given profile of $n(\vec{r})$ there are many solutions to the eikonal equation, and moreover, each solution has many streamlines (systems of rays). These freedoms allow for various initial conditions on the rays.

Following that lead, W. Hamilton found a surprising analogy in 1834. Suppose that $U(\vec{r})$ is the potential energy function for a newtonian particle in a time-independent system and $E$ is a constant. Hamilton found that if $W(\vec{r})$ solves

$$|\nabla W|^2 = 2m(E - U), \quad \text{Hamilton–Jacobi equation} \quad (21.16)$$

then the streamlines of $\nabla W$, traversed at variable speed $E/|\nabla W|$, are possible newtonian trajectories of the mechanical system defined by $U$ with total energy $E$. Thus, “Hamilton’s principal function” $W$ plays a role analogous to that of the eikonal. For a given profile of $E - U(\vec{r})$, there are many solutions to Equation 21.16, and moreover, each solution has many streamlines (solutions to Newton’s law of motion). These freedoms allow for various initial conditions on the trajectories.

Hamilton’s observation helped with the solution to some esoteric mechanics problems. But apparently he did not pursue a question that appears obvious with the benefit of hindsight: What wave equation has a geometrical-optics limit leading to Equation 21.16? This step had to await the experimental discovery of wavelike aspects of the electron. Indeed, following this logic, Schrödinger proposed

$$(-\frac{\hbar^2}{2m} \nabla^2 + U)\Psi = E\Psi.$$ 

To check this proposal, make a trial solution of the Schrödinger equation that is of eikonal form, $\Psi(\vec{r}, t) = \exp((i/h)(W - Et))$, and consider the behavior when $\hbar^2 W \ll |\nabla W|^2$. This gives

$$\frac{1}{2m}|\nabla W|^2 + U - E = \frac{i\hbar}{2m} \nabla^2 W \to 0,$$

which indeed is Equation 21.16. The argument can be extended to nonstationary wavefunctions as well.

The interpretation of the wavefunction $\Psi$ is not as simple as in optics; for example, it is an intrinsically complex-valued function. Max Born proposed that instead of being directly observable, like the electric field, its modulus squared has a probabilistic interpretation.
21.1  *Poor wandering one*

Figure 21.17 shows light shone from a laser pointer into a tank of—mostly—water. The surface of the water is near the top of the tank (that is, far from the beam of light). What do you think might cause the light to take this bizarre, wandering path?

21.2  *Rainbow caustic*

Use the geometrical-optics approximation for this problem. Figure 21.18 shows a cross-section through the center of a transparent sphere of radius $B$ and refractive index $n$.

a. Consider a set of parallel incoming rays in the plane, initially traveling horizontally. One such ray arrives at distance $y_0$ from the centerline, as shown. Find the angle $\theta$ shown in terms of $y_0$ and the sphere radius. This is also the ray’s angle of incidence.

b. Use the law of refraction to find the angle $\psi$.

c. The triangle shown is isosceles. Use that fact to find the reflection point $Q$.

d. Use the law of reflection to conclude that the triangles $PQO$ and $QRO$ are congruent, and hence to find the point of exit, $R$.

e. Use the law of refraction again to find the angle $\gamma$ that the ray makes to the horizontal axis after exiting the glass (the “scattering angle”).

Figure 21.17: See Problem 21.1 and Media 12.

Figure 21.18: [Ray diagram.] See Problem 21.2.
f. Use a computer to draw the four segments of this ray, for each of several \( y_0 \) values. Use the refractive index value \( n = 1.33 \) appropriate for 650 nm (red) light in water.

g. Make a graph showing the angle of the exiting ray as a function of \( y_0/B \). Superimpose another curve showing the corresponding result for \( n = 1.34 \) appropriate for 400 nm (blue) light.

h. Suppose that incoming light is uniformly distributed across the water drop, that is, \( y_0/B \) is a uniform random variable in the range \((0, 1)\). Explain qualitatively why the exit angle you found in (g) is nonuniformly distributed.

i. In fact, the probability density for \( \gamma \) is \( \varphi(\gamma) = \varphi(y_0)/|dy_0/d\gamma| \). Explain why this function has a sharp peak, and at what value of \( \gamma \).

j. If the incoming light is monochromatic, what would you expect to see projected onto a screen to the left of the droplet? How does that predicted intensity distribution differ between red and blue light?

21.3 \( \text{\textcolor{red}{T_2}} \) Secondary bow

Figure 21.9 pointed out that not all the light arriving at \( R \) will emerge from the droplet there; some will suffer a second reflection, then emerge from the droplet the next time it encounters the boundary.

a. Get a computer to calculate and draw the appropriate generalizations of Figure 21.10a and Figure 21.10b.

b. Then sketch the appropriate generalization of Figure 21.12 (page 320). Where in the sky should we look for this rainbow?

c. Blue light bends more upon entering or leaving water than does red. Get your computer to draw the caustic rays for both red and blue and compare your result to the situation for the primary bow.

d. Use your results from (b,c) to explain the reversed order of colors in the secondary bow (Figure 21.8, page 317).

21.4 \( \text{\textcolor{red}{T_2}} \) Mirage

Section 21.6.1 mentioned the problem of light passing through a layer of air that is heated at the bottom, leading to a temperature gradient, hence a density gradient, hence a gradient in the refractive index. This is the special case of a “gradient-index” material, in which the refractive index depends only on height \( z \).

Section 21.5.6 worked out a general formula for the angle \( \theta \) that a ray’s trajectory makes with the \( z \)-axis (Equation 21.13). This condition has two unsurprising solutions (Figure 21.19): One is a straight, horizontal line: \( z = z_0, \theta(x) = \pi/2 \). The other is a straight, vertical line: \( \theta(x) = 0 \). But there can also be ray solutions that are curved.

Suppose that the density profile \( n(z) \) is strictly increasing as \( z \) increases (cooler air away from the hot pavement), and that \( \theta \) starts out tilted downward \((0 < \theta < \pi/2)\). Then \( \theta \) can increase as \( z \) decreases, potentially even leveling off \((\theta \to \pi/2)\), as shown in Figure 21.19.

Suppose that light is emitted by a source \( A \) at height \( z_0 \), and detected at \( B \), also at height \( z_0 \) but a distance \( D \) away. We can characterize a curve in the \( xz \) plane by its height function,
Figure 21.19: Curved light rays in an inhomogeneous medium. The x and z axes are not drawn to the same scale. Dashed lines show the simple solutions mentioned in Problem 21.4. An observer who assumes straight-line propagation will interpret some of the light from A as having come from C. The figure is analogous to what we imagined for shortwave radio skip in Section 21.5.6 (page 327), but upside down.

\[ z = h(x), \quad h(\pm D/2) = z_0. \]  
We wish to find functions \( h(x) \) that give solutions to Equation 21.13 subject to these boundary conditions.

a. To be specific, suppose that \( n(z) = n_\infty(1 - \alpha e^{-z/L}), \) where \( n_\infty \) is the index of air at 30°C, \( n_\infty(1 - \alpha) \) is the index of air at 50°C, \( L = 20 \text{ cm}, \) and your eyes are \( z_0 = 2 \text{ m} \) off the ground. One can look up these values for the two indices of refraction for visible light:

- 30°C: \( n = 1.000262; \)
- 50°C: \( n = 1.000244. \)

Use Equation 21.13 to see how close \( \theta_0 \) must be to \( \pi/2 \) in order for the ray’s trajectory to level off before hitting the ground. Then estimate how far away the mirage will appear to be (horizontal distance from B to C in the figure).

b. Reformulate Equation 21.13 (page 328) as a differential equation determining the entire curve; that is, an equation involving \( dh/dx. \) Solve it analytically or numerically for the situation discussed above. If any simplifying approximations are valid, go ahead and use them. Use the smallest value of \( \theta_0 \) that you found in (a), and use a computer to make a graph showing your solution. (Use different scales for the x and z axes, to show the shape of your solution clearly.)

21.5 Gradient-index (“GRIN”) lens

Use the geometrical-optics approximation for this problem, and neglect reflection at interfaces. Section 21.5.6 considered light ray trajectories in a nonuniform medium whose refractive index depends on only one cartesian coordinate, the height. In the present problem, you’ll adapt the approach to a spherical geometry, where refractive index depends only on radius, that is, the distance \( r \) to the center of a lens. Section 21.6.2 mentioned that this situation holds for the eyeflenses of fish, and claimed that such nonuniformity

---

41 At 633 nm, 101.3 kPa pressure, 50% relative humidity.
42 See also Problem 21.4.
can eliminate much of the aberration created by a uniform spherical lens (compare Figure 21.6a to Figure 21.15).

You may scale all lengths by the radius $B$ of the sphere, that is, work in terms of $\tilde{r} = r/B$ and so on. Let $n_c = n(0)$ be the index at the center, $n_p = n(1)$ its value at the periphery, and $K = n_p/n_c - 1$. Fish eyellenses typically have $n_c \approx 1.52$ and $n_p \approx 1.38$, and are immersed in a watery medium ($n_w \approx 1.33$). W. Jagger investigated a nonuniform but spherically symmetric refractive index profile:

$$n(\tilde{r}) \approx n_c \left(1 + K (0.82\tilde{r}^2 + 0.30\tilde{r}^6 - 0.12\tilde{r}^8)\right).$$

Choose cartesian coordinates with the lens center at the origin and a plane passing through that origin, say the $xy$ plane.

Generate a picture similar to Figure 21.15, by taking the following steps to construct a collection of solutions to the ray equation.

a. Write out both components of the ray equation (Equation 21.12, page 327), which determines the streamlines $\tilde{t}(s)$. It’s a pair of coupled, second-order ordinary differential equations in the two coordinates of a curve lying in the chosen plane, $\tilde{t}_x(s)$ and $\tilde{t}_y(s)$. As in the main text, parameterize the curve by arc length $s$, so that $||d\tilde{t}/ds|| = 1$. [Hint: It may be convenient to define $g(\tilde{r}) = n^{-1}(dn/d\tilde{r})$.]

b. Each ray initially starts outside the lens, traveling parallel to the $x$ axis at some height $y_+$ above the axis. Find the $x$ value at which each incoming ray enters the lens, and the angle it makes relative to the perpendicular (the “angle of incidence”), in terms of $y_+$.

c. Use the law of refraction to find the tangent vector to the ray just after it enters the lens.

d. Use your results in (b,c) to get the required four initial conditions for the ray equation, then use a computer to solve it numerically.

e. Follow your solution to find the value $\tilde{x}_{\text{exit}}$ at which $\tilde{r}$ once again reaches the value 1.

f. The tangent vector $d\tilde{t}/ds|_{\text{exit}}$ then tells you the angle of incidence as the ray crosses the lens→water interface. Use the law of refraction again to find its angle after it leaves the lens.

g. After leaving the lens, the ray is once again straight. Find the point where it hits the $x$ axis, then have your computer draw all three segments (straight→curved→straight).

Construct several such rays with various entry heights.

21.6 $\mathbf{T_2}$ Gravitational lensing

Use the geometrical-optics approximation for this problem. First review Your Turn 21E (page 331). Section 21.5.6 considered light ray trajectories in a nonuniform medium whose refractive index depends on only one cartesian coordinate, the height. In the present problem, you’ll adapt the approach to a nonuniform “medium” (a static gravitational field) whose “refractive index” depends only on radius, that is, the distance $r$ to the location of a point mass (Equation 21.14).

---

See also Problem 21.4.
Choose cartesian coordinates with the point mass at the origin, and a plane passing through that origin, say the xy plane.

a. Write out both components of the ray equation (Equation 21.12, page 327), which determines paths \( \vec{r}(s) \). It’s a pair of coupled, second-order ordinary differential equations in the two cartesian coordinates \( \vec{r}_x(s) \) and \( \vec{r}_y(s) \) of a curve (ray) lying in the chosen plane parameterized by arc length \( s \).

It’s convenient to scale all lengths by the radius \( r_* \) that you found in Your Turn 21E (page 331), that is, to work in terms of \( \tilde{s} = s/r_* \) and so on. It may also be convenient to define \( g(\tilde{r}) = n^{-1}(dn/d\tilde{r}) \).

b. Show that \( g(\tilde{r}) = -1/(\tilde{r}^2(1 + 1/\tilde{r})) \).

c. Consider a series of rays that each start at \( \tilde{x}_0 = -10 \), initially traveling parallel to the x axis at various \( y \) values. The initial position and direction of each ray amounts to the four initial conditions needed in order to solve the ray equation. Use a computer to solve it numerically for several values of \( \tilde{y}_0 \). Because Equation 21.14 is only valid for weak gravitational fields, only examine values of \( \tilde{y}_0 \) that are greater than (say) 5. Continue each ray out far enough to see its final behavior after the flyby.

d. Now generate a picture analogous to Figure 21.15, by having your computer draw your solutions.

e. Your trajectories are distinguished by their \( \tilde{y}_0 \) values. Each one becomes a straight line far from the point mass. Each bends toward the point mass, and eventually hits the x axis at some \( x_* \). Graph \( x_* \) as a function of \( \tilde{y}_0 \).

f. From your numerical result in (c), find the angle of approach \( \tilde{\theta}_a \) at that intercept with the symmetry axis. This gives an apparent angular location in the sky for an observer sitting at \( (x_*,0) \). By the problem’s axial symmetry, the background star appears as a ring with angular radius \( \tilde{\theta}_a \), called an Einstein ring (Figure 21.16, page 330). Make a graph of \( \tilde{\theta}_a \) as a function of \( \tilde{y}_0 \).

g. Finally, combine your results in (e,f) to graph \( \tilde{\theta}_a \) as a function of \( x_* \), that is, apparent angular radius of the Einstein ring as a function of rescaled distance from observer to the lensing object, for a background star at infinity.

[Note: There is a more elegant way to handle trajectories in a spherically symmetric field. However, the method recommended in this problem remains useful in an arbitrary gravitational potential, not just the field near a point mass.]
CHAPTER 22

Diffraction and Semiclassical Approximation

From the purity of an ethereal vapor, ideas may seem to condense into everyday matter; conversely, ideas seem to sublime from the solid, quotidian world into air.

— Peter Galison

22.1 FRAMING: STATIONARY-PHASE POINTS AND THEIR BIFURCATIONS

Chapter 21 explored a limiting case, in which a short-wavelength, nearly-plane wave propagates through a medium. In the case where the medium was piecewise-uniform, and the boundaries between regions were nearly flat on the length scale of the wavelength, we found that the light’s energy can be thought of as traveling along rays that were themselves piecewise-straight. But we know that waves can also diffract—water waves wrap around a breakwater, and we can carry on mobile-phone conversations even when large conducting objects block our line of sight to the nearest antenna tower. Similarly, our calculation of rainbow scattering predicted a completely dark angular zone between primary and secondary bows, and yet observation shows that some light “leaks” into that region.¹ We’ll define diffraction generically as deviations from geometrical-optics behavior when light encounters objects much larger than the light’s wavelength.²

This chapter will begin to improve interfering-ray optics by viewing it as a lowest-order approximation to a fuller, though still approximate, scheme that we will call “semiclassical.” Continuing into the following chapter, we will recover diffractive phenomena as consequences of bifurcations, in which geometrical rays are abruptly lost as we pass a critical viewing point or other parameter value.

Electromagnetic phenomenon: Light can go to places forbidden by geometrical optics.
Physical idea: Even a nearly stationary-phase path is enough to give illumination. The semiclassical approach gives the leading-order effects of not-quite stationary phase paths, explaining edge diffraction and more.

22.2 KNIFE-EDGE DIFFRACTION

¹The experimental data in Figures 23.1 (page 364) and 23.7 (page 374) will show this.
²For objects comparable to, or smaller than, the wavelength, generally the term “scattering” is used instead. This chapter will study solutions created by superposing many simple (spherical) waves. The special case of diffraction that can be understood as arising from the combination of just a few waves is instead called “interference.”
22.2 Knife-Edge Diffraction

Diffraction from a conducting knife-edge. Dots: Monochromatic microwave radiation with wavelength $\lambda = 3.18$ cm arrived from a distance of 60 m, traveling along $\hat{y}$ and polarized along $\hat{z}$. The wave impinged on a thin copper half-plane that was perpendicular to $\hat{y}$ with edge $\{x = y = 0\}$. The time-averaged electric field squared (intensity) was measured on another, parallel plane at $y = 3\lambda$. Solid curve: Prediction of Sommerfeld’s exact solution to this problem. Dashed line: Geometrical optics predicts a sharp shadow. Dotted line: semiclassical approximation, to be discussed in Section 22.4.4. All three curves have been separately normalized to the same area; apart from that, none had any free fitting parameters. [Experimental data from Harden, 1952; see Dataset 2.]

22.2.1 Sommerfeld’s exact solution reproduces classic diffractive features

Throughout this chapter, we continue our study of waves with time-independent, linear boundary conditions. Thus, we may again assume that all fields depend on time harmonically, all with the same angular frequency $\omega$.

After the infinite plane boundaries studied in Sections 21.3.1–21.3.2, the next simplest situation we could imagine is that of electromagnetic radiation impinging on a conducting half-plane, or more realistically a thin conductor (“knife-edge”). An experiment of this sort yielded the data points shown in Figure 22.1. The diffractive effects include:

- Light where none was predicted by geometrical optics (the “shadow” region, positive $x$ in the figure), and also
- Modulation of light where geometrical optics predicts uniform intensity ($x < 0$).

Chapter 23 will also discuss a third key diffractive phenomenon:

- Finite intensity when geometrical optics predicts it to be infinite.

When faced with Electromagnetic Phenomena of this sort, our first impulse is to solve the Maxwell equations with appropriate boundary conditions. Indeed, A. Sommerfeld found an exact solution for the situation in the figure. Although the mathematical justification of his formula is technical, and even the formula itself is rather involved,$^3$ the

---

$^3$ Readers who dare will find the complete expression in Equations 22.33–22.34 (page 357).
result is shown here to point out how successfully the solution accounts for the data, in
particular the first two features mentioned in the preceding paragraph. The main text of
this chapter, however, will take a different approach to problems of this sort: Instead of
an exact solution, we will develop a simpler, and more broadly applicable, approximation
scheme. The figure also shows the result from this “emiclassical” approximation, which
is often surprisingly easy, physically meaningful, and accurate, even when the wavelength
is not much shorter than the characteristic size scale of the system.

Section 22.2.1’ (page 354) gives a derivation of the exact solution that is more direct than
Sommerfeld’s original one.

22.2.2 The solution is expressed in terms of Fresnel integrals

Sommerfeld’s problem is difficult because it requires that we self-consistently find not only
the fields everywhere, but also the induced currents in the conductor that arise to satisfy
the boundary conditions on its surface. Moreover, the boundary condition at infinity is that
the fields become the incident plane wave—not that they vanish. Although the derivation
is lengthy, what may surprise us is that the problem is analytically tractable at all. In part,
this relative simplicity arises from the fact that Sommerfeld’s problem contained no length
scale other than the wavelength and the distance to the observation plane: There is no
“aperture size” for a single straight edge.

One key feature of Sommerfeld’s solution is that it is expressed in terms of functions
now called “Fresnel integrals.” We will soon see that these functions already arise even
in a much simpler, approximate approach, so we now pause to study them.

22.3 STATIONARY-PHASE APPROXIMATION OF OSCILLATORY
INTEGRALS

Sommerfeld’s exact solution was a tour de force, but only a handful of diffraction problems
admit such a complete resolution. Even problems with a lot of symmetry, such as light
impinging on a sphere, are often very difficult, and even when “solved” yield a solution in
the form of a slowly converging, infinite sum of terms. The semiclassical approximation
avoids such problems. Before we can formulate it, however, we need a mathematical
interlude to introduce “stationary phase approximation.” It’s worthwhile because the
method has many other applications beyond diffraction. Luckily, much of what needs to
be said can be said graphically.

22.3.1 The Fresnel integral illustrates the stationary-phase principle

In a diffraction problem, incoming light faces a fixed array of absorbers, mirrors, and/or
dielectrics, and electrons in the materials respond to the incoming radiation. That is,
a continuous distribution of fields interacts with a distribution of charges, also taken

---

4A. Fresnel investigated diffraction around 1819 (long before Maxwell).
22.3 Stationary-phase Approximation of Oscillatory Integrals

(Figure 22.2) Fresnel integral. (a) The real part of the function exp(iξ²). When we integrate it over a region away from ξ = 0, most of the positive contributions (for example, the upper red area) approximately cancel negative ones in pairs (lower red area). (b) A similar cancellation occurs with the imaginary part of the function. In each case, however, the contributions from regions close to 0 do not cancel (hatched areas).

(Figure 22.3) Approximation of an integral by a finite sum. For each value of ξ, the value of the integrand in the Fresnel integral has been represented by a thin arrow in the complex plane. Each arrow has the same length Δξ, but a different angle ξ² relative to the real axis. The sum of those arrows is computed by finding the resultant (heavy arrow). For clarity, this figure divides the range 0.66 < ξ < 1.2 into just three terms; Figure 22.4 subdivides the ranges more finely.

to be continuous. That is a complicated situation, but essentially we must add a lot of contributions to get the complex amplitude of the outgoing light. We first consider a simple calculation of this type: evaluation of the complete Fresnel integral $\int_{-\infty}^{\infty} d\xi \, e^{i\xi^2}$. Once we see how it works, the extensions we need for diffraction will be straightforward.

Evaluating the Fresnel integral may seem to be a nightmare. Normally we are taught that an integral to infinity is meaningless unless the integrand falls off rapidly with ξ; in contrast, this integrand is everywhere a complex number of modulus one! But take a closer look at Figure 22.2. In any region far away from ξ = 0, for example, ξ > 5, the real part of the integrand is positive about as often as it is negative (Figure 22.2a), so we may hope that the integral in such regions will largely cancel, leaving a finite and well-defined net value from the region near ξ = 0. Similar remarks apply to the imaginary part (Figure 22.2b). That is, we may hope that the integral will “oscillate to death” for large |ξ|, and so be well defined. Let’s see if this hope is justified.

Figure 22.3 gives a different graphical representation of an integral of this sort. We imagine taking an interval and dividing it up into small regions of width Δξ. The integral
Chapter 22  Diffraction and Semiclassical Approximation

Figure 22.4: Complex number sums. Evaluation of Fresnel integrals over various ranges of $\xi$, illustrating Idea 22.2. Each panel shows the same sort of sum as in Figure 22.3, but with a larger integration region and more finely subdivided. Each small arrow is a contribution to the integral in the complex plane; the long arrow shows the resultant (sum) representing the value of the integral. The full curve for $-\infty < \xi < +\infty$ resembles panel (a) and is called the Cornu spiral. Panels (a–d) depict integrals over various ranges, all of with the same width $\xi_{\text{max}} - \xi_{\text{min}} = 10$. They show that when the range of integration is large, the magnitude of the total depends strongly on whether the integration range includes the stationary-phase point $\xi = 0$ (panels (a,b)) or does not (panel (d)). Even if the stationary-phase point is nearly in the integration range, its presence off-stage affects the result (panel (c)).

The remaining panels of Figure 22.4 show some other important features of the Fresnel integral. Each contribution to this sum is a complex number. The figure shows arrows with angle equal to $\xi^2$, for each of three values of $\xi$. The sum is obtained by laying successive arrows tail-to-head as usual. The long arrow is the resultant.

Applying this procedure to the full Fresnel integral gives the picture shown in Figure 22.4a. The region near $\xi = 0$ appears at the center of this panel. Here the phase $\xi^2$ is changing slowly, so all the arrows point nearly in the same direction, and give a large excursion (“swan’s neck”) across the page. But at large $|\xi|$, the curve curls up tightly. It doesn’t much matter if we stop the integration at $\xi = \pm 5$ (as shown), or 50, or 5000; the length of the colored arrow will be little changed, as we see from the inward spirals near the endpoints in the figure. Hence, there is a limit as we send the range of integration to infinity: The full integral equals $(1 + i)\sqrt{\pi}/2$. More generally, change of variables shows that for constant real $p$

$$
\int_{-\infty}^{\infty} d\xi \ e^{i\xi^2} = (1 + i)\sqrt{\pi/(2p)} \quad \text{and} \quad \int_{-\infty}^{\infty} d\xi \ e^{-i\xi^2} = (1 - i)\sqrt{\pi/(2p)}.
$$

The second result also arises from taking the complex conjugate of the first.

Ex. Connect the visual features of Figure 22.4a to the observation in Figure 22.2 that the colored regions make nearly canceling contributions to the integral.

Solution: The successive endpoints of arrows in Figure 22.4a represent the integral up to various endpoints. Each curly end of the curve keeps returning to nearly the same point. Figures 22.2a–b show the same behavior for the area under the curves: The positive and negative excursions cancel more and more nearly in pairs as we go farther from $\xi = 0$.

The remaining panels of Figure 22.4 show some other important features of the Fres-
nel integral by evaluating it for various combinations of end points:

- If we evaluate the Fresnel integral only over a finite range of $\xi$ that includes $\xi = 0$ (Figure 22.4a,b), then we get a larger resultant than if we evaluate over an equally large range that does not include 0 [panel (d)].

- However, even when the point $\xi = 0$ is not included, its nearby presence can give a significant contribution [panel (c)].

Again, the point $\xi = 0$ is special, because that’s where the phase $\xi^2$ doesn’t change rapidly with $\xi$. In fact, its rate of change, $d(\xi^2)/d\xi$, is zero at $\xi = 0$. Thus, the angular rotation of $e^{i\xi^2}$ pauses at this point: We call $\xi = 0$ a stationary-phase point of the integrand. Near a stationary-phase point, the arrows representing the integrand align, and can make a large resultant (vector sum in the complex plane). This reasoning suggests an extension of Idea 22.2:

Whenever an integrand oscillates many times over its range of integration, its value will be dominated by contributions from regions close to stationary, or nearly-stationary, phase points.

(22.3)

More specifically, the integral is approximately a sum of terms, one for each stationary-phase point. In the contrary case (slow oscillation or narrow range of integration), then the stationary-phase points have no special significance.

There is another useful way to think about stationary-phase points: If we look near a generic point, say $\xi_0 = 1$, then we find that on one side ($\xi$ slightly greater than 1), the phase function $\xi^2$ is slightly greater than at $\xi_0$, whereas on the other side it is slightly smaller, leading to a curled-up region of the arrow plot. The stationary-phase point is different: The phase function is slightly greater on either side of $\xi = 0$ than its value right at 0, leading to an S-shaped region. To connect those visual ideas to our earlier results, recall from calculus that a place where the first derivative equals zero is generally also an extremal point (in this case, a local minimum).

Section 22.3.1’ (page 358) compares Fresnel and gaussian integrals.

### 22.3.2 Extensions

A key takeaway from Figure 22.2 is that most of the integral arises from a limited range of $\xi$. Taking this one step further, when we examine an integral of the form $\int_{-\infty}^{\infty} d\xi e^{i\xi^2}$ for large constant $p$, then a change of variables shows that most of the integral arises from a narrow range of $\xi$, of width $\approx 1/\sqrt{p}$.

Now suppose that we need to compute $\int_{-\infty}^{\infty} d\xi e^{i\xi^2} g(\xi)$, where $g$ is some function that is held fixed as we consider large $p$. Then the integral still largely cancels except in a narrow range of $\xi$ values close to the stationary-phase value ($\xi = 0$). So we may approximate the full integral by simply evaluating $g$ at that point and regarding it as a constant, that is, pulling it outside the integral:

$$
\int_{-\infty}^{\infty} d\xi e^{i\xi^2} g(\xi) \approx \int_{-\infty}^{\infty} d\xi e^{i\xi^2} (g(0) + \xi g'(0) + \frac{1}{2} \xi^2 g''(0) + \cdots) \approx g(0)(1 + i)(\pi/(2p))^{1/2}.
$$
The crossed-out term is zero. The following terms are suppressed by two or more powers of $p^{-1/2}$.

Indeed, even integrals of the form

$$
\int_{-\infty}^{\infty} d\xi \, e^{ipf(\xi)} g(\xi)
$$

(22.4)

can be approached in this way for $p \to \infty$, fixed real $f$, and fixed $g$. Because only a narrow neighborhood of the stationary-phase point $\xi_*$ contributes, we may Taylor-expand the phase function $f$ as a constant, plus a term quadratic in $(\xi - \xi_*)$, plus higher-order terms. The exponential of the constant is a common factor that can be moved outside the integral; the linear term is missing because we are looking at a stationary point; the quadratic term determines a Fresnel integral; and in lowest approximation we may neglect the rest as before.

Section 22.3.2 outlines a systematic series of corrections to stationary-phase approximation.

### 22.3.3 Breakdown of stationary-phase approximation

The function $f(\xi) = \xi^3$ has a stationary point at $\xi = 0$, but its Taylor expansion about that point has no quadratic term. We will call the integrand superstationary at such points. It may seem as though the factor $p^{-1/2}$ in Equation 22.1 would then make the corresponding Fresnel-type integral be infinite. However, this problem is only a breakdown of the approximation scheme described below Equation 22.4; actually, the integral is perfectly finite by the same sort of graphical observation used earlier.

## 22.4 SEMICLASSICAL APPROXIMATION TO SCALAR DIFFRACTION

### 22.4.1 Overview

The rest of this chapter will outline an extension to geometrical optics that we will call semiclassical approximation and apply it to the knife-edge problem. Chapter 23 will return to lingering questions about rainbows.

Briefly, we will use stationary-phase approximation to reformulate optics problems in terms of oscillatory integrals. Then:

- Often it is enough to consider only the immediate neighborhoods of stationary-phase points, approximating the phase of the integrand near each such point as in Section 22.3.2,\(^5\) and evaluating the resulting Fresnel integrals as though they were over infinite ranges.
- Geometrical optics then reemerges when we square each such contribution to the field and add the results.

\(^5\)At a caustic, there is a stationary-phase point, but Chapter 23 will show that the integral cannot be approximated as Fresnel. Instead we’ll use some other, more accurate, means to estimate the integral at and near such special points.
If the appropriate integration range contains no stationary-phase point, but one is lurking nearby, then there will be a nearly-stationary phase region. Semiclassical approximation amounts to integrating over that region, which as we have seen still gives a nonzero result (Figure 22.4c).

In short, the semiclassical approximation to optical problems says that

When wavelength is much shorter than the length scale of the objects encountered, we may identify the stationary-phase points in or near the region of integration and approximate their contributions by Fresnel-type integrals. (22.5)

A complete justification of this approach is very technical, but we will see how it subsumes geometrical and interfering-ray optics, extends them just far enough beyond their range of validity be useful for diffraction problems, and is quite accurate for two problems with much more difficult (and much less intuitive) exact solutions.

Section 22.4.1’ (page 359) connects semiclassical approximation as used here with the one used in quantum mechanics.

22.4.2 Vector waves resemble scalar waves at large distances

Chapter 38 will explore the electromagnetic waves created by a point object, and find that, when viewed many wavelengths from the source, all components of the wave behave like solutions of the scalar wave equation. Although Sommerfeld’s solution kept the full vector character, the rest of this chapter will simplify by pretending that we are studying a scalar field. (Chapter 24 and the rest of the book will reinstate the full vector character of light.)

22.4.3 The Huygens principle

In free space

We next establish a useful property about traveling scalar plane waves. Suppose that such a wave propagates through vacuum with wavevector \( \mathbf{k} = k(-\hat{y} \cos \theta + \hat{x} \sin \theta) \) and consider the plane \( \{ y = 0 \} \). In words, the wave encounters the plane, traveling downward with angle of incidence \( \theta \). But there is nothing located at the plane; the wave passes right through it. The useful property is that

Everything above the plane, including whatever distant source created the wave, can be replaced by an array of point sources right at the plane (Equation 22.7). Those sources, radiating downward, reproduce the original wave when viewed many wavelengths downstream from the plane. \( \text{Huygens principle} \) (22.6)

Specifically, center the \( x, z \) coordinates such that the observer is at \( (0, h, 0) \), where \( h < 0 \) and \( |h| \gg k^{-1} \) (Figure 22.5). Let the incoming plane wave be \( \frac{1}{2} \Phi(\mathbf{r}) e^{-ikx} \) (c.c.), where

---

*See for example Your Turn 38B (page 585) and Problem 38.3 (page 588).*
\[ \Phi(x, y, z) = e^{i\vec{k}\cdot\vec{r}}. \] Then the required sources in each small region \( \mathrm{d}x \mathrm{d}z \) of the plane contribute

\[ I(h) = \int \mathrm{d}x \mathrm{d}z \left( -\frac{i k}{2\pi} \cos \theta \Phi(x, 0, z) R^{-1} e^{ikR} \right), \quad \text{where} \quad R = \sqrt{x^2 + z^2 + h^2}. \quad (22.7) \]

Intuitively, each point source creates an outgoing spherical wave \( (R^{-1} e^{ikR}) \). The density factor \( \cos \theta \) adjusts for foreshortening when the incidence is not perpendicular. It and the other prefactors will be justified in the following paragraphs. Again, the claim is that Equation 22.7 reproduces the original field evaluated at \((0, h, 0)\) (Figure 22.5). By translation invariance, we get similar results for observers elsewhere in the region \( \{ h < 0 \} \).

To establish Idea 22.6, first note that the oscillatory integral over the plane will be dominated by its stationary-phase contribution because \( k |h| \) is large. That contribution comes from sources close to a point \((x_*, 0, z_*)\) in the plane. Setting the \( x \) and \( z \) derivatives of the phase to zero gives \( z_* = 0 \) and \( x_* = -R_* \sin \theta \), or

\[ x_* = -|h| \tan \theta. \]

That result certainly makes geometrical sense: the ray from the original plane wave that intercepts the observer passes through the imagined plane at \((x_*, 0, z_*)\). Then

\[ R_* = -|h| \sqrt{\tan^2 \theta + 1} = |h| / \cos \theta. \]

The integrand in Equation 22.7 has some rapidly varying factors: \( e^{ik(\cdot)} \). Section 22.3.2 instructs us to Taylor-expand the corresponding phase function,

\[ \phi = \vec{k}_x x + kR(x, z), \quad (22.8) \]

about the stationary-phase point. The first derivatives equal zero by our choice of \((x_*, z_*)\). Next,

\[ \left. \frac{\partial^2 \phi}{\partial x^2} \right|_{x_*} = k \left( R_*^{-1} - \frac{x_*^2}{R_*^3} \right) = kR_*^{-1} (1 - \sin^2 \theta) = (k/R_*)(h/R_*)^2, \]

\[ \left. \frac{\partial^2 \phi}{\partial z \partial x} \right|_{x_*} = 0, \]

\[ \left. \frac{\partial^2 \phi}{\partial z^2} \right|_{x_*} = k \left( R_*^{-1} - \frac{z_*^2}{R_*^3} \right) = k/R_*. \quad (22.9) \]

---

7 Chapter 38 explores spherical waves in greater generality.
22.4 Semiclassical Approximation to Scalar Diffraction

Figure 22.6: Huygens for a mirror. A plane wave reflected at the $xz$ plane can be thought of as being created by a distribution of sources located in that plane, for purposes of observation at a point above that plane ($h > 0$).

Because the cross-term controlled by Equation 22.9 is zero, Equation 22.7 is approximately the product of two Fresnel integrals.

**Ex.** Approximate the integrals in Equation 22.7 as outlined in Section 22.3.2 and simplify the result.

**Solution:**

$$I(h) \approx -\frac{ik}{2\pi R_s} \cos \theta \int dx dz e^{ik(x_s x_s) + \phi(x)} e^{i(h/2)(x - x_s)^2k/2R_s} e^{i(h/2)(z - z_s)^2k/2R_s}.$$

In the first exponential, note that

$$x_s \sin \theta + R_s = -|h| \frac{\sin^2 \theta}{\cos \theta} + |h| \frac{1}{\cos \theta} = |h| \cos \theta.$$

So

$$I(h) \approx -\frac{ik}{2\pi R_s} \cos \theta e^{ik|h| \cos \theta} \int dx e^{i(h/2)(x - x_s)^2k/2R_s} \int dz e^{i(h/2)(z - z_s)^2k/2R_s}.$$

Equation 22.1 gives the $z$ integral as $(1 + i)\sqrt{\pi R_s/k}$ and the $x$ integral as $(1 + i)\sqrt{\pi R_s^3/(kh^2)}$. Bringing everything together gives

$$I(h) \approx e^{ik|h| \cos \theta}.$$

The approximate sign reminds us that we have used stationary-phase approximation.

Recalling that $h < 0$, we can write this answer as

$$I(h) \approx e^{-ikh \cos \theta} = e^{ik(h - h)}.$$

This is indeed the original wave $\Phi$ evaluated at the observation point $(0, h, 0)$, establishing Idea 22.6.

In a dielectric medium, we may use the same logic, everywhere replacing $k = \omega/c$ by $\omega/(c/n)$, where $n$ is the refractive index.

**Mirror**

We may seem to have taken a long road just to recover the free propagation of a wave. But now suppose that instead of nothing present at the surface $y = 0$, we have a mirror. Now the physical region is above the plane, and the sources are not imagined, but real. The same derivation as the preceding case, but evaluated for $h > 0$, now gives us a superposition of the incoming wave with a reflected wave that, when evaluated many wavelengths away,
agrees with that from the law of reflection,\(^8\) that is,
\[
ext \cos \theta = e^{-i \vec{k} \cdot \hat{n}} - e^{+i \vec{k}' \cdot \hat{n}},
\]
where \(\vec{k}'\) is defined in Figure 22.6. Note that the net flux of energy through the surface is zero, because the component of the incoming wave’s energy flux perpendicular to the surface is canceled by that of the reflected wave: Both waves have the same amplitude and \(\hat{y} \cdot \vec{k} = -\hat{y} \cdot \vec{k}'.\)

By attributing the reflected wave to an array of localized sources at the boundary, we can plausibly extend the reasoning to describe reflection by a nearly-flat mirror as well: Now the angle of incidence \(\theta\) varies as our integral sweeps over the surface, resulting in a nonparallel bundle of outgoing rays. This qualitative statement can be used to understand distortion of images seen in a curved mirror; Section 23.2 will treat a related problem quantitatively.

**Interface**

For future use, next consider a wave that propagates through vacuum, encountering a dielectric at the plane \(\{y = 0\}\). This time, we expect both reflection (an additional contribution to the incoming field in the region \(y > 0\)) and refraction (when \(y < 0\)). Again, we would like not to have to follow every charge throughout the dielectric, but rather to supplement the usual reduction of wave speed in the bulk with an array of local sources at the interface that will reproduce the laws of reflection and refraction by radiating both upward and downward.

Reproducing the reflected wave is a straightforward repeat of the mirror case, with one change: The amplitude will be decreased by a reflection factor \(R\), which depends on details of the dielectric medium. We are not yet ready to work out this factor; for now, only note that Chapter 49 will show that it depends on the angle \(\theta\) (and on polarization). Its magnitude must be less than 1 by energy conservation.

For the transmitted wave, substitute wavenumber \(\omega/c \rightarrow \omega/(c/n)\) for the rebroadcast wave but not for the incoming wave \(\Phi\) that excites the charges: Equation 22.7 becomes
\[
\frac{ikn}{2\pi} \int dxdz \cos \theta' \Phi(x,0,z)R^{-1}e^{inkR}, \quad \text{where } R = \sqrt{x^2 + z^2 + h^2}, \quad \sin \theta' = (\sin \theta)/n,
\]
and \(T\) is another constant called the transmission factor. As in the case of no interface, we now ask whether the combined effect of everything above and including the interface can be mimicked by this array of sources at the surface \(\{y = 0\}\):

\(^8\)See Section 21.3.1 (page 308).
22.4 Semiclassical Approximation to Scalar Diffraction

Figure 22.7. Diffraction past a knife edge. Light from a distant, monochromatic source arrives traveling perpendicular to a barrier. The text describes how to calculate the screen illumination at various observation positions \((x, h, 0)\), where \(x\) is distance to the right of the dotted line. We can identify a specific path by giving the value of \(u\), which must be less than zero. Five possible paths are shown for observation at \(x < 0\). Only one of these is a stationary-phase path (heavy line). For observation points with \(x > 0\), there is a similar family of paths, but none of them is stationary-phase.

Your Turn 22A

a. Show that this time, the downward wave generated by Equation 22.10 gives a transmitted field at \((0, h, 0)\) that at large distances approaches \(\mathcal{F} e^{ikh} \cos \theta\), and hence reproduces the law of refraction (Section 21.3.2).

b. Find a version of Equation 22.10 appropriate for a wave exiting from medium to vacuum [not entering as in (a)]. Again show that it reproduces the law of refraction.

22.4.4 Knife-edge diffraction and its disappearing stationary-phase path

Let’s apply these abstract ideas to a concrete diffraction problem. Figure 22.7 shows Sommerfeld’s knife-edge geometry. We now apply the Huygens principle along a plane that includes the barrier and extends leftward from it to minus infinity. An observer measures intensity at various points on the plane at the top of the figure.

We can name points on our plane by their distance \(u\) from the edge of the barrier and position \(v\) parallel to the edge. Rather than carefully considering realistic boundary conditions on the barrier, we will simply omit all paths that pass through it by integrating \(u\) only from \(-\infty\) to 0. Thus, instead of integrating Equation 22.7 over the entire \(xz\) plane, we integrate over a half-plane.

A light source is located infinitely far to the bottom of the figure, so all the lower path segments shown are of essentially equal length. However, the upper segments converging on the observation point have varying lengths. For observer positions \(x < 0\), there is one stationary-phase path labeled by \((u_+, v_+) = (x, 0)\). As the observer moves rightward, however (\(x\) passes 0), that point migrates out of the integration region: There is a bifurcation, or sudden change from one stationary-phase point to none. Stationary-phase approximation therefore predicts an abrupt transition from some illumination to none as we cross \(x = 0\), and at first, that may sound reasonable: As we move into the shadow region, we...
should stop seeing the light. But experiment, and the exact solution, disagree with this prediction (Figure 22.1, page 341), so let’s evaluate the integral by other means.

The Huygens principle (Idea 22.6) tells us to evaluate

$$I(x, h, 0) = \int_{-\infty}^{0} du \int_{-\infty}^{\infty} dv \, R^{-1} e^{2\pi i R / \lambda}$$

where

$$R = \sqrt{(u - x)^2 + v^2 + \hbar^2}.$$  \hspace{1cm} (22.11)

Actually, the $v$ integral does always have a stationary-phase point:

$$\frac{dR}{dv} = R^{-1} v = 0 \text{ at } v_s = 0.$$

Section 22.3.2 outlined how to do this integral by stationary-phase approximation: First approximate the exponential by its Taylor series about $v_s$, whose quadratic term involves

$$\frac{d^2R}{dv^2}\bigg|_{v_s} = R_s^{-1}, \text{ where } R_s = \sqrt{(u - x)^2 + \hbar^2}.$$  

The inner integral is then

$$\approx \int_{-\infty}^{\infty} dv \, R_s^{-1} \exp((2\pi i / \lambda)(R_s + \frac{1}{2} v^2 / R_s + \cdots))$$

$$= R_s^{-1} \left( \frac{2\pi}{2R_s \lambda} \right)^{-1/2} \left( \frac{\pi}{2} \right)^{1/2} (1 + i) e^{2\pi i R_s / \lambda}.$$  

The remaining outer integral is the tricky one, because it has no stationary phase point when $x > 0$, and hence no stationary-phase approximation. Instead we will evaluate it numerically. Dropping some constant prefactors gives

$$\int_{-\infty}^{0} du \, R_s^{-1/2} e^{2\pi i R_s / \lambda}.$$  

Define dimensionless barred quantities by expressing $x, u$, and $h$ as multiples of $\lambda = 2\pi / k$. Then the complex amplitude at $\tilde{x} \lambda$ is a constant times

$$\int_{-\infty}^{0} d\tilde{u} \, \tilde{R}_s^{-1/2} e^{2\pi i \tilde{R}_s} = \text{constant}, \text{ where } \tilde{R}_s = \sqrt{(\tilde{u} - \tilde{x})^2 + \tilde{h}^2}.$$  

Figure 22.1 (page 341) shows that numerically evaluating this integral gives a function of $\tilde{x}$ that agrees fairly well both with experimental data and with the exact calculation, even when $\hbar$ is not very large.

Section 22.4.4 (page 360) mentions an approximation scheme that is distinct from the one used in this section.

22.4.5 Qualitative origin of diffractive phenomena

Returning to the two phenomena listed in Section 22.2.1, we can summarize by saying that:

- Light can appear where geometrical optics predicts none, because the Huygens integral is nonzero when a stationary-phase point is just outside the integration region (Figure 22.4c).
Light can be modulated where geometrical optics predicts uniformity, because the incomplete Fresnel integral circles around its limiting value (compare the left ends of Figures 22.4a,b). Note, however, that the illumination does not drop to zero between peaks (Figure 22.1), because the resultant in the Fresnel integral is never of length zero (again see Figures 22.4a,b).

**FURTHER READING**

*Semipopular:*

*Intermediate:*
Feynman et al., 2010b, chaps. 21 and 30.

*Technical:*
Exact treatment of diffraction from a knife-edge: We have followed Lamb, 1907.
Other derivations include Baker & Copson, 1950; Sommerfeld, 1964b; Born & Wolf, 1999; Schwinger et al., 1998. [[K. McDonald’s notes are a very helpful resource: http://kirkmcd.princeton.edu/examples/sommerfeld.pdf.]]
22.2.1 Lamb’s exact solution of knife-edge diffraction

Certainly today we could instead solve this problem, and many others, by advanced numerical methods, but analytic solutions, when available, can sometimes give us more insight into why the problem turned out as it did.

Consider an infinitesimally thin, conducting barrier lying in the xz plane in the region $x > 0$ (Figure 22.7). A plane wave is incident perpendicular to the barrier’s surface from below. The incident electric field is of the form $E_0 e^{i(ky - \omega t)}$. We wish to find the electric field everywhere in space. This problem was first solved by Arnold Sommerfeld in 1895 (Sommerfeld, 1964b). Here we follow a more straightforward, though still challenging, approach given by H. Lamb (Lamb, 1907). Figure 22.1 (page 341) compares the results to experimental data.

As usual, we’ll assume harmonic time dependence, that is, the electric field is

$$E(x, y, z, t) = \frac{1}{2} E(x, y, z) e^{-i\omega t} + c.c.$$ 

Nothing depends on $z$, so we will drop it from our notation. We will consider only electric fields with $E_z = \tilde{E}_y = 0$.

We must solve the Maxwell equations subject to boundary conditions. We require that the electric field vanishes on the barrier and that far from the edge, we recover the incident plane wave:

$$\tilde{E}(x > 0, y = 0) = 0 \quad (22.12)$$
$$\tilde{E}(x \to -\infty, y) = E_0 e^{i ky} \hat{z}. \quad (22.13)$$

Note that the field in the region of finite, negative $x$ may not look like a single plane wave: For example, in front of the barrier we expect a standing wave, including reflection that spreads to $x < 0$. Instead, Lamb constructed a trial solution of the form

$$\tilde{E}(x, y) = \tilde{E}_T(x, y)e^{i ky} \hat{z} + \tilde{E}_R(x, y)e^{-i ky} \hat{z}. \quad (22.14)$$

Here the transmitted and reflected terms each separately solve the wave equation and propagate mainly upward and downward, respectively. Lamb realized that such a separation might simplify the treatment of boundary conditions.

For example, geometrical optics predicts zero electric field $\tilde{E}_{GO}$ in the first quadrant (the shadow region), an incident ray in the second and third quadrants, and both an incident and reflected ray in the fourth quadrant:

$$\tilde{E}_{T}(x, y) = \begin{cases} 0 & x > 0, y > 0 \\ E_0 & x < 0 \\ E_0 & x > 0, y < 0 \end{cases} \quad \tilde{E}_{R}(x, y) = \begin{cases} 0 & x > 0, y > 0 \\ 0 & x < 0 \\ -E_0 & x > 0, y < 0 \end{cases} \quad \mbox{(geometrical optics)} \quad (22.15)$$

Actually, we expect that light will diffract around the edge of the barrier, causing deviations from these expressions, but still this observation suggests that Lamb’s trial solution Equation 22.14 will be useful.

To impose appropriate boundary conditions requires separate analysis of the solution in different regions. A natural coordinate system for that purpose is parabolic-cylindrical coordinates.

---

Written with Aaron Winn.

November 14, 2023; Contents Index Notation Glossary
Figure 22.8: **Parabolic coordinates on the xy plane.** The constant-coordinate lines are parabolas with focus at the origin. The region \( \{ \eta \geq 0 \} \) covers the full plane. (Compare the elliptical coordinate system in Figure 5.1, page 72, where the lines are ellipses or hyperbolas.)

To define them, we introduce the dimensionless complex coordinate \( w = k(x + iy) \) and let \( u \) be its (complex) square root. Expressing \( u \) as \( \xi + i\eta \) then gives \((\xi, \eta)\) related to the familiar cartesian system by

\[
x + iy = re^{i\varphi} = \left(\frac{\xi + i\eta}{k}\right)^2
\]

\[
x = \frac{\xi^2 - \eta^2}{k}, \quad y = \frac{2\xi\eta}{k}
\]

\[
\xi = \sqrt{k r \cos(\varphi/2)}, \quad \eta = \sqrt{k r \sin(\varphi/2)}.
\]

The virtue of the square-root transformation is that it unfolds the \(+x\) axis: Instead of boundary conditions just above and below that half-line, we have a single condition all along the axis \( \eta = 0 \) (Figure 22.8):

\[
\vec{E}(\xi, \eta = 0) = 0.
\]

Note that both sides of the conducting plane correspond to \( \eta = 0 \), unlike in polar coordinates where we would need separate conditions for \( \varphi = 0 \) and \( \varphi = 2\pi \). Next, note that \( k\sqrt{x^2 + y^2} = \xi^2 + \eta^2 \); The large-\( \eta \) region stays far from the edge. So we will also look for solutions satisfying

\[
\vec{E}(\xi, \eta \to \infty) \to E_0e^{iky} = E_0e^{2i\varphi}.
\]

It is straightforward to cast the 2D laplacian operator in parabolic-cylindrical coordinates, by noting that square root is a complex-analytic map, and\(^{10}\)

\[
\nabla^2 = 4k^2 \frac{\partial^2}{\partial w \partial w^*} = 4 \left| \frac{k}{2u} \right|^2 \frac{\partial^2}{\partial u \partial u^*} = \frac{k^2}{4(\xi^2 + \eta^2)} \left( \frac{\partial^2}{\partial \xi^2} + \frac{\partial^2}{\partial \eta^2} \right).
\]

Although any function of \( x, y \) can be written in the form Equation 22.14, Lamb added a restriction that, while intuitive, must be confirmed later: He sought transmitted and reflected solutions that

\(^{10}\)This circumstance saves us some of the labor needed in other curvilinear systems (Section 5.5, page 72).
We see that wave in Equation 22.15, so we look for a solution of the form

\[ \frac{\partial^2 \mathcal{E}_r}{\partial \xi^2} + 2i \xi \frac{\partial \mathcal{E}_r}{\partial \eta} + \frac{\partial^2 \mathcal{E}_r}{\partial \eta^2} = 0 \quad (22.22) \]

Although we have confirmed that our trial solution yields a solution to the Maxwell equations, we must still investigate its boundary behavior.

That observation motivates us to seek a solution of the form \( \mathcal{E}_r(\xi, \eta) = \mathcal{E}_r(\xi_{\pm}) \), where

\[ \frac{d^2 \mathcal{E}_r}{d \xi^2} - 2i \xi \frac{d \mathcal{E}_r}{d \xi} = 0 \quad (22.26) \]

Similarly, \( \xi_{\pm} \) behaves differently in the fourth quadrant, a behavior reminiscent of the reflected wave in Equation 22.15, so we look for a solution of the form \( \mathcal{E}_g(\xi, \eta) = \mathcal{E}_g(\xi_{\pm}) \), where

\[ \frac{d^2 \mathcal{E}_g}{d \xi^2} - 2i \xi \frac{d \mathcal{E}_g}{d \xi} = 0 \quad (22.27) \]

All together, these restrictions on the trial solution replace two PDEs (Equations 22.22–22.23) by two separable ODEs (Equations 22.26–22.27), whose general solutions are easy:

\[ \mathcal{E}_r(\xi_{\pm}) = C_1 + C_2 \int_0^{\xi_{\pm}} e^{i \xi' d \xi'} \quad (22.28) \]

\[ \mathcal{E}_g(\xi_{\pm}) = C_3 + C_4 \int_0^{\xi_{\pm}} e^{i \xi' d \xi'} \quad (22.29) \]

Although we have confirmed that our trial solution yields a solution to the Maxwell equations, we must still investigate its boundary behavior.

The total electric field Equation 22.14 has now become

\[ \mathcal{E}(\xi, \eta) = \left( C_1 + C_2 \int_0^{\xi_{\pm}} e^{i \xi' d \xi'} \right) e^{i \xi_{\pm} \eta} + \left( C_3 + C_4 \int_0^{\xi_{\pm}} e^{i \xi' d \xi'} \right) e^{-2i \xi_{\pm} \eta}. \quad (22.30) \]

Next, recall from Section 22.3.1 (page 342)\(^{12}\) that

\[ \int_0^\infty e^{i \xi' d \xi'} = -\int_{-\infty}^0 e^{i \xi' d \xi'} = \frac{\sqrt{\pi}}{2} e^{i \xi/4}. \quad (22.31) \]

\(^{11}\) See Equation 18.19 (page 276).

\(^{12}\) See also Section 22.3.1' (page 358).
Applying the boundary conditions Equations 22.19–22.20:

\[
\vec{E}(\xi, \eta = 0) = \left( C_1 + C_2 f_0^\xi e^{i\xi^2 d_\xi} \right) \hat{z} + \left( C_3 + C_4 f_0^\xi e^{i\xi^2 d_\xi} \right) \hat{\xi} = 0 \Rightarrow \begin{cases} C_1 = -C_1, \\ C_4 = -C_2. \end{cases}
\]

\[
\vec{E}(\xi, \eta \rightarrow \infty) = \left( C_1 + C_2 f_0^{\infty} e^{i\xi^2 d_\xi} \right) e^{2i\eta d_\xi} - \left( C_1 + C_1 f_0^{\infty} e^{i\xi^2 d_\xi} \right) e^{-2i\eta d_\xi} \hat{z} = E_0 e^{2i\eta d_\xi}
\]

\Rightarrow \begin{cases} C_1 - C_2(\sqrt{\pi}/2) e^{i\pi/4} = E_0 \\ C_1 + C_2(\sqrt{\pi}/2) e^{i\pi/4} = 0 \end{cases} \Rightarrow \begin{cases} C_1 = \frac{1}{2}E_0 \\ C_2 = -\frac{1}{\sqrt{\pi}}E_0. \end{cases}
\]

Putting it all together, the complex amplitude in parabolic-cylindrical coordinates is given by

\[
\vec{E}(\xi, \eta) = E_0 \left[ \frac{1}{2} - \frac{e^{i\pi/4}}{\sqrt{\pi}} \int_0^{\xi-\eta} e^{i\xi^2 d_\xi} \right] e^{2i\eta d_\xi} - \left( \frac{1}{2} - \frac{e^{-i\pi/4}}{\sqrt{\pi}} \int_0^{\xi+\eta} e^{i\xi^2 d_\xi} \right) e^{-2i\eta d_\xi} \hat{z}. \tag{22.32}
\]

We can re-express this in polar or cartesian coordinates:

\[
\vec{E}(r, \varphi) = E_0 \left[ \frac{1}{2} - \frac{e^{i\varphi/4}}{\sqrt{\pi}} \int_0^{\sqrt{\pi}/2} e^{i\xi^2 d_\xi} \right] e^{ikr \sin \varphi} - \left( \frac{1}{2} - \frac{e^{-i\varphi/4}}{\sqrt{\pi}} \int_0^{\sqrt{\pi}/2} e^{i\xi^2 d_\xi} \right) e^{-ikr \sin \varphi} \hat{z}, \tag{22.33}
\]

\[
\vec{E}(x, y) = E_0 \left[ \frac{1}{2} - \text{sgn}(\xi) \frac{e^{-i\varphi/4}}{\sqrt{\pi}} \int_0^{\sqrt{\pi}/2} e^{i\xi^2 d_\xi} \right] e^{iky} - \left( \frac{1}{2} - \text{sgn}(\xi) \frac{e^{i\varphi/4}}{\sqrt{\pi}} \int_0^{\sqrt{\pi}/2} e^{i\xi^2 d_\xi} \right) e^{-iky} \hat{z}. \tag{22.34}
\]

As a check, the exact result agrees with the geometrical optics solution far away from the barrier. Many computer math packages offer the incomplete Fresnel integral as built-in functions:

\[
\int_0^\infty e^{i\xi^2 d_\xi} = \sqrt{\pi/2} \int_0^{\sqrt{\pi}/2} e^{i\tau^2/\tau} d\tau = \sqrt{\pi/2} \left[ C \left( \sqrt{\frac{2}{\pi}} u \right) + iS \left( \sqrt{\frac{2}{\pi}} u \right) \right] \tag{22.35}
\]

where

\[
C(x) \equiv \int_0^x \cos(\pi t^2/2) dt, \quad S(x) \equiv \int_0^x \sin(\pi t^2/2) dt, \tag{22.36}
\]

are the Fresnel cosine and sine integrals.

In order to obtain the time-averaged energy density everywhere in space, we use a tactic introduced in Section 20.2.1 (page 299):\(^{13}\)

\[
\langle |\vec{E}|^2 \rangle = \frac{1}{2} \vec{E} \cdot \vec{E}^*. \tag{22.37}
\]

Figure 22.9 shows the energy density as a heatmap.

Lamb’s strategy may seem to have relied on several brilliant hunches. To some extent, these may have come naturally to a scientist with deep experience in partial differential equations. We should note, however, that Lamb had the benefit of already knowing the result from Sommerfeld’s much more involved derivation.

\(^{13}\)Also see the Example on page 576.
Figure 22.9: [Mathematical function.] **Sommerfeld’s solution throughout space.** Electric field intensity obtained by substituting Equation 22.34 into Equation 22.37. Compare Figure 22.1 (page 341), which displays a slice at constant y. [Calculation by Aaron Winn.]

Figure 22.10: **Contour deformation** in the complex $\xi$ plane. The bracket shows a region where $u$ is approximately real (see text).

### 22.3.1’ Another approach to the Fresnel integral

The main text approached the complete Fresnel integral by a graphical means. But it looks superficially similar to a better known integral, the gaussian: $\int_0^\infty du e^{-u^2} = \sqrt{\pi}$.

To see the connection, first note that we may deform the contour from the real $\xi$ axis to anything else with the same starting and ending values without changing the integral, because exponential and square are a complex-analytic functions. If we make the choice shown in Figure 22.10, then throughout the central region $u = \xi (1 - i)/\sqrt{2}$ will run over real values from negative to positive.\(^{14}\) Next, change variables in the integral from $\xi$ to $u$ and neglect the outlying regions because they are oscillatory.

We have now transformed the Fresnel integral into a constant times the gaussian integral:

$$\int d\xi e^{i\xi^2} \approx \int_{-\infty}^{\infty} du \sqrt{2 \pi} e^{-u^2} = \sqrt{2 \pi} = (1 + i)\sqrt{\pi / 2}, \quad (22.38)$$

which becomes exact in the limit of large central region of the deformed contour. Thus we recover the main text’s result.

\(^{14}\)The contour obtained from the one shown by reflecting through the real axis would also start and end at the same places, but is not helpful for evaluating the integral.
22.3.2' Corrections to stationary-phase approximation

Like any approximation, stationary-phase is just the leading term in an expansion. Suppose that we wish to evaluate

$$I = \int_{-\infty}^{\infty} d\xi \exp(i(p\xi^2 + A\xi^4 + B\xi^6 + \cdots))$$

approximately in the limit of large $p$. Section 22.3.2 (page 345) suggested that we neglect the $A$ and $B$ terms altogether, because the integration range that contributes significantly to the answer is narrow, $|\xi| \lesssim 1/\sqrt{p}$, and so the higher order terms will be suppressed. But instead of dropping those terms, let's now use their smallness to do a Taylor series expansion of the exponential:

$$I = \int_{-\infty}^{\infty} d\xi \exp(i(p\xi^2 + IA\xi^4 + IB\xi^6 - 1/2 A^2\xi^6 + \cdots)).$$

We may drop all odd-power terms, because the integral of an odd function over a symmetric range is zero. To evaluate the remaining terms, we can use the same trick as is used to find the variance of the gaussian distribution:

$$\int d\xi \xi^2 e^{ip\xi^2} = -\frac{d^2}{dp^2} \int d\xi \exp(i(p\xi^2)) = -\frac{d^2}{dp^2}(1 + i)\left(\frac{\pi}{2p}\right)^{1/2}.$$ 

Hence the $B$ term is suppressed relative to the leading term by $B/p^2$, and so on for higher terms. Similar considerations apply when there is another slowly-varying function $g(\xi)$ under the integral (Equation 22.4, page 346).

22.4.1' About the term “semiclassical”

Although we have been studying only classical (non-quantum) electrodynamics, we have used the term “semiclassical.” The term acknowledges an analogy to quantum mechanics.

Classical trajectories are extrema of the action functional. In some situations, quantum corrections are small and may be understood via a path integral over just the neighborhood of a classical trajectory. For example, in this way we may find a nonzero quantum amplitude for a particle to tunnel to a region where it is classically forbidden (no classical trajectory leads there). One way to approximate such “tunneling” amplitudes is via an approach dominated by a trajectory that is almost, but not quite, a solution of the classical equations of motion, for example, because it surmounts an energetically forbidden barrier.

The analogy in the main text of this chapter was that:

- We considered rays of light, which are extrema (stationary points) of a different functional (the transit time, Section 21.5.4’, page 333).
- Diffraction involves light penetrating to regions where no solution of the ray equation, and hence no stationary-phase path, connects source to observer.
- Nevertheless, we were able to approximate the corresponding illumination via an approach dominated by a light path that is almost stationary-phase.

15See Chapter 40.
22.4.4 Vista: physical optics
The approach we have called “semiclassical” also included some reasonable but nonrigorous steps, such as our treatment of the barrier in Section 22.4.4 (page 351), instead of finding a fully self-consistent solution to the Maxwell equations. A more formal, but still approximate, treatment called “physical optics” codifies these shortcuts and attempts to estimate the size of errors that they introduce.
22.1  DIY Cornu spiral
[[Not ready]]

22.2  [[Not ready]]
CHAPTER 23

Diffraction at Caustics

It is remarkable that just by looking up in the sky at [a] fine detail (visible in about half of natural rainbows), one sees directly the replacement of the theory of light in terms of rays—geometrical optics—by the deeper and more fundamental wave theory.

— Michael Berry

23.1 FRAMING: SUPERSTATIONARITY

Chapter 21 showed in an approximate scheme (geometrical optics) how an inhomogeneous medium can focus light via the formation of caustics. The result can be æsthetically pleasing and unexpected (the rainbow), or practically useful for image formation (in optical instruments including telescopes and microscopes).

But Chapter 22 pointed out limitations of geometrical optics, in situations without transparent media, such as edge diffraction. In particular, light can appear where geometrical optics predicts none (shadow regions), because even a nearly stationary-phase path can lead to some illumination. It is now time to investigate such effects in situations with media. We’ll see how the diffractive blurring of edges affects the performance of optical instruments.

Another shortcoming of geometrical optics was its unphysical prediction of infinite light intensity at the rainbow caustic, whereas the observed peak intensity of the primary rainbow is finite (and also not quite located at the caustic angle). Finally, geometrical optics failed to predict the supernumerary bows seen at slightly larger scattering angle (Figure 21.8). Thomas Young’s interfering-ray optics goes part of the way, qualitatively explaining the supernumerary bows, but does not get their locations quantitatively correct. The present chapter will begin by revisiting all of these problems.

Section 23.3 will then turn to a more singular form of caustic, the one produced when a lens focuses light to a point. Again, we will see that the geometrical optics prediction of an infinitely narrow, infinitely bright focus is an artifact of using stationary-phase approximation where it is not justified, and that diffraction replaces those predictions by something more reasonable.

Electromagnetic phenomenon: Even when spherical and chromatic aberrations are small, diffraction still prevents perfect focusing by a lens.

Physical idea: Focusing arises from a superstationary-phase path to a special observation

---

1 See Problem 21.2 (page 335). Here “intensity” means power per solid angle. The singularity is integrable, so the total power delivered is finite; nevertheless, the prediction of infinite concentration is unphysical.

2 See Section 21.4.5 (page 321).
point, but nearby points will also be reachable via nearly-stationary phase paths.

23.2 THE SEMICLASSICAL RAINBOW

The next sections introduce many symbols, which are summarized here for reference:

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td>total scattering angle, defined in Figure 23.2; $\gamma_*$, a special value (the caustic angle)</td>
</tr>
<tr>
<td>$\theta, \alpha, \beta$</td>
<td>angles defining transition points, defined in Figure 23.2</td>
</tr>
<tr>
<td>$\theta_0, \alpha_0, \beta_0$</td>
<td>values about which to Taylor expand</td>
</tr>
<tr>
<td>$\delta \theta, \delta \alpha, \delta \beta$</td>
<td>small variations</td>
</tr>
<tr>
<td>$\theta_<em>, \alpha_</em>, \beta_*$</td>
<td>stationary-phase values when $\gamma = \gamma_*$</td>
</tr>
<tr>
<td>$L_0$</td>
<td>very large distance to a point source of light (Figure 23.2)</td>
</tr>
<tr>
<td>$d$</td>
<td>very large distance to a projection screen</td>
</tr>
<tr>
<td>$B$</td>
<td>droplet radius</td>
</tr>
<tr>
<td>$R_1, R_2, R_3, R_4$</td>
<td>path segment lengths, defined in Figure 23.2</td>
</tr>
<tr>
<td>$n$</td>
<td>refractive index</td>
</tr>
<tr>
<td>$L_{\text{eff}}$</td>
<td>effective path length</td>
</tr>
</tbody>
</table>

23.2.1 Somewhere under the rainbow

Chapter 21 showed that geometrical optics can explain the most basic aspects of rainbows, but Section 22.1 (page 340) also pointed out some failures.

Raindrops are approximately spheres, generally with radius at least a thousand times larger than the wavelength of visible light, and hence the successive refraction/reflection/refraction events in rainbow scattering are all many wavelengths distant from each other.\(^3\) So let us try to capture the essential physics missing from the geometrical optics approach via a semiclassical calculation.\(^4\) Certainly the intensity profiles from rainbow scattering are experimentally accessible in controlled experiments (Figure 23.1).

23.2.2 Earlier results rederived as stationary-phase applied to the Huygens integral

Figure 23.2 defines quantities discussed in this section. We will not require that the angles in the figure obey the rules of geometrical optics (laws of reflection or refraction). Instead, we will generalize the approach in Section 22.4.4, applying the Huygens principle (Idea 22.6, page 347) three times in succession:

- At the top of the diagram, we integrate over all entry points (angle $\theta$) using Equation 22.10 (page 350) to find the refracted field.
- At the right, we integrate the refracted field over all reflection points (angle $\theta + \alpha$) using Equation 22.7 (page 348) to find the reflected field.

\(^3\)The next section will point out that two stationary-phase paths can merge, but the kinks on each path remain well separated even near the caustic.

\(^4\)Section 22.4.1 (page 346) introduced this approach.
Figure 23.1: **Light intensity patterns from a single water droplet** of radius ≈ 2 mm, illuminated with monochromatic light of wavelength 639 nm, in such a way that reflection from the front surface of the droplet does not contribute. The resulting circular rainbow appears as straight bands in the narrow slice shown. Both the primary bow (right) and the secondary (left) are divided into a main peak and a series of supernumeraries that trail off toward the edges of the photo. Although the center region is dark, some light leaks over into the region forbidden by geometrical-optics approximation ("Alexander’s dark band," left of arrow). Quantitative data appear in Figure 23.7. [From Walker, 1980.]

Figure 23.2: [Diagram.] **Rainbow geometry.** Light from a distant point source encounters a droplet, then lands on a distant screen, where its intensity is observed at various points. The diagram is not to scale; the main text considers only the case \( L_0 \gg B \) and \( d \gg B \). In that limiting case, the angles \( \delta' \) and \( \delta \) become equal and the angle \( \gamma \) shown becomes the scattering angle. Angles \( \delta, \alpha \), and \( \beta \) are generic; instead of taking them to be determined by geometrical optics, the main text integrates over them. Paths that do not contribute to the primary rainbow are not shown; some of them are analogous to rays seen in Figure 21.7 (page 314).

- At the bottom, we integrate the reflected field over all exit points (angle \( \theta + \alpha + \beta \)) using your result from Your Turn 22Ab (page 351) to find the scattered field.

We will now see how this scheme reproduces geometrical optics approximation when that is valid, predicts when it will break down and why, and tells us what to do then.

When we include the final path segment from exit to the observer, Huygens gives us four exponential factors, one from each path segment. These combine into a single
and $n$ is the index of refraction of water at vacuum wavelength $\lambda = 2\pi/k$. The factor $e^{i\theta}$ is to be multiplied by various $1/R$ factors from the Huygens principle, then integrated over the variables $\theta$, $\alpha$, and $\beta$ characterizing a path, while holding fixed the parameters $L_0$, $d$, and observer position (or equivalently the angle $\gamma$).

We will illustrate the method in two dimensions for simplicity; that is, we neglect the third dimension coming out of the page. Again, our task is to compute the light intensity at a particular point on the projection screen.

We are interested in the situation where the source distance $L_0$ is much larger than the droplet radius $B$. We may Taylor expand $7$

$$R_1 = \left( (L_0 - B \cos \theta)^2 + B^2 \sin^2 \theta \right)^{1/2} = L_0(1 - 2(B/L_0) \cos \theta + (B/L_0)^2)^{1/2}$$

$$\approx L_0(1 - (B/L_0) \cos \theta + \frac{1}{2}(B/L_0)^2 - \frac{1}{8}(2B/L_0)^2 \cos^2 \theta + \cdots) = L_0 - B \cos \theta + (B/L_0) \sin^2 \theta + \cdots.$$

We assume that $L_0$ is so large that we may drop the last term (and higher terms), and so

$$R_1 \approx \text{const} - B \cos \theta. \quad (23.2)$$

**Ex.** How large must $L_0$ be if $B \approx 0.5 \text{ mm}$ and $\lambda \approx 500 \text{ nm}$?

**Solution:** The first term we wish to drop contributes a factor of $\exp \left( \frac{2\pi B^2}{2L_0} \sin^2 \theta \right)$. This factor will be close to 1 if its argument is small, so we require

$$2\pi(0.5 \cdot 10^{-3} \text{ m})^2 \ll (5 \cdot 10^{-7} \text{ m})L_0,$$

that is, $L_0 \gg 3.2 \text{ m}$.

To find $R_4$, first note that this path segment connects the point with $(x, y) = (-B \cos(\theta + \alpha + \beta), B \sin(\theta + \alpha + \beta))$ to the point $(d \cos \gamma, -d \sin \gamma)$ on the screen. Thus,

$$R_4 = \sqrt{(d \cos \gamma - (-B \cos(\theta + \alpha + \beta)))^2 + (-d \sin \gamma - (B \sin(\theta + \alpha + \beta)))^2}.$$

Because $d \gg B$, we have

$$R_4 \approx d + B \cos(-\gamma + \theta + \alpha + \beta) + \cdots, \quad (23.3)$$

where again the ellipsis indicates terms we may drop. The two isosceles triangles in Figure 23.2 have

$$R_2 = 2B \sin \left( \frac{\alpha}{2} \right) \quad \text{and} \quad R_3 = 2B \sin \left( \frac{\beta}{2} \right). \quad (23.4)$$

---

5 The phase can also be regarded as frequency times transit time. When we find its critical paths later in this section, we are therefore extremizing transit time, consistent with Fermat's principle (Section 21.5.4’ page 333).

6 Some experiments are done with cylindrical streams of water, for which the 2D calculation is literally applicable.

7 Section 3.4 (page 40) introduced such expansions.
Chapter 23 Diffraction at Caustics

To find the stationary-phase path(s), if any, we now fix a point on the observation screen (that is, $\gamma$ is a constant) and consider small variations about a particular path:

$$\theta = \theta_0 + \delta\theta, \quad \alpha = \alpha_0 + \delta\alpha, \quad \beta = \beta_0 + \delta\beta. \quad (23.5)$$

Next, expand $L_{\text{eff}}$ as a Taylor series in $\delta\theta, \delta\alpha, \text{ and } \delta\beta$:

$$L_{\text{eff}} = L_{\text{eff}}^{[0]} + L_{\text{eff}}^{[1]} + L_{\text{eff}}^{[2]} + \cdots. \quad (23.6)$$

The leading (zeroth-order) term is an irrelevant constant, because $\exp(iL_{\text{eff}}^{[0]}/(2\pi\lambda))$ is a common factor that we may pull out of the integral, and ultimately we only want the modulus of the complex amplitude. It will also be convenient to define

$$\Lambda = -\gamma + \theta_0 + \alpha_0 + \beta_0. \quad (23.7)$$

Combining Equations 23.1 and 23.2–23.4 then gives the first-order terms as

$$L_{\text{eff}}^{[1]} = B[\delta\theta (\sin\theta_0 - \sin \Lambda) + \delta\alpha (n\cos(\alpha_0/2) - \sin \Lambda) + \delta\beta(n\cos(\beta_0/2) - \sin \Lambda)]. \quad (23.8)$$

For the path under consideration to be stationary-phase, the coefficients of $\delta\theta, \delta\alpha,$ and $\delta\beta$ must all separately be zero. Thus, $\sin\theta_0, n\cos(\alpha_0/2),$ and $n\cos(\beta_0/2)$ must all equal $\sin \Lambda$ (and hence must equal each other), or equivalently

$$\sin\theta_0 = n\cos(\alpha_0/2) \quad (23.9)$$

$$\alpha_0 = \beta_0 \quad (23.10)$$

$$\theta_0 = \text{either } \Lambda \text{ or } (\pi - \Lambda). \quad (23.11)$$

We can interpret Equations 23.9–23.11 physically as follows:

- Figure 23.2 shows $\psi_0 = (\pi - \alpha_0)/2$, so Equation 23.9 is equivalent to the law of refraction at the entry point.
- Equation 23.10 says that the two triangles bordered by the light ray are congruent, which is equivalent to the law of reflection.
- The same congruence of triangles also implies that upon exit from the sphere, the angle of incidence equals the same angle $\psi_0$ describing refraction upon entrance. Hence the angle of refraction upon exit equals $\theta_0$.
- Equation 23.11 has two cases because $\sin(\pi - x) = \sin x$, but the first is spurious; we are interested in the other one, which says that

$$2\theta_0 = \pi - (-\gamma + 2\alpha_0). \quad (23.12)$$

**Your Turn 23A**

Show that Equation 23.12 is therefore equivalent to the law of refraction upon exiting the sphere.

**Section 23.2.2' (page 381) gives a vista onto more general caustics.**

---

8It says that $\gamma_0 = 2\alpha_0$, which contradicts the fact that the ray passing through the center of the sphere has $\alpha = 0$ but $\gamma = \pi$. 

November 14, 2023; Contents Index Notation Glossary
23.2.3 Stationary-phase paths can merge and annihilate as parameters change

In short, the requirement of stationary phase recapitulates the rules of geometrical optics. We therefore know that Equations 23.9–23.11 have two solutions for large angles $\gamma$, and no solutions for smaller $\gamma$. That is, as we move on the projection screen from larger to smaller scattering angles ($\gamma$ decreasing from 180 deg), the two stationary-phase points in $\partial \alpha \beta$ space approach each other, merge, and disappear: The phase function $\phi$ in Equation 23.1 undergoes a bifurcation. Unlike the knife-edge situation, however, it goes abruptly from two stationary-phase points to none as we scan over $\gamma$.

We are particularly interested in the special value $\gamma_*$ at which the two stationary-phase paths merge. This is the caustic, and it is where geometrical optics approximation gave the result of infinite intensity (Problem 21.2). We need to replace that approximation by a different approach that is well behaved at such points.

To find the caustic angle $\gamma_*$, combine Equations 23.7 and 23.9–23.11 to get

$$\gamma = 2 \theta_0 + 4 \cos^{-1}(n^{-1} \sin \theta_0) - \pi. \quad (23.13)$$

Figure 23.3a shows this function. It is minimum at $\theta_0 = \theta_*$, where

$$\frac{1}{2} n \sin(\alpha_*/2) = \cos \theta_* . \quad (23.14)$$

Use Equation 23.9 to eliminate $\alpha_*$:

$$2 n^{-1} \cos \theta_* = \sqrt{1 - n^{-2} \sin^2 \theta_*}$$

$$\sin \theta_* = \sqrt{(4 - n^2)/3} . \quad (23.15)$$

---

9See Section 21.4.3 (page 317) and Problem 21.2.
10See Section 22.4.4 (page 351).
For water in air, evaluating this expression and substituting in Equation 23.13 gives the caustic angle as $\gamma_s \approx 138 \text{ deg}$, as seen in Figure 23.3a.

### 23.2.4 A semiclassical calculation addresses the shortcomings of geometrical optics

Figure 23.3b shows a simplified version for the mathematical behavior that we have found: The horizontal axis represents a single variable $u$ schematically representing all three integration variables $\delta \theta, \delta \alpha$, and $\delta \beta$. The three curves show imagined behavior above, at, and below a critical value of the control parameter $\gamma$. We see two stationary points (1,2) that move together, merge (3), and then disappear for $\gamma$ below a critical value.

What happens to the intensity pattern when this merger occurs? Above the caustic angle, the function in Figure 23.3b has two stationary-phase points (marked 1 and 2 on the dashed curve). When we perform the integral, each stationary-phase point leads to a large excursion (“swan’s neck”), flanked by tight coils in the complex plane (Figure 22.4a). Depending on how much coiling occurs between the two swan’s necks, the large excursions will alternate between reinforcement (Figure 23.4c,e) or cancellation [panel (d)]. Because the length-squared of the resultant controls light intensity, we find an interference pattern. The details depend on wavelength, potentially explaining the colored supernumerary bows (Figure 21.8).

Exactly at the caustic angle, there is only a single stationary-phase point (analogous to the solid curve in Figure 23.3b), but the phase function is flatter there than in the generic case. Section 22.3.3 called it superstationary: As a function of $u$, the leading behavior is $u^3$, not $u^2$, implying that the phase stays small out to larger values of $|u|$ than in the generic case. Accordingly, the integral is especially big there (Figure 23.4b), giving rise to the bright caustic ring. However, in contrast to the geometrical optics result, the resulting intensity (amplitude squared) is finite at the caustic, because the integrand is everywhere finite, and so is the range of integration. Indeed, an even larger resultant arises slightly above $\gamma_s$ (Figure 23.4c).

For scattering below the caustic angle, there is no stationary-phase point at all (analogous to the dotted curve in Figure 23.3b). The integral that yields the amplitude consists mostly of tight spirals, so its modulus squared is small (Figure 23.4a). Thus we recover Alexander’s dark band. Unlike in geometrical optics, however, there is some illumination below the caustic angle: When $\gamma$ is just slightly less than $\gamma_s$, the influence of the lost superstationary point nearby gives a significant contribution.

The following sections give details about the calculations whose results are shown in Figures 23.4 and 23.5.

### 23.2.5 The second-order variation of phase degenerates at the caustic

The preceding section set out some qualitative expectations. To confirm them, we must look at the higher-order terms in $L_{\text{eff}}$ (Equation 23.1). Equation 23.8 only expanded this...
Contributions to rainbow scattering, for an increasing series of scattering angles. Each panel represents a one-dimensional complex integral in the same style as Figure 22.4 (page 344). In each case, the long arrow is the resultant from adding many contributions (small arrows). Number labels refer to corresponding features in Figure 23.3b. (a) At scattering angles slightly below $\gamma_*$, there is no stationary-phase path at all; nevertheless, the resultant is nonzero. (b) Exactly at the caustic angle, we get superstationary phase and hence strong constructive interference. (c–e) Beyond the caustic angle, the two stationary-phase paths each make contributions similar to Figure 22.4b (page 344). These can interfere constructively (c,e) or destructively (d). Details of the integration paths chosen in $\theta\alpha\beta$ space are discussed in Section 23.2.7. See also Problem 23.1.

Semiclassical calculation of rainbow scattering. Dashed curve: Geometrical optics predicts zero light below the caustic angle $\gamma_*$ and infinite intensity right at it (see Problem 21.2). Points a–e on the solid curve correspond to the five panels of Figure 23.4: a: small but nonzero intensity below the caustic angle; b: finite intensity at $\gamma_*$; and c–e: an interference pattern for $\gamma > \gamma_*$. The calculation assumed an index of refraction 1.33 and droplet radius $B$ related to light wavelength by $B = 367\lambda$, and neglected the contribution from the path that reflects off the droplet surface instead of entering it.

function to first order in small variations around chosen starting values $\theta_0$, $\alpha_0$, and $\beta_0$. That sufficed to find the bifurcation, but now we need to go further.

To begin, we examine the phase function for an observer located exactly at the caustic
angle, that is, set $\gamma$ equal to the value$^{12}$ $\gamma_s$. The solid curve in Figure 23.3b suggests that we are looking for a flat region of the phase function, but there are three integration variables—not just one as in the figure. At least we know where to look: The stationary-phase path is at $\hat{\theta}_s$ and $\alpha_s = \beta_s$ given by Equations 23.14 and 23.15.

As in Equation 23.5, let $\delta \theta$, $\delta \alpha$, and $\delta \beta$ be small deviations of $\theta$, $\alpha$, and $\beta$ about the starred values. We already know that the phase function’s variation vanishes to first order in these deviations, so we now collect the second-order terms. Again using Taylor expansion, they are

$$L_{\text{eff}}^{[2]} = \frac{1}{2} B \left[ \delta \theta^2 \cos \theta_s + (\delta \alpha^2 + \delta \beta^2) \frac{-n}{2} \sin(\alpha_s/2) - (\delta \theta + \delta \alpha + \delta \beta)^2 \cos \Lambda_s \right],$$

where $\Lambda_s$ denotes $-\gamma_s + \theta_s + \alpha_s + \beta_s$. For the second term of Equation 23.16, Equation 23.14 says that $\sin(\alpha_s/2) = (2/n) \cos \theta_s$. Also, Equation 23.15 says that

$$\cos \theta_s = \left(1 - \frac{4 - n^2}{3}\right)^{1/2} = \left(\frac{n^2 - 1}{3}\right)^{1/2}.$$ 

Combining with other formulas gives

$$L_{\text{eff}}^{[2]} = \frac{1}{2} B \left[ (2 \delta \theta^2 + \delta \alpha^2 + \delta \beta^2 + 2 \delta \theta \delta \alpha + 2 \delta \theta \delta \beta + 2 \delta \alpha \delta \beta) \cos \theta_s - (\delta \alpha^2 + \delta \beta^2) \frac{1}{2} \sqrt{n^2 - \sin^2 \theta_s} \right].$$

Next, note that Equation 23.15 says $(n/2)\sqrt{1 - n^2 \sin^2 \theta_s} = \sqrt{(n^2 - 1)/3}$. Thus, we have the quadratic function

$$L_{\text{eff}}^{[2]} = \frac{1}{2} B \sqrt{(n^2 - 1)/3} \left[ \begin{array}{c} \delta \theta \\ \delta \alpha \\ \delta \beta \end{array} \right] \left[ \begin{array}{ccc} 2 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{array} \right] \left[ \begin{array}{c} \delta \theta \\ \delta \alpha \\ \delta \beta \end{array} \right].$$

We must now pause to see how to handle the required integral.

### 23.2.6 Extension of stationary phase to a function of several variables

We now need to generalize the Fresnel integral to multiple variables: Consider

$$I = \int d^N u \ e^{i T u/2},$$

where $T$ is a real, symmetric matrix and the integral is over all $N$-dimensional space. To get started, recall that any symmetric matrix can be brought to diagonal form: There is always an orthogonal matrix $R$ such that $D = R^T T R$ is real and diagonal.$^{13}$ Changing variables from $u$ to $v = Ru/\sqrt{2}$ gives

$$\int 2^{N/2} |\det R|^{-1} d^N u \ e^{i v^T v/2}.$$  

---

$^{12}$Section 23.2.3 found a formula for $\gamma_s$ given the index of refraction of the medium (here water).

$^{13}$This fact also came up when classifying quadrupoles (Sections 3.6.8, page 44 and 17.4.1, page 254).
But the determinant of an orthogonal matrix is $\pm 1$, so Equation 23.18 is just the product of $N$ ordinary Fresnel integrals. Moreover, the product of the diagonal elements of $D$ is the determinant of $T$, so

$$I = (1 + i)^{N_+} (1 - i)^{N_-} |\det T|^{-1/2},$$

(23.19)

where $N_\pm$ are the number of positive and negative eigenvalues of $T$ respectively (see Equations 22.1).

For a phase function $\phi$ that is not quadratic, we can now repeat the logic of Section 22.3.2 (page 345). The Taylor expansion now involves the matrix of second derivatives of $\phi$ near a stationary point $u_\pm$ (also called the hessian matrix). Stationary phase approximation is the result of replacing the integral of $e^{i\phi}$ by $e^{i\phi(u_\pm)}$ times Equation 23.19 (or by a sum of such terms, one for each isolated stationary-phase point).

Similarly to Section 22.3.3, however, it can happen that the hessian matrix at a stationary-phase point has one or more zero eigenvalues, and so the expression in Equation 23.19 is infinite. Indeed, Equation 23.17 shows explicitly that at the caustic, one eigenvalue of the hessian matrix is zero, so $(\delta_\Theta, \delta_\alpha, \delta_\beta)$ is a superstationary phase point. The corresponding eigenvector has $\delta_\alpha = \delta_\beta = -\delta_\Theta$.

We can now say more precisely what is happening at the caustic angle. The approach to optical problems in Chapter 21 implicitly identified the stationary-phase path and approximated the Huygens integral as a many-dimensional Fresnel integral about that path.\(^\text{14}\) This approach breaks down at the caustic, because in one integration direction the integrand is constant when we truncate to second-order variations: The corresponding Fresnel integral appears to be infinite!\(^\text{15}\) But this infinity is a breakdown of the approximation scheme, not a real physical phenomenon;\(^\text{16}\) it merely says that for $\gamma$ at (or near) the caustic value, we cannot stop with $\gamma^{[2]}_{\text{eff}}$. Certainly $\gamma_{\text{eff}}$ has no reason to be constant at cubic and higher order, as in the analogy of Figure 23.3b.\(^\text{17}\) In short, at and near the caustic angle we must proceed without using stationary-phase approximation.

Section 23.2.6 (page 381) and Problem 23.5 outline how to substantiate the qualitative picture without the numerical integration used in the next section.

### 23.2.7 Reduction to a single integral

It may now seem that the calculation we must do is quite complicated, but the preceding discussion points us toward a great simplification: Now that we have found the one “dangerous” (flat) direction in the space of $\Theta, \alpha,$ and $\beta$, we will replace the full three-dimensional integral by a single integral along that line. However, we will make no other use of the Taylor expansion that revealed the dangerous direction (Equation 23.17): The integral we will perform will involve the exact phase function from Section 23.2.2.

\(^\text{14}\)See Section 22.4.1 (page 346).
\(^\text{15}\) Even when we acknowledge that the ranges of integration for the angles $\Theta, \alpha,$ and $\beta$ are finite, nevertheless they are large enough to invalidate the quadratic truncation for $\gamma_{\text{eff}}$.
\(^\text{16}\) See Section 22.3.3 (page 346).
\(^\text{17}\) You’ll look more closely in Problem 23.5.
Chapter 23  Diffraction at Caustics

Specifically, for $\gamma \leq \gamma_*$ we numerically integrate $\exp(2\pi i L_{\text{eff}}(\theta, \alpha, \beta; \gamma)/\lambda)$ using the
exact $L_{\text{eff}}$ (not its quadratic truncation) along the line that passes through $(\theta_*, \alpha_*, \beta_*)$ and
is directed along the singular direction found in Equation 23.17. Because the other
two eigenvectors of $L_{\text{eff}}^{[2]}$ are nonzero, we may hope that the remaining two “safe” integrals
(in directions transverse to the dangerous one) will contribute a roughly constant
factor, which contributes to the overall normalization but is nonsingular.\(^{18}\) Even though
the phase is superstationary in the special direction, Figure 23.4b shows that the inten-
sity not only is infinite, but in fact is not as large as at another point nearby [panel (c)].
Indeed, experimentally the observed peak of intensity is slightly above the caustic angle
(Figure 23.7).

For $\gamma > \gamma_*$, there are two stationary-phase points in $\theta \alpha \beta$ space. In this regime, we
do a one-dimensional integral along the line that passes through those two points.
That line coincides with the “dangerous” line when we approach $\gamma_*$ from above, so our
calculation smoothly joins the two regimes.

In short, although the Huygens principle instructs us to do several integrals, we have
approximated the answer by a constant times a single integral, which we will perform
numerically.

23.2.8 Other simplifications

Our derivation is restricted to the short-wavelength limit,\(^{19}\) for example, because we used
the Huygens principle. Because our limited goal is just to understand behavior close to
the caustic angle, we have also assumed that the light intensity pattern is dominated by
the presence of one or two stationary-phase points, or at least a nearly stationary point.

Thus, the integral will be dominated by narrow ranges of the angles $\theta, \alpha,$ and $\beta$ close
to the stationary-phase values. Accordingly, we now make the further approximation of
treating other factors, which are slowly-varying functions of the angles, as approximately
constant. As such, those factors just contribute to overall normalization. Generally in
experiments the total illumination intensity is not measured, so we may drop all such
factors and normalize the predicted intensity curve arbitrarily at the end.

The near-constant, nonsingular factors include:

- All of the $1/R$ factors arising from the Huygens principle, analogous to the one in
Equation 22.11 (page 352).\(^{20}\)
- All of the cosine factors arising from the Huygens principle, analogous to the one in
Equation 22.7 (page 348).
- The transmission and reflection factors associated to all three encounters with the
air–water interface.\(^{21}\)

---

\(^{18}\)Actually, the entry, reflection, and exit points can vary out of the cross-sectional plane in Figure 23.2 (page 364), and so there are really five “safe” directions whose integrations we are neglecting.

\(^{19}\)Hence our calculations are not applicable to scattering from extremely small droplets. Indeed, such scattering does not yield a colorful rainbow, but rather a white “fogbow.”

\(^{20}\)It’s not hard to include these and confirm that they don’t change the results significantly.

\(^{21}\)In a fortunate circumstance, the reflection factor is nearly zero for one polarization of light, further justifying
23.2 The Semiclassical Rainbow

Figure 23.6: **Primary rainbow** from a water droplet of radius 0.16 mm for light of wavelength 400 nm (solid curve) and 650 nm (dashed curve). Each curve was separately normalized to peak at 1. The two curves differ both because of the wavelength dependence of interference and also due to dispersion (analogous to the prism, Figure 21.7, page 314). Comparing the main peaks shows that the red part of the primary bow is predicted to be at slightly smaller scattering angle than the blue, as seen in Figure 21.8 (page 317). Moreover, red’s first supernumerary peak (asterisk) lies beyond blue’s main peak (dagger).

- Integrals over the “safe” directions, which Section 23.2.7 already suggested contribute a nearly constant factor over the limited ranges of interest.

23.2.9 Expectations and results

Our goal was to overcome the failures of interfering-ray optics. Certainly we may expect that diffraction will smooth out singular concentrations geometrically predicted to arise at the caustic. But there is a more subtle failure pointed out (and resolved) by G. Airy. Airy found that above the caustic angle, the two interfering rays have an “extra” relative phase factor of $e^{i\pi/2}$. This relative factor, not known to Thomas Young, may not be dropped, unlike the common factors mentioned in Section 23.2.8. Neglecting it would alter the interference, invalidating the predicted positions of the supernumerary bows.

Actually, there is nothing ad hoc or “extra” about this factor. Whenever a cubic function develops two stationary points, Figure 23.3b shows that one is concave-up, the other concave-down. Equation 23.19 (page 371) then gives a relative factor of $(1 + i)/(1 - i) = i$ to the respective contributions of these points. The argument of $i = e^{i\pi/2}$ is Airy’s $\pi/2$ phase. One benefit of evaluating our integral numerically, instead of by stationary-phase, is that we need not put in this factor by hand.

Figure 23.5 shows the modulus squared of the predicted electric field amplitude at various values of scattering angle. Labeled points correspond to the examples shown in Figure 23.4. Figure 23.6 compares the illumination patterns at two different wavelengths, and illustrates how one color’s first supernumerary can lie beyond another’s principal bow, and hence both can be separately visible.

Airy carried out a related analysis in 1838, though from a different viewpoint.\(^\text{22}\) He concluded, as we did, that diffractive scattering near the caustic was dominated by an

---

\(^{22}\)Thus, Airy’s work actually predated the Maxwell equations! He based his calculation on imagery involving “wavefronts”; he had to guess at the amplitude distribution along a chosen initial wavefront.
oscillatory integral whose phase is a family of cubic functions. His expression for the intensity as a function of the control parameter is now called the Airy integral.\textsuperscript{23}

In a natural (meteorological) setting, various real-world complications may obscure the supernumeraries. For example, one must average over the distribution of droplet sizes (and shapes); too wide a range will blur the supernumeraries together. Also, as mentioned the droplet size must be small enough to place the first supernumerary outside all the main bands for other colors, so that it doesn’t overlap them, yet also large enough that the first supernumerary is not so broad (and hence faint) as to be swamped by background light. Finally, the Sun is hardly a point source of light at infinity: Its angular width is 0.5°. Our intensity predictions must also be blurred over that range before they can be compared to observation. Finally, to predict perceptual color, the full predicted spectrum of light at each angular direction must be fed into the wavelength sensitivity functions of our eyes’ three classes of photoreceptors. The resulting three neural activation levels determine the perceived color at that point in the sky.

23.2.10 Comparison with experiment

W. Miller tested Airy’s predictions in 1841, using cylindrical streams of water of various diameters. He observed up to 30 supernumeraries (appearing in this situation as bars of light) associated to the primary bow, up to 25 associated to the secondary, and so on, and reported agreement with Airy’s predictions. Figure 23.7 shows more modern data, in which not only the positions but also the shapes and relative heights of each peak were measured. Although our calculation was not expected to be very accurate away from the primary maximum, due to the many angle-dependent factors we omitted, nevertheless in the region relevant for rainbows it proves to be remarkably successful.

Section 23.2.10' (page 381) compares our semiclassical results with more detailed calculations.

\textsuperscript{23}See Problem 23.1. Confusingly, the Airy function to be introduced in Section 23.3.3 refers to a different concept.
23.2.11 Compare and contrast with knife-edge diffraction

Figures 22.1 and 23.7 look superficially similar, but there is a key difference: Knife-edge scattering has only partial modulation in the bright region, whereas in rainbow scattering the illumination falls to zero between the supernumeraries. This is because the supernumeraries are not diffractive at all; instead, they show the interference of just two rays. In contrast, Section 22.4.5 (page 352) identified the knife-edge fringes as arising from an incomplete Fresnel integral.

Both situations do, however, share a common origin for the “forbidden” light: A Fresnel-type integral can be nonzero even if it has no stationary-phase point, either because of a restricted range of integration (in the knife-edge case) or because the integrand does come close to leveling off (in the rainbow case). In the rainbow, this phenomenon appears even with no opaque obstruction: All that is needed is a forbidden region, here supplied by the caustic.

23.3 LENS FOCUSING, POINT CAUSTICS, AND THE RAYLEIGH CRITERION

23.3.1 Geometrical optics of 2D lens focusing

So far, this chapter has studied caustics as a generic phenomenon. For example, a rain droplet subjected to uniform illumination led to a bright ring regardless of the distance to the observer. Similarly, random ripples on the surface of a swimming pool focus uniform sunlight to a network of bright lines on the bottom of the pool, regardless of its depth. Let us now turn to more carefully engineered devices, see what other caustics may arise, and again explore the role of diffraction.

When light passes through an aperture, geometrical optics predicts that the resulting illumination forms a shadow image. The most extreme deviation we could imagine is for all the light to instead arrive at one point on a distant screen. Is this possible?

Figure 23.8 shows rays arriving at a barrier from a distant source. The barrier has an aperture, but we have imagined filling it with an unspecified thin, transparent object. This object modifies the phase of the light by an amount that depends on distance \( u \) from the centerline. For example, it could be a flat plate with variable refractive index,\(^\text{24}\) a nonflat plate of ordinary glass, or a combination of both. We have already seen that geometrical optics predicts focusing in such situations.\(^\text{25}\)

To go beyond a geometrical optics treatment, we now modify the Huygens formula, Equation 22.7 (page 348), by introducing an extra phase to account for the slowing of light by the medium (Equation 22.10, page 350). To bring out the essential features, we will consider a limiting case (which is realistic for many situations):\(^\text{26}\)

\(^{24}\)A gradient-index, or “GRIN,” lens (Problem 21.5, page 337).

\(^{25}\)See Figure 21.6 (page 313) and Figure 21.15 (page 330).

\(^{26}\)You’ll explore more general situations in Problems 23.2, 23.3, and 23.4.
Figure 23.8: [Path diagram.] **Lens-based focusing.** In this cross section, a distant point source of light lies on the centerline (dotted), that is, the line perpendicular to the opaque barrier (middle) and passing through the center of an aperture of width \(W\). The aperture contains a thin, transparent element (the lens, represented by a question mark), which changes the phase of each light path passing through it. Two typical paths are shown, each passing through the lens at a distance \(u\) from the center and ending at observation point \(B\) on the projection screen, a distance \(|x|\) below that line (here \(x\) is a negative quantity).

- We assume that the source is at infinity, so that all path segments on the left side of Figure 23.8 are of equal length.
- Initially, we will work only in the two-dimensional plane shown in the figure. (A real situation of that sort could be an infinite slit aperture containing a cylindrical lens.)
- We also simplify by considering the common situation where \(d \gg W\) (narrow aperture) and \(d \gg |x|\) (small angle), so that we may Taylor expand to get

\[
R(u) = \sqrt{d^2 + (u - x)^2} \approx d + (u - x)^2/(2d) + \cdots.
\]

In this limiting case, the \(1/R\) factor in the Huygens formula equals the constant \(1/d\) plus higher-order terms (suppressed by more powers of \(1/d\)), which we drop.

The phase factor is \(e^{i(2\pi/d)(R + \Delta t(u))}\), where \(\Delta t(u)\) is the extra delay introduced by the lens. Equation 23.20 shows that the condition for stationary phase is

\[
\frac{u - x}{d} + c \frac{d\Delta t}{du} = 0.
\]

Let us simplify further by adding two more specializations representing common situations:

- Assume that the lens is symmetric about the centerline. Then \(c\Delta t(u)\) has a Taylor expansion with only even powers.
- Assume that the lens is only slightly curved, so that we may drop higher terms in

\[
c\Delta t(u) = A - \frac{1}{2f} u^2 + \cdots.
\]

Here \(A\) and \(f\) are constants with units of length that describe relevant features of the lens’s shape and composition. We will soon see why \(f\) is called the lens’s **focal length**. The sign
convention is such that positive \( f \) corresponds to a convex lens, that is, one whose time delay decreases as we move outward from the center.

The stationary-phase condition then says
\[
\frac{u - x}{d} - \frac{u}{f} = 0.
\]

If we set the screen distance \( d \) equal to the lens characteristic \( f \), then there is a dramatic bifurcation: For observation point \( x = 0 \), every path is of stationary phase, whereas for \( x \neq 0 \) no path is of stationary phase. Geometrical optics then predicts no illumination at \( x \neq 0 \), but infinite intensity at \( x = 0 \), as all the light passing through the aperture gets concentrated to the focus. We also say that \( x = 0 \) is the image point of the point source at infinity.

Lens focusing is both more fragile and more extreme than the rainbow caustic:

- **Fragile**: Instead of one condition \( \gamma = \gamma_s \), we now have two that must be satisfied: \( d = f \) and \( x = 0 \). That is, the caustic occupies a curve in space, not a surface.
- **Extreme**: Instead of a superstationary phase path where the quadratic bit of the phase variation varies but cubic and higher terms remain, our assumed \( u \rightarrow -u \) symmetry forbids cubic terms; the leading term at the caustic is therefore quartic. Moreover, in the limiting case we have studied, even the quartic and higher terms are negligible, so the phase was exactly constant throughout the integration region, leading to a huge intensity at \( x = 0 \).

**Your Turn 23B**

Discuss how the situation changes when the system is defocused, that is, when \( d \neq f \).

### 23.3.2 Diffraction blurs the point focus

Prior experience leads us to suspect that the behavior just outlined may be an artifact of geometrical optics approximation, which also falsely predicted infinite intensity at the rainbow caustic, again because of the appearance of a superstationary point. In the rainbow case, the intensity exactly at the caustic was finite essentially because the cubic and higher terms in Equation 23.6 made the integral still oscillatory, and hence convergent (Figure 23.4b). Although such cubic terms are negligible in the present situation, still our integral is cut off by its finite range, which is set by the aperture width \( W \). Thus, the Huygens integral has a finite integrand integrated over a finite region; let us just carry it out in the same large-\( d \) approximation we have been using so far:

\[
\int_{-W/2}^{W/2} du \exp((2\pi i/\lambda)((u - x)^2/(2d) - u^2/(2f))).
\]

\(^{27}\)When those terms may not be neglected, they give rise to spherical aberration. See Sections 21.3.7 and 21.3.5.
Setting \( d = f \) makes the \( u^2 \) term in the exponential cancel, but there remain a linear term, \((2\pi/\lambda)(-ux/d)\), as well as a \( u \)-independent term, \((2\pi/\lambda)x^2/(2d)\). We may pull the exponential of the latter outside the integral, and then drop it because its modulus is one.

**Your Turn 23C**

Show that the modulus-squared of the integral may be written as

\[
\text{const} \times \frac{\sin^2 \bar{x}}{\bar{x}^2},
\]

where \( \bar{x} = x/(\lambda d/(\pi W)) \). Make a graph of this function. Find the range of \( \bar{x} \) where it exceeds half its maximal value.

Your result is called the **point spread function** of this simple imaging system, because it is the response to a point source. Any imaging system, not necessarily subject to the simplifications in this example, has its own characteristic point spread function.

We conclude that even for a perfect lens, even for monochromatic light (and hence no chromatic aberration), even in the large-\( d \) limit (and hence no spherical aberration), even exactly at the focal plane \( d = f \), nevertheless **diffraction spreads the image over a width** \( \Delta x \approx 2.8\lambda d/(\pi W) \). If an observer were to interpret the illumination pattern naïvely via geometrical optics, they would falsely conclude that the point source of light was spread over an angular size \( \Delta \theta \approx \Delta x/d \).

We have found a fundamental limitation on lens imaging. When generalized to sources that are not at infinity, and to more complex lens systems (including those with curved mirrors instead of lenses), it applies to microscopes, telescopes, and even radio telescopes. Stated differently, it also implies that

Two distant point sources whose angular directions are spaced more closely than \( \sim \lambda/W \) will give overlapping point spread functions, and hence will not be resolvable. \hspace{1cm} (23.23)

Despite some superficial differences, Equation 23.22 illustrates themes we have already encountered in the context of knife-edge and rainbow diffraction:

- Light appears where geometrical optics says it cannot, due to a nearly-stationary phase path.
- Light intensity is finite at the focus where geometrical optics says otherwise, because stationary-phase approximation breaks down at a caustic.

### 23.3.3 Extension to a circular lens

Although we illustrated the calculation in two dimensions, it works similarly in 3D, for example, for a lens that has spherical (not cylindrical) surfaces. In this situation, two eigenvalues of the second-order variation of effective path length both become singular at the focus, and so the caustic occupies one point.\(^{28}\) The point spread function that

\[^{28}\text{Problem 23.3}\]
corresponds to Equation 23.22 is called the Airy function, two images of nearby point light sources are deemed to be resolvable if their angular separation is

$$\Delta \theta \geq 1.2\lambda / W.$$  

the Rayleigh criterion

### 23.4 PLUS ULTRA

Nature uses only the longest threads to weave her patterns, so that each small piece of her fabric reveals the organization of the entire tapestry.

— Richard Feynman

The examples we studied all involved piecewise-uniform objects (sharp boundaries), but gradient-index materials can also form diffracting caustics.

The rainbow reminds us that mathematical concepts from one domain can be unexpectedly relevant in an apparently distant context: For example, the mathematical ideas of bifurcation from dynamical systems, and semiclassical approximation from quantum mechanics, proved to be key to unraveling the structure of diffraction at the rainbow caustic, and then generalized to lens focusing.

Even at the level of raw phenomena, we may find something salient (a beautiful rainbow), think about it, and extract a more general class of phenomena (caustic focusing) that we had not previously thought to explore systematically. For example, nonuniformity in atmospheric temperature leads to small nonuniformities of the index of refraction that shift constantly over time. These in turn lead to caustic surfaces in the light arriving from a distant star. Each time such a surface sweeps across our eye, we experience a momentary sharp increase in the star’s apparent brightness, which we call the “twinkling” of stars.

Even high energy physics turned out to have many examples of scattering amplitudes whose angular dependence is essentially similar to the rainbow calculation.

### FURTHER READING

**Intermediate:**

**Technical:**

Many of this chapter’s ideas are concrete, specific applications of a general theory developed in Berry & Upstill, 1980; some aspects are summarized in Berry, 2015a.

---

29 The Airy integral in Section 23.2.9 is a different concept.
30 Section 54.3.4 will mention an astrophysical analog of this phenomenon.

23.2.2’ Catastrophe theory
A wide range of phenomena involve bifurcations of the stationary points of a function. For example, each curve of Figure 23.3b could represent the potential energy of a mechanical system; as the control parameter changes, the stable minimum disappears altogether. A branch of math called catastrophe theory classifies such bifurcations (which could indeed have catastrophic consequences) into universal (structurally stable) classes.

The rainbow caustic is an avatar of the simplest such degeneration (the “fold catastrophe”), represented by the family of functions \( f(u, C) = u^3 + C u \) (Figure 23.3b with \( C = \gamma - \gamma_* \)). A handful of other, more complex, catastrophes exist, each with its own beautiful diffraction patterns described by a low-dimensional integral generalizing the one we studied.

23.2.6’ Reduction of the rainbow integral to a cubic phase function
Section 23.2.3 (page 367) suggested that the phase function at the caustic angle would be purely cubic in excursions from the stationary phase point, and that as we move away from the caustic angle, new linear terms would arise, leading to the scenario in Figure 23.3 (page 367). In Problem 23.5 you’ll substantiate these claims by explicitly finding the linear and cubic contributions to the Taylor expansion.

23.2.10’ High-tech rainbow calculations
Figure 23.9 compares the semiclassical calculation from this chapter to much more elaborate, and accurate, strategies.
Figure 23.9: Comparison to partial-wave and other calculations. Solid blue curve: Semilog plot of the complex amplitude squared, calculated by the method of the main text (scalar diffraction, semiclassical approximation) and evaluated with $n = 1.33$ and droplet radius equal to $1500\lambda / 2\pi$. Solid red: Airy’s approximation. Other curves: The same quantity evaluated by more detailed methods, for the dominant polarization of light, excluding front-surface reflection. Dashed: Solution to the Maxwell equations, in a multipole-type expansion scheme (Mie theory). For spheres much larger than the wavelength of light, the sum over all contributions converges slowly; several thousand terms must be kept. Dotted: Evaluation via analytic continuation of the Mie solution to complex angular momentum. [Last three curves: Data from Khare & Nussenzveig, 1974.]

### PROBLEMS

23.1 Light in Alexander’s darkness

Mathematicians define the **Airy integral** function as

$$\text{Ai}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} du \, e^{iu^3/3 + xu}.$$  

The main text suggested that the modulus squared of this function could approximately represent some features of rainbow scattering.

a. Get a computer to evaluate this integral numerically via a simple Riemann sum. That is, take a finite but large range, such as $-50 < u < 50$, and subdivide it into intervals of width $\Delta u = 0.001$. For a range of $x$ values $-10 < x < 10$, evaluate the integrand on this grid of $u$ values, add them up, and multiply by $\Delta u$. Then plot the result versus $x$.

b. If your computer math system offers a built-in Ai function, plot it alongside your result.

c. Does your graph resemble any of the ones in this chapter? Discuss the qualitative features of your graphs in terms of the physical ideas in the chapter.

d. Numerical evaluation can be tricky. Because the integral is really over an infinite range, it may seem like a good idea to repeat the calculation with $-200 < u < 200$ and the same $\Delta u$. What goes wrong and why?

e. Just to drive the point home, repeat with $-500 < u < 500$ and the same $\Delta u$. 
23.2 Focusing without Taylor approximation

a. Generalize Section 23.3.1 (page 375) to the case of a source at distance $L_0$ that may not be $\gg d$, though it is still $\gg W$. Hence the geometric path length has variable terms $R_{\text{left}}$ and $R_{\text{right}}$ corresponding to either side of the thin lens. Use the usual Taylor expansion for $R_{\text{right}}$ and something similar for $R_{\text{left}}$. With the same assumption for the lens delay as Equation 23.21 (page 376), derive a famous condition on $L_0, d$, and $f$ to have focusing.

b. Use a computer to make graphs of the exact phase function, without using the Taylor series approximation. Try the values $\lambda = 500$ nm, $f = 0.25$ m, and $L_0 = 1$ m. Make separate graphs of the phase as a function of $u$ for a projection screen at $d_0 = (f^{-1} - L_0^{-1})^{-1}$ or at $d_1 = 1.01d_0$ and at each consider screen position $x = 0$ (the focus) and $x = 0.1$ mm (slightly off the focus). In those four graphs, display a wide enough range of $u$ values to reveal any stationary-phase paths that may exist.

c. Discuss on your result in (b) qualitatively.

23.3 Spherical lens

Upgrade the discussion in the main text from two to three dimensions. That is, allow two transverse directions and replace Equations 23.20 and 23.21 by

$$R(\vec{u}) \approx d + ||\vec{u} - \vec{x}||^2/(2d), \quad c\Delta t = A - ||\vec{u}||^2/(2f).$$

Find the stationary phase point $\vec{u}_s$ for given $\vec{x}, d$, and $f \neq d$. Then expand the phase in Taylor series about $\vec{u}_s$ and comment on the behavior of its quadratic terms in $\delta\vec{u}$ as $d \to f$.

23.4 Stationary and superstationary

If you haven’t done Problem 22.1 yet, do it first. Consider 2D lens focusing, summarized by Equations 23.20 and 23.21. In this problem, use the wavelength $\lambda = 500$ nm, focal length $f = 25$ cm, and aperture $W = 2$ cm. In addition, the distance to the screen is $d = f(1 + \epsilon)$; thus, nonzero $\epsilon$ indicates defocus. Make and interpret Cornu-type curves\(^{31}\) for the following situations:

a. Focused, and we observe at the point $x = 0$.

b. Focused, but we observe at the point $x = 0.1$ mm.

c. Defocused ($\epsilon = 0.2$), and we observe at the point $x = 0$.

d. Defocused, and $x = 0.1$ mm.

e. Defocused, and $x = 1$ mm.

23.5 Cubic phase function

In this problem, you’ll explore the behavior of the phase function near the rainbow caustic.

a. Confirm Equation 23.8 (page 366): That is, for arbitrary $\gamma$ expand $L_{\text{eff}}$ about arbitrary $\theta_0, \alpha_0, \beta_0$ and find the first-order variation.

b. Now specialize by setting $\theta_0 = \theta_*$ and $\alpha_0 = \beta_0 = \alpha_*$, but $\gamma$ not necessarily equal to $\gamma_*$. In fact, in the following take $\gamma = \gamma_* + \epsilon$ and expand only to first order in $\epsilon$.

---

\(^{31}\) Recall Figures 22.4 (page 344) and 23.4 (page 369).
Also substitute $\delta \theta = -x, \delta \alpha = \delta \beta = x$ (the “dangerous” direction discussed in Section 23.2.5, page 368). We already know that $I_{\text{eff}}^{(1)} = 0$ when $\epsilon = 0$; find its $O(\epsilon)$ contribution and comment.

c. Confirm and extend Equation 23.17 (page 370): That is, find the variation of $I_{\text{eff}}$ to second order in $x$ and order $\leq 1$ in $\epsilon$.

d. Similarly, find the variation of $I_{\text{eff}}$ to third order in $x$ and order $\leq 1$ in $\epsilon$. Confirm that it is nonzero even when $\epsilon = 0$, and comment on the relevance of Figure 23.3b.

23.6 *T2* Diffractive rainbow

Figure 23.4 (page 369) shows calculations of the Huygens integral, simplified to a single numerical integral, for scattering angles 137, 138.4, 139.1, and 139.5 deg. Other assumptions were that the droplet diameter was 0.465 mm, wavelength was 633 nm, and refractive index was 1.332. Recreate these figures by the method described in the text, but also try some angles farther above and below the caustic.
CHAPTER 24

Partial Polarization

24.1 FRAMING: STOKES PARAMETERS

We found plane-wave solutions to the Maxwell equations. Each such solution had a single, definite wavevector $k$, and hence a definite frequency: That is, they described monochromatic light, such as might be obtained from a laser. Each also had a single, definite polarization vector. So plane waves are too restrictive to describe light from real sources. For example, natural light is usually unpolarized (like sunlight), or partially polarized (like the blue sky). This chapter will explore a widely-used way to characterize partial polarization, via Stokes parameters.

Electromagnetic phenomenon: For optical purposes, polarization of light may be described by a point inside an abstract sphere.

Physical idea: Optical instruments ultimately measure light intensities after various linear filters have been applied.

24.2 LIGHT AS AN ENSEMBLE

24.2.1 Most sources give chaotic light

A single atom, making a transition between definite states, gives off a pulse of light of finite duration, so it has some spread in frequency. Even a supposedly monochromatic filter transmits a finite range of frequencies. In addition, the superposed light from zillions of independent atoms (for example, in the Sun) will be a jumble of many polarizations. To model such light classically, we now consider a superposition of plane waves in a narrow but finite range of frequencies. Assuming for simplicity that each wave is traveling in the same direction $\hat{z}$, such a superposition may look like

$$\vec{E}(t, \vec{r}) = \frac{1}{2} \vec{E}(t) e^{-i\omega(t-z/c)} + \text{c.c.} \quad (24.1)$$

In this expression, $\vec{E}(t)$ is the sum of the profiles of many pulses, transverse to $\hat{z}$. Because we pulled out the mean frequency, $\vec{E}$ varies more slowly in time than $\vec{E}$. Each pulse may have a phase shift relative to the others, and each may be polarized in a different way.

24.2.2 Optical instruments ultimately measure energy deposition

In practice, optical instruments in millimeter wavelength and shorter don’t measure the detailed time dependence of the electric field. They measure averages over a time that’s

---

1 Radiotelescopes can in principle measure this, and so pick up more detailed information about the waves they detect than instruments like bolometers or cameras.
long compared to the time scale over which $\vec{E}$ varies, and hence also much longer than $2\pi/\omega$.

Moreover, most optical detectors measure only the time average of energy flux delivered by a light source. We may place various filters between the source and detector, to restrict to various polarization or frequency ranges, but ultimately what’s measured are energy fluxes of the filtered lights. Section 20.2.1 (page 299) argued that energy flux is a constant times the square of the electric field of the (possibly filtered) light.

In most optics applications, the filters we might use generally perform linear operations. For example, an ideal color filter may extract a single frequency component of $\vec{E}$. A polarizer applies a linear operator that doesn’t depend (much) on frequency, but that has one eigenvalue much smaller than the other one (high absorption for one polarization), and so on.

The preceding logic implies that, in optics, anything we can really measure via a filter/detector combination can be extracted from twelve time-averaged quantities:

$$
\langle \vec{E}_i e^{-i\omega t} \vec{E}_j e^{+i\omega t} \rangle, \text{ their conjugates, and } \langle \vec{E}_i e^{-i\omega t} \vec{E}_j^* e^{+i\omega t} \rangle \text{ where } i, j = 1, 2.
$$

Of these, the first eight average to zero because of their fast time variation.

The remaining four quantities constitute a $2 \times 2$ hermitian matrix:

$$
\vec{I}_{ij} = \langle \vec{E}_i \vec{E}_j^* \rangle \text{ or } \vec{I} = \langle \vec{E} \otimes \vec{E}^* \rangle.
$$

Although this matrix does not contain enough information to reconstruct $\vec{E}$ completely, it does characterize a beam of nearly monochromatic light well enough to specify what it will do when it passes through linear optical elements and lands on an intensity detector.

The most general $2 \times 2$ hermitian matrix can be written in terms of four real quantities. A traditional choice is to introduce the four Stokes parameters:

$$
\vec{I} = \frac{1}{2} \begin{pmatrix} s_0 + s_1 & s_2 - is_3 \\ s_2 + is_3 & s_0 - s_1 \end{pmatrix}.
$$

Again: The Stokes parameters describe light for the purposes of detectors of the sort used in most optics experiments.\(^3\) Note that

$$
\det \vec{I} = (s_0^2 - s_1^2 - s_2^2 - s_3^2)/4.
$$

Thus, for given $s_0$ the remaining Stokes parameters must lie in the region $(s_1^2 + s_2^2 - s_3^2) \leq (s_0)^2$, called the Poincaré sphere for the given overall intensity.

\(^2\)The magnetic field of a plane wave just tracks the electric field, so we would learn nothing more by considering terms with $\vec{B}$.

\(^3\)Some books factor out the overall normalization and define Stokes parameters as $\xi_1 = s_2/s_0$, $\xi_2 = s_1/s_0$, $\xi_3 = s_3/s_0$.\(^4\)
24.2.3 Steady sources: Replace time average by ensemble average

Much as in equilibrium statistical mechanics, we can introduce a notion of steady light source, in which time averages are replaced by ensemble averages over a probability distribution of electric field vectors. In that language, we propose a classical model of unpolarized light in which the two complex coefficients $\tilde{E}_1$ and $\tilde{E}_2$ are random variables that are as uncorrelated as possible, subject to having a specified mean intensity. That is, their probability distribution will take the form:

$$\varphi(\tilde{E}_1, \tilde{E}_1^*, \tilde{E}_2, \tilde{E}_2^*) = f(||\tilde{E}||^2).$$  \hspace{1cm} \text{unpolarized light} \hspace{1cm} (24.5)

Here the length-squared, $||\tilde{E}||^2$, of a complex vector is understood to mean $\sum_i \tilde{E}_i \tilde{E}_i^*$. The real function $f$ may be chosen such that $\langle ||\tilde{E}||^2 \rangle$ gives the desired intensity; for example, that appropriate to a thermal radiation spectrum at some temperature and the wavelength under consideration.

When we substitute Equation 24.5 into the definition Equation 24.2, we find

$$\bar{I} = \frac{1}{2} V \int_0^\infty dx f(x^2) x^2. \hspace{1cm} \text{unpolarized light} \hspace{1cm} (24.6)$$

In this expression, $\bar{I}$ is the identity tensor in the 2D space of transverse directions and $V$ is the volume of the 3D surface of the unit sphere in four dimensions. For our purposes, the main point is that $\bar{I}$ is proportional to the identity matrix. For example, $I_{12}$ is zero by the invariance of Equation 24.5 under reflections in $y$. Also, symmetry under exchange of $x$ and $y$ gives $I_{11} = I_{22}$. Thus, unpolarized light sits at $s_1 = s_2 = s_3 = 0$.

Note that the distribution Equation 24.5 contains all polarizations, including all linear polarizations, both circular polarizations, and all the elliptical polarizations in between. The distribution takes the same form if we rotate in the $xy$ plane; or if we re-express the fields in a circular-polarized basis; or indeed if we perform any other unitary change of polarization basis. For light that is also chaotic in direction, for example thermal radiation in a cavity, we can further average the ensemble over uniformly distributed rotations of the direction of propagation and the polarization vector $\tilde{E}$.

We can then think of partially polarized light as having a more informative distribution of polarization vectors than Equation 24.5, and fully polarized light as the extreme case where the distribution is a delta function selecting some definite $\tilde{E}$.

24.3 SOME CONVENIENT MODELS OF LIGHT

24.3.1 Fully polarized light corresponds to the periphery of the Poincaré sphere

Note that the average of a product is not in general the same as the product of the corresponding averages. So although $\bar{I}$ is the average of a dyad product, still it need not itself be expressible as such a dyad. If, however, the light in question is truly monochromatic, then $\tilde{E}$ is a single complex vector, we may drop the averages, and so we do have a dyad.
Your Turn 24A

For such a wave traveling along \( \hat{z} \), substitute the complex polarization vector \( \vec{E} = A\hat{x} + B e^{i\delta}\hat{y} \) into the definition of \( \vec{I} \) and find the Stokes parameters\(^4\) in terms of \( A \), \( B \), and \( \delta \). The determinant of a dyad product always equals zero; confirm that your answer has that property.

Thus, for fully polarized light \( s_1, s_2, \) and \( s_3 \) always sit on the surface of a sphere of radius \( s_0 \) (see Equation 24.4).

Your Turn 24B

Comment on what parts of the Poincaré sphere surface correspond to linearly polarized light, and what parts to circular polarization.

Beware: Although we speak of the Stokes parameters \( s_1, s_2, \) and \( s_3 \) as lying on a sphere, they do not constitute a “vector” in the sense of pointing somewhere in ordinary 3-space. That is, they do not define a rank-one 3-tensor.\(^5\) The Poincaré sphere is an abstract, though sometimes useful, representation of \( \vec{I} \), a complex, rank-2, 2D tensor.

24.3.2 Simplified model of unpolarized light

Section 24.2.3 showed that unpolarized light gives rise to the opposite extreme situation. A simpler realization than the one given there is often helpful, however. Consider an ensemble of \( \vec{E} \) vectors that are each linearly polarized, with directions that are uniformly distributed over the circle perpendicular to \( \hat{z} \). For simplicity, assume that each vector has the same amplitude \( A \). Then

\[
\vec{I}_{11} = \langle A \cos \theta A \cos \vartheta \rangle = \frac{1}{2} A^2,
\]

(24.7)

\[
\vec{I}_{12} = \langle A \cos \theta A \sin \vartheta \rangle = 0,
\]

(24.8)

and so on. Thus,

\[
\vec{I} = \frac{A^2}{2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},
\]

(24.9)

so this ensemble indeed serves as a model of unpolarized light. Although not as complete as Equation 24.5 (we omitted circular and elliptically polarized states), this model is easy to think about and equivalent if we restrict to the limited measurements outlined in Section 24.2.2.

24.3.3 Partial polarization

The limiting cases just discussed motivate us to define the degree of polarization as \((s_1^2 + s_2^2 + s_3^2)/s_0^2\). It ranges from zero (unpolarized) to one (fully polarized).

---

\(^4\) \( \vec{E} \) is sometimes called the Jones vector. The tensor \( \vec{I} \) discards any overall phase, so we don’t need to give \( A \) and \( B \) separate phases.

\(^5\) Nor do the full set of four Stokes parameters constitute a 4-vector!
24.4 HOW TO MEASURE THE STOKES PARAMETERS

It’s straightforward to measure $s_0$, because it’s a constant times the total intensity (energy flux) of the light.

To see how to measure the others (and indeed, why they are needed), let’s first think about the sorts of filters that we could apply to a light source. Section 24.2.2 pointed out that an ideal polarizing filter performs a projection on the electric field, that is, the operation $\vec{E} \rightarrow \hat{\vartheta} (\hat{\vartheta}^* \cdot \vec{E})$. Then the corresponding transformation on the polarization tensor $\vec{I}$ is

$$\vec{I} \rightarrow \{ \hat{\vartheta} (\hat{\vartheta}^* \cdot \vec{E}) (\vec{E}^* \cdot \hat{\vartheta}) \} = (\hat{\vartheta} \otimes \hat{\vartheta}^*) \cdot (\vec{E} \cdot \vec{E}^*) .$$

**Your Turn 24C**

a. Consider the case of a linear polarizer, that is, $\hat{\vartheta} = [\begin{array}{c} 1 \\ 0 \end{array}]$, acting on unpolarized light. Interpret the new polarization tensor.

b. Repeat for a circular polarizer.

Think about how applying various filters to an arbitrary $\vec{I}$, then finding the intensity of the filtered light, lets us deduce the various matrix elements of $\vec{I}$, and hence the Stokes parameters.

**FURTHER READING**

*Intermediate:*

*Technical:*
Thompson et al., 2017.
the $xy$ plane, and describe that ellipse. That is, give its semimajor and semiminor axes, and the angle that the semimajor axis makes with the $x$ axis.

b. Repeat with $\{s_\alpha\} = 25, 0, 24, 7$. 
CHAPTER 25

Generation of Radiation: First Look

If with the aid of our electric waves we can directly exhibit the phenomena of light, we shall need no theory as interpreter; the experiments themselves will clearly demonstrate the relationship between the two things. As a matter of fact, such experiments can be performed.

— Heinrich Hertz, 1889

25.1 FRAMING: SLOW Falloff

Section 18.8.2 (page 281) formulated the Maxwell equations in terms of potentials, then specialized to the situation where the vector potential satisfied \( \vec{\nabla} \cdot \vec{A} = 0 \) (Coulomb gauge):

\[
\nabla^2 \psi = -\frac{\rho_q}{\varepsilon_0} \quad (25.1)
\]
\[
\nabla^2 \vec{A} - c^{-2} \left( \frac{\partial^2 \vec{A}}{\partial t^2} + \vec{\nabla} \cdot \frac{\partial \vec{A}}{\partial t} \psi \right) = -\mu_0 \vec{j}. \quad (25.2)
\]

To keep things simple, this chapter will assume that the charge density is everywhere zero. In Your Turn 18F, you showed that in this case, we may also assume \( \psi = 0 \).

However, we’ll now allow regions in space where the charge flux \( \vec{j} \neq 0 \). The continuity equation requires that \( \vec{\nabla} \cdot \vec{j} = 0 \), but this can be satisfied, for example, by having current in a closed loop of wire that is uniform along the wire’s length. Equation 25.2 reduces to three decoupled copies of the D’Alembert equation:

\[
\nabla^2 \vec{A} - c^{-2} \frac{\partial^2 \vec{A}}{\partial t^2} = -\mu_0 \vec{j}. \quad \text{Coulomb gauge, no net charge} \quad (25.3)
\]

In empty space, we found some simple solutions to this equation: the plane waves. But certainly empty space could instead contain no radiation (fields everywhere zero). We’ll now see how, in the presence of accelerating charges, electromagnetic waves are obligatory. Moreover, we’ll see that they exhibit slow falloff with distance, compared to the fields of analogous static charge or current arrays.

Electromagnetic phenomenon: An antenna emits energy with a specific directional pattern.
Physical idea: Far from the source, the fields must be transversely polarized, and this condition depends on angle.

---

1 Sometimes called the inhomogeneous wave equation.
25.2 REVIEW: GREEN FUNCTION SOLUTIONS TO ELECTRO- AND MAGNETOSTATICS

We already encountered the special case of Equation 25.3 in which the charge flux \( \vec{j} \) is time independent. In that case, we had three independent (decoupled) copies of the Poisson equation, each of which had the same solution as in electrostatics:

\[
\vec{A}(\vec{r}) = \frac{\mu_0}{4\pi} \int d^3 r_s \frac{\vec{j}(\vec{r}_s)}{||\vec{r} - \vec{r}_s||^3}, \quad \text{static case} \quad \text{[15.18, page 229]}
\]

Chapter 15 called this expression the Green function solution to the Poisson equation. As usual, call \( \vec{r} \) the “field point” and \( \vec{r}_s \) the “source point.” Also define \( \vec{R} = \vec{r} - \vec{r}_s \), and denote its length by \( R \) (no arrow). Then the function \( G(\vec{r}, \vec{r}_s) = (4\pi R)^{-1} \) is the Green function of the operator \(-\nabla^2\).

We’d like to find a similar solution for the time-dependent case.

25.3 A PHYSICALLY MOTIVATED TRIAL SOLUTION FOR THE RADIATION GREEN FUNCTION

With time-varying currents, we might expect that the vector potential at a spatial position \( \vec{r} \) would again be determined by currents at \( \vec{r}_s \), with a \( 1/R \) falloff. But we also expect that signals will travel from source point to field point at the finite speed \( c \), and hence \( \vec{A}(t, \vec{r}) \) can only involve the behavior of a source element time in the past. Based on that hunch, a simple trial solution is to generalize Equation 15.18 as

\[
\vec{A}(t, \vec{r}) \sim \frac{\mu_0}{4\pi} \int d^3 r_s \frac{1}{R} \vec{j}(t - R/c, \vec{r}_s). \quad \text{trial solution, Coulomb gauge} \quad (25.4)
\]

In the case of stationary currents, \( \vec{j} \) is time independent, and our guess reduces to the known answer for that case. For time-dependent currents, Equation 25.4 says we must look back in time to the moment \( t - R/c \), when a source point’s current could have influenced our observer’s field point \( \vec{r} \) at time \( t \).

**Your Turn 25A**

Verify that the proposed Equation 25.4 really obeys the Coulomb gauge condition \( \nabla \cdot \vec{A} = 0 \). [Hint: Adapt the approach used in magnetostatics (Section 15.5.4, page 229), and remember that we are still assuming zero charge density everywhere.]

The form of our trial solution suggests part of the answer to Hanging Question #H (page 32): The fields observed at some time \( t \) have nothing to do with the source at that time. We may have turned off the apparatus; a radiating star may have died out; an electron/positron pair may have annihilated by the time radiation gets to our detector. Once

---

\[2\text{See Equation 2.7 (page 31).}\]
formed, radiation proceeds autonomously through space. It reflects only the behavior of currents at the \textit{retarded time} $t - R/c$.

### 25.4 CHECK THE TRIAL SOLUTION

We now apply the \textbf{wave operator} $\Box = \nabla^2 - c^{-2} \partial^2 / \partial t^2$ to our trial solution, to see whether we indeed recover $-\mu_0 j$ (Equation 25.3). The wave operator involves derivatives with respect to the field point, so throughout this section $\tilde{\nabla}$ will denote $\partial / \partial \tilde{r}_i$ (not $\partial / \partial \tilde{r}_{i\ell}$).

**Your Turn 25B**

Show that (or review why)

$$\tilde{\nabla} R = \tilde{R}; \quad \tilde{\nabla} \cdot \tilde{R} = 3; \quad \tilde{\nabla}(R^{-p}) = -pR^{-(p+1)}\tilde{R}; \quad \nabla^2(R^{-1}) = -4\pi \delta^{(3)}(\tilde{R}).$$

[Hint: Recall Section 1.3 (page 18).]

To save writing, let $\phi$ denote any component of $4\pi \tilde{A}/\mu_0$, and $J$ the corresponding component of $\tilde{j}$. So our proposed Green function solution, Equation 25.4, says

$$\phi(t, \tilde{r}) = \int d^3 r_s \frac{1}{\tilde{R}} J(t - R/c, \tilde{r}_s),$$

and we wish to show

$$\nabla^2 \phi - c^{-2} \frac{\partial^2}{\partial t^2} \phi = -4\pi J.$$

The gradient of Equation 25.5 is

$$\tilde{\nabla} \phi = \int d^3 r_s \left[ (\tilde{\nabla}(R^{-1}))J(t - R/c, \tilde{r}_s) - \frac{1}{cR}(\tilde{\nabla}R) \frac{\partial J}{\partial t} \bigg|_{\text{ret}} \right].$$

Here the subscript “ret” means to evaluate at the retarded time $t - R/c$, after taking any indicated derivatives.

The second derivative is then

$$\nabla^2 \phi(t, \tilde{r}) = \int d^3 r_s \left[ (\nabla^2 R^{-1})J \bigg|_{\text{ret}} - c^{-1}(\tilde{\nabla}R^{-1}) \cdot (\tilde{\nabla}R) \frac{\partial J}{\partial t} \bigg|_{\text{ret}} - c^{-1} \tilde{\nabla} \cdot (R^{-1} \frac{\partial J}{\partial t} \bigg|_{\text{ret}}) \right]$$

$$= \int d^3 r_s \left[ -4\pi \delta^{(3)}(\tilde{R})J(t - R/c, \tilde{r}_s) - c^{-1}(-R^{-2}\tilde{R}) \cdot \tilde{R} \frac{\partial J}{\partial t} \bigg|_{\text{ret}} - c^{-1} \tilde{\nabla} \cdot (R^{-2} \frac{\partial J}{\partial t} \bigg|_{\text{ret}}) \right].$$

The three delta functions eliminate the integral over $\tilde{r}_s$ and set $\tilde{r}_s = \tilde{r}$, so continuing,

$$= -4\pi J(t, \tilde{r}) + \int d^3 r_s \left[ (cR^2)^{-1} \frac{\partial J}{\partial t} \bigg|_{\text{ret}} + c^{-1} 2R^{-3} \tilde{R} \cdot \tilde{R} \frac{\partial J}{\partial t} \bigg|_{\text{ret}} - (cR^2)^{-1} 3 \frac{\partial J}{\partial t} \bigg|_{\text{ret}} \right]$$

$$+ (cR)^{-2} \tilde{R} \cdot (\tilde{\nabla}R) \frac{\partial J}{\partial t} \bigg|_{\text{ret}}.$$

---

3This traditional terminology may cause confusion; note that $t - R/c$ is \textit{earlier} than the observation time $t$.

4$\Box$ is also called the D’Alembert operator, or “dalembertian.”
Chapter 25 Generation of Radiation: First Look

The three terms in the brace cancel.

Bringing the last term on the right to the other side, we have shown that Equation 25.5 solves Equation 25.6 for any \( \mathbf{J} \). Reinstating the vector character of \( \mathbf{A} \) and multiplying by \( \mu_0/(4\pi) \) establishes Equation 25.4: The Green function solution for the Coulomb-gauge vector potential created by a specified current distribution with net charge everywhere zero.

25.5 OUR FIRST ANTENNA

25.5.1 A closed current loop can carry current without charge building up anywhere

Ultimately, we would like to see whether and how radiation can be emitted from a dipole microwave antenna (Figure 43.2). We’re not ready for that yet, because whenever charge flows into one arm of the antenna, it piles up, violating the neutrality condition that we’ve assumed so far (Section 25.1). Instead, consider a circular loop of wire in the \( xy \) plane, centered on the origin, with radius \( a \) (Figure 25.1). As usual, assume that the charge \( I(t) \) falls with distance \( 1/r^3 \). Now we want to explore what changes when the current alternates.

Section 25.5.1’ (page 399) will discuss a more realistic treatment.

25.5.2 Far from the source, the fields fall as \( 1/r \)

We know the fields far from a static magnetic dipole: \( \vec{E} = 0 \), and \( \vec{B} \) falls with distance to the antenna. According to a contemporary, FitzGerald previously presented an erroneous paper in 1879 on the “impossibility” of producing electric waves, but struck out the “im” afterward.

\(^5\)This example appears to be due to G. FitzGerald in 1883. FitzGerald also derived the \( \omega^4 \) rule for power emission and suggested a spark gap as a generator of high-frequency alternating current to drive the antenna. According to a contemporary, FitzGerald previously presented an erroneous paper in 1879 on the “impossibility” of producing electric waves, but struk out the “im” afterward.

\(^6\)See Your Turn 17A (page 254).
Imagine sitting somewhere far away along the +x axis, at position \( \vec{r} = (L, 0, 0) \). What are the electromagnetic fields there, to leading nontrivial order in powers of \( 1/L \)?

We parameterize the wire loop by azimuthal angle \( \varphi_+ \), which runs from zero (closest point to our observer) to \( 2\pi \) (same point). At any point on the loop, the current points in the azimuthal direction \( \pm \phi \). So Equation 25.4 gives:

\[
\vec{A}(t, \vec{r}) = \frac{\mu_0 I}{4\pi} \int_0^{2\pi} (a d\varphi_+) R^{-1} \left[ \frac{1}{2} e^{-i\omega(t-R/c)} \hat{\phi} + c.c. \right].
\] (25.7)

In this formula, \( R = \sqrt{(L - a \cos \varphi_+)^2 + a^2 \sin^2 \varphi_+} \), and \( \hat{\phi} \) is the unit tangent vector to the loop at angular position \( \varphi_+ \).

Our answer can be simplified a lot if we start by studying only the leading-order behavior in powers of \( a/L \) (the small source regime\(^9\)). Thus, \( R^{-1} = L^{-1} + \cdots \), where the ellipsis contains only terms that we have agreed to drop. This factor is independent of \( \varphi_+ \), so it can be moved outside the integral, along with the time dependence:

\[
\vec{A} = \frac{\mu_0 I}{4\pi} \frac{a}{L} \frac{1}{2} e^{-i\omega t} \int_0^{2\pi} d\varphi_+ (-\hat{x} \sin \varphi_+ + \hat{y} \cos \varphi_+) \exp\left[ i \frac{\omega}{c} L (1 - \frac{a}{L} \cos \varphi_+ + \cdots) \right] + c.c. \] (25.8)

We must be careful with the last exponential. Inside it, the first subleading term may not be dropped. Even though it is smaller than the leading term, nevertheless it is not small in an absolute sense, because the \( L \) factors cancel.

**Your Turn 25C**

However, show that the terms even higher than this one may be neglected as \( L \to \infty \).

(That is why Equation 25.8 abbreviated them by an ellipsis.)

We therefore find the vector potential in the small source regime to be

\[
\vec{A} \to \frac{\mu_0 I}{4\pi} \frac{a}{L} \frac{1}{2} e^{-i\omega t - L/c} \int_0^{2\pi} d\varphi_+ (-\hat{x} \sin \varphi_+ + \hat{y} \cos \varphi_+) \exp\left[ -i(\omega a/c) \cos \varphi_+ \right] + c.c.
\] (25.9)

The term that points along \( \hat{x} \) integrates to zero by a symmetry argument: It is an odd function of \( \varphi_+ \), which may be integrated over the symmetric range \(( -\pi, \pi )\). The \( \hat{y} \) term need not be zero, however. Thus, the vector potential far away from the loop has a contribution that, at nonzero frequency, falls slowly with distance (as \( L^{-1} \)).

Note that the \( \hat{y} \) term of Equation 25.9 would also integrate to zero in the static case \(( \omega = 0 )\); more generally, however, it need not vanish.

---

\(^7\)By rotational symmetry, we get a similar result when we go far away in any direction in the plane of the loop. In Problem 25.1, you’ll study the far fields in other directions.

\(^8\)Substitute the charge flux in the thin-wire approximation (Equation 15.21, page 232) into Equation 25.4.

\(^9\)In statics, Chapters 3 and 17 called this limit “far field,” but in dynamics we will reserve that term for a stricter condition. Chapter 43 will explore this and other regimes systematically.
Your Turn 25D

a. Suppose that $\omega$ is small but nonzero; use a Taylor expansion of the exponential in Equation 25.9 to get an approximate answer for the integral.

b. Apply stationary phase approximation (Section 22.3) to get an answer in the opposite limiting case, where $\omega$ is large.

c. Ask an analytic math assistant about the needed integral, for example, by asking for \texttt{Integrate[Cos[x]*E^(-I*p*Cos[x]),{x,-Pi,Pi}]} . Graph the answer (a function of $p$), along with the approximate form for the small-$p$ limit found in part (a) and the other extreme in (b).

Because we used restricted Coulomb gauge, the scalar potential is zero. Thus, the electric field is simply $-\partial \vec{A}/\partial t$. The time derivative just introduces a factor of $(-i\omega)$, so

$$\vec{E} \to (\text{const}) \frac{1}{L} e^{-i\omega t - L/c} \hat{y} + \text{c.c.} \quad \text{along } x \text{ axis as } L \to \infty.$$  

Although there is no net charge anywhere, we nevertheless get an electric field, in contrast to the case of a static magnetic dipole. Moreover, the field falls off slowly with distance, as $1/L$, in contrast to a static dipole. For a distant observer anywhere in the $xy$ plane, $\vec{E}$ is polarized in the plane perpendicular to the magnetic dipole moment.

We also get a prediction that the outgoing wave observed along the $+x$ axis is nearly a plane wave traveling along $+\hat{x}$ and linearly polarized along $\hat{y}$. Thus, it is polarized transversely to the “line of sight,” in this case $\hat{x}$.

What about the magnetic field, given by the curl of $\vec{A}$? We might naïvely imagine that it must fall as $L^{-2}$ (the derivative of $L^{-1}$), but think about the factor $e^{i\omega L/c}$ in Equation 25.9. When we differentiate in the $\hat{x}$ direction, this factor introduces $i\omega/c$, and no additional $L^{-1}$. Thus, the leading behavior of $\vec{B}$ at $L \to \infty$ is

$$\vec{B} \to (\text{const}) \frac{1}{L} e^{-i\omega t - L/c} \hat{z} + \text{c.c.},$$

again a slower falloff with distance than in the case of a static magnetic dipole. The magnetic field is perpendicular to $\hat{x}$ and also to the electric field, similarly to a plane wave propagating along the $\hat{x}$ direction. It points along $\pm \hat{z}$.

In small-source approximation, $\vec{E}$ and $\vec{B}$ form an approximately plane wave moving toward the observer. For a distant observer anywhere in the $xy$ plane, the magnetic field is parallel to the magnetic dipole moment whose oscillation gave rise to the wave.
Your Turn 25E

Keep track of factors that were dropped in the preceding formulas and confirm two other key features:

For an oscillating magnetic dipole source in the limit of low frequency, the fields are proportional to the amplitude of the dipole moment (here $\pi a^2 \vec{I}$), and to the frequency squared.

Media 11 shows the streamlines of $\vec{B}$ (also called magnetic field lines).

25.5.3 Net energy escapes to infinity

The slow field falloffs in $\vec{E}$ and $\vec{B}$ are the hallmark of radiation. They imply that energy is being continually sent out to infinity, if the frequency $\omega \neq 0$. To see this, recall from Chapter 20 that a test charge can extract power proportional to $||\vec{E}||^2$. Although the direction of $\vec{E}$ oscillates, its mean-square value is nonzero. Imagine a spherical shell of such receivers, all at distance $L$ from the source. The area of that shell increases as $L^2$, whereas the energy flux falls like $||\vec{E}||^2$, that is, as $L^{-2}$. So the total energy sent out from the source is $\propto L^2L^{-2}$—it is independent of $L$. In other words, our antenna sends energy out all the way to infinity: It radiates, just as a candle or a star radiates light. However, you’ll see in Problem 25.1 that the loop antenna is directional.

Although we have not yet found a fully general analysis, Section 25.5 has shown how the main features of radiation emerge from the Maxwell equations.

25.6 PLUS ULTRA

This is the end of Part III. In a sense, we could stop here: We know most of what’s needed to understand the second Industrial Revolution.\(^\text{10}\) We have also found an unexpected Electromagnetic Phenomenon (waves), also with vast technological implications (including radio). We have seen one way for waves to be generated and how they deliver energy, momentum, and even angular momentum across empty space.

But there is a lot more work to do! Many antennas of interest don’t have zero net charge (for example, the transmitting antenna in the microwave generator demonstration, Media 1, is not a closed loop), so we’ll need a more general formalism. However, every complicated thing that we’ll do later is just a variation on the straightforward calculation in Section 25.5.2.

More importantly, although the derivation given in this chapter was straightforward, it relied on too much magic. We should develop a more sophisticated formalism, and accompanying physical intuitions, that will make it clear that Equation 25.4 is correct, without all the messy verification. To find that deeper understanding, Parts IV and V will

---

\(^{10}\)That’s the one involving electrical technology. The next revolution (semiconductors) required quantum mechanics.
uncover an important aspect of the Maxwell equations that has been hiding in plain sight ever since we introduced Maxwell’s correction to Ampère’s law.

FURTHER READING

Intermediate:
An alternative to the derivation in Section 25.4, in Fourier space, appears in Pollack & Stump, 2002, §15.1.1.

25.5.1 Realistic antenna theory requires a self-consistent solution

The main text made the assumption that the current through the loop took a simple form. Really, however, when we connect a loop of wire to a signal generator, the resulting current must be calculated by self-consistently solving the Maxwell equations for the field along with:

- the Lorentz force law for charges,
- some characterization of the signal generator, and
- an ohmic assumption about the wire.

At high frequency, the finite capacitance of the loop will permit nonzero charge pileup, invalidating our assumption. In other words, “antenna theory” is a large branch of electromagnetic engineering that we will gloss over, both here and in Chapter 43.
25.1 Directionality of antenna

A circular loop of wire, carrying an oscillating current, lies in the $xy$ plane (Figure 25.2). Equation 25.7 gives the vector potential, and $\psi$ is zero. In this formula, the field point $\vec{r}$ is the position of the observer. The angle $\varphi_*$ specifies an element of the loop located at $\vec{r}_* = a \hat{r}$. The unit vectors $\hat{r}$ and $\hat{\varphi}$ are evaluated on the loop at $\varphi_*$. The distance $R(\varphi_*) = ||\vec{r} - \vec{r}_*||$. The current in the loop is everywhere $I(t) = I \cos \omega t$.

The main text examined the far fields at distant points in the $xy$ plane. Instead, now find the vector potential, this time for an observer located along the $z$ axis. Then characterize the far electric and magnetic fields in words, and contrast with their far-field behavior when viewed at points along the $x$ axis.

25.2 Square loop

Repeat the analysis of Section 25.5 for an antenna that is a square loop of wire with side $a$. That is, evaluate the far fields for the limiting case of low frequency and compare to the result in Your Turn 25Da (page 396). Can you make a statement that covers both of these cases?

25.3 From far to near

Background: The main text derived an exact expression for the vector potential outside an arbitrary current distribution, for the situation with zero charge density everywhere. Section 25.5 (page 394) specialized to the case of an oscillating current confined to a loop of wire, and to an observer located on the $x$ axis. Then we made an approximation: The observer was assumed to be far away, so we discarded $O(L^{-2})$ terms. Your Turn 25Da made the additional approximation of long wavelength (low frequency; source motion follows newtonian mechanics). That was useful for specialized situations. In this problem, you’ll get your digital assistant to compute the fields without either of these approximations.

Let’s begin with some intuition about the full solution. One of Faraday’s insights was

![Figure 25.2: A current loop in the $xy$ plane. See Problem 25.1.](image)
that unlike electric field lines, magnetic field lines must be closed loops (no ends). When viewed close to a singular source, such as a thin wire, the field lines at any moment of time should just wrap around the wire in a direction based on the current at that time (each one “links” with the current loop). As we move away from the wire, however, more interesting things can happen: The field lines may detach from the source and move outward as closed curves that don’t link the current loop. We’d like to know if this detachment really occurs.

In the following steps you’ll again consider a circular loop of wire of radius \( a \) in the \( xy \) plane, carrying a prescribed, harmonically oscillating current \( I(t) = \frac{1}{2} |I| e^{-i\omega t} + \text{c.c.} \) (see Figure 25.1, page 394). Thus, \( I \) is one half of the peak-to-peak current amplitude. You are to find and plot the magnetic field \( \vec{B}(t, \vec{r}) \) everywhere, at various times. This involves numerically evaluating a formula obtained in Section 25.5.2:

\[
\vec{A}(t, \vec{r}) = \frac{\mu_0 I}{4\pi} \int_0^{2\pi} (ad\varphi_s) R^{-1} \left[ \frac{1}{2} e^{-i\omega(t-R/c)} \vec{\phi} + \text{c.c.} \right]. \tag{25.10}
\]

In this expression,

- \( R = ||\vec{r} - \vec{r}_s(\varphi_s)|| \), where \( \vec{r}_s \) is the point on the loop at angular position \( \varphi_s \), and
- \( \vec{\phi} \) is the unit tangent vector to the loop at angular position \( \varphi_s \).

You’ll work out the curl of Equation 25.10, and only then evaluate it numerically. Actually, it’s enough to find \( \vec{B}(t, \vec{r}) \) for \( \vec{r} \) in the \( xz \) plane, and indeed to look only at \( x > 0 \), because of the axial symmetry. But unlike in the main text, don’t restrict to \( \vec{r} \) just along the \( x \) axis.

**Do:**

Measure all lengths in units of the loop radius \( a \). (Or equivalently, measure lengths in meters and take \( a = 1 \) m.) Measure time in units of \( a/c \). The numerical value of \( c \) is 1 in these units (that is, in units of \( a/(a/c) \)).

a. Write a symbolic expression for the curl of \( \vec{A} \), specialized for the situation in the problem. [Hint: Remember that you must evaluate any \( y \) derivatives before setting \( y = 0 \).]

Leave the integral unevaluated for now. Using your analytic expression, show that one of the three cartesian components of \( \vec{B} \) equals zero throughout the \( xz \) plane.

That last point is convenient: It means that every integral curve (streamline, Section 0.3.1, page 7) of \( \vec{B} \) that starts in the \( xz \) plane will remain completely in that plane. These curves are Faraday’s magnetic “field lines.”

b. Set \( \omega = 0 \), and check your analytic results by comparing to a case that you know: the far fields in the magnetic multipole expansion. That is, expand your result for large distance \( r \gg a \), then do the integrals explicitly. Next, look up the well-known static magnetic dipole field, compare, and if necessary reconcile your result.

c. When you are confident in your result, still with \( \omega = 0 \), numerically evaluate your complete formulas for \( \vec{B} \) on a grid of points with \( y = 0 \) and say, \( 0 < x < 5a \) and

---

11 Chapter 36 will obtain this from \( \nabla \cdot \vec{B} = 0 \).
12 Or use your result in Your Turn 17A (page 254).
Use a computer to display the streamlines of this vector field. \(^{13}\) Try
telling the software specifically to make streamlines that pass through a set of points
\(x(t), 0\), that is, points along the \(x\) axis at an evenly spaced series of values \(\{x(t)\}\). (Just
make sure none of your choices is \((1, 0)\), because the fields are singular exactly on the
wire.)

Note that overall factors like \(\mu_0 I/(4\pi)\) aren’t needed when we want only the stream-
lines. Your computer math system may choose different scales for the \(x\) and \(z\) axes in
your plot. So if necessary, figure out how to override that behavior.

d. As mentioned above, some or all of your integral curves may have the property that
they link the current loop.\(^{14}\) We say they are “attached to the source.” Find which ones
have this property and comment.

e. Move on to nonzero angular frequency \(\omega = 2\pi c/(3a)\). Again find and plot the \(\vec{B}\) field
lines at time \(t = 0\). This time, we expect the far fields to be waves with wavelength \(3a\).
Comment on the behavior you observe both close to and far from the origin; on the
\(z\) axis versus on the equatorial plane; and so on. If some of the integral curves (field
lines) are not linked with the source loop, estimate the locus separating the attached
lines from the detached ones.

f. A picture may be worth a thousand words, and \(N\) pictures may be worth \(N\) thousand
words, but a movie of those \(N\) pictures would be better still. After all, we are studying a
spacetime phenomenon. So get your computer to make an animation, covering many
moments throughout a period \(2\pi/\omega\). [Hint: You’ll get a smoother movie if you choose
appropriate initial points when finding the streamlines: At time \(t\), ask your software
for streamlines that pass through \((x(t) + ct, 0)\), where \(\{x(t)\}\) are the same points you
used in (b).]

Show some initiative. Suppose these are figures in a paper you’re trying to publish—make
some improvements in presentation, informative labels, and so on. If you think that the
range from 0 to \(5a\) doesn’t show the physics optimally, choose some better range. Play.

g. Finally, the easy part: Write a formula for the electric field, again containing an integral.
Without explicitly evaluating it, find the direction that \(\vec{E}\) points at any point in space.
Describe qualitatively the corresponding electric field lines. Then comment on their
geometrical relation to the magnetic field lines that you found previously.

25.4 \textbf{Emergence of transversality}

Continuing Problem 25.3, now compute the longitudinal part of \(\vec{B}\), that is, \((\hat{r} \cdot \vec{B})/||\vec{B}||\)
at time zero, and plot it in some way that shows how the field becomes transverse as an
observer moves away from the loop. [Remark: One way to convey this information is to
plot the requested quantity as the observation point moves outward along some ray, for
example the diagonal \(x = z\). Or better, find a way to plot it throughout the \(xz\) plane.]

\(^{13}\)See Problem 3.10. For example, Python has a function \texttt{plt.streamplot} that accomplishes this. Problem 15.7
discussed the fields created by a stationary current loop of finite size.

\(^{14}\)You are plotting a slice, that is the field in part of the \(xz\) plane, so the current loop just looks like the two points
\((\pm a, 0)\), one of which is outside the range you are plotting. Indicate the other one in your plot by a dot. A curve
in the \(xz\) plane therefore links the current loop if it encircles that dot.
25.5 **Twist it up**

First do Problem 25.3 parts (a–e). But then consider a current source consisting of two circular loops of wire. One lies in the $xy$ plane and again carries sinusoidal current with angular frequency $\omega$ the same as in part (d) of that problem. The other lies in the $xz$ plane and carries sinusoidal current with the same frequency and amplitude, but shifted in phase by $1/4$ cycle relative to the first one. In this situation, we may *not* restrict everything to the $xz$ plane.

a. Write a superposition of two formulas, each similar to the one you used in Problem 25.3 part (d).

b. Choose a moment of time at which the current in the $xz$ loop equals zero (and hence the current in the $xy$ loop is maximum). Write a function that can evaluate $\vec{B}$ anywhere in space at the one instant of time you chose.

c. Make a three-dimensional streamplot of some representative magnetic field lines that pass through a collection of starting points lying along the $+x$ axis. Rotate your plot to gain some perspective. Print one or two good-looking views, but describe in words how they look as three-dimensional curves, and how they interpolate between what you expected at short and at long distances.

*Optional*: If you think this would be better as a movie—nobody’s stopping you.]
Ole Rømer’s astronomical measurement of the speed of light began with the idea that the orbits of Jupiter’s moons are periodic phenomena reported to us by light, and hence must suffer a Doppler-type change in frequency controlled by our motion relative to Jupiter’s.
False views, if supported by some evidence, do little harm, as everyone takes a salutary pleasure in proving their falseness.

— Charles Darwin

26.1 FRAMING: THE PRINCIPLE OF RELATIVITY

This chapter’s goal is to rephrase some familiar ideas in a useful way. Although later chapters will overturn these ideas, we will set up a general framework that will survive that revision.

Galileo believed that the Earth moved around the sun, while also spinning on its axis. Many found this proposition absurd. If the Earth moves, why doesn’t it feel like we’re moving? Why aren’t we thrown off? Galileo patiently constructed arguments about how, below decks on a ship moving uniformly on a calm sea, butteries will fly with the same speed in all directions, and so on. While he didn’t have it completely straight, his successors (Huygens and Newton) eventually elevated this idea to the status of a fundamental principle, which we now call the Principle of Relativity:¹

No experiment done within an isolated system can determine whether or how fast that system is moving. More precisely, if we put all our apparatus in a box and measure time and space via instruments anchored to that box, then the results recorded in any experiment will be the same regardless of whether that box is at rest or moving in a straight line at uniform speed.

Einstein didn’t introduce the Principle of Relativity. Nor did he discard it: We still believe it to be experimentally correct. What Einstein said was that newtonian physics implements the principle in a way that is demonstrably wrong. Before we get into that, this chapter will review the newtonian situation.

Electromagnetic phenomenon: Light from distant objects arrives at Earth with a delay related to distance, but not velocity, of the source.

Physical idea: Light cannot be interpreted as a stream of tiny material particles emitted from a source and then following newtonian physics.

¹Section 26.1’ (page 417) discusses the notion of “isolated system.”

¹Henri Poincaré seems to have introduced this phrase, centuries later. A “principle” is not a firm starting point that you can use to prove other things. Nor is it itself a provable proposition. Think of a “principle” as a generator of interesting hypotheses.
26.2 AN ILLUSTRATION FROM MECHANICS

Let’s see how the Principle of Relativity plays out in a concrete situation. Consider two equal point masses $m$ joined by a spring with equilibrium length $L$ and spring constant $k$, floating freely in outer space without rotating.\(^2\) Newtonian mechanics says that their motions are solutions to the equations

\[
\frac{d^2 x_1}{dt^2} = -\frac{k}{m} (x_1 - (x_2 - L)) \quad \frac{d^2 x_2}{dt^2} = -\frac{k}{m} ((x_2) - L - x_1). \tag{26.2}
\]

Although these are familiar equations, let’s unpack their content a bit.\(^3\)

Classical mechanics is formulated in terms of events. An event is specified by a location in space and a moment in time. A trajectory is a continuous chain of events, for example, the locations of a particle at various times.\(^4\) We think of events as points in a four-dimensional space, called spacetime, and trajectories as curves in spacetime. To do analytical work, we must uniquely assign four numbers to each event; that is, we must impose a choice of coordinate system on spacetime. In this language, Equations 26.2 implicitly claim that:

It is possible to label events (points in spacetime), in such a way that every allowed motion of this system corresponds to a pair of curves in spacetime whose coordinate representations are solutions to Equation 26.2.

The following sections review a key fact about newtonian mechanics in this context:

Newton’s laws of motion have a mathematical property called galilean invariance, which guarantees that the physics they predict will obey the Principle of Relativity.

Our ultimate goal is to investigate the same claim about the Maxwell equations and show it’s not valid. However, we’ll find a different, true, property that again guarantees the Principle of Relativity. First we will review how it works in newtonian physics, in two equivalent formulations.

26.3 ACTIVE VIEWPOINT: SYMMETRY

Here is one solution to Equations 26.2:

\[
x_1(t) = C \cos(\omega t) \quad x_2(t) = L - C \cos(\omega t). \tag{26.3}
\]

Here $C$ is any constant and $\omega = \sqrt{2k/m}$. Starting from one such solution, we can manufacture many others by adding any constant $A$ to both $x_1$ and $x_2$:

\[
\tilde{x}_1(t) = C \cos(\omega t) + A \quad \tilde{x}_2(t) = L - C \cos(\omega t) + A. \tag{26.4}
\]

\(^2\)Alternatively, the masses could be free to move in 1D along a frictionless air track in the lab, because the acceleration of Earth’s surface is small.

\(^3\)See also page 22.

\(^4\)Some books use the term “world-line” for this concept.
Such transformations are called **active**, because the new solution is a **physically different motion** from the original. The operation in Equation 26.4 transforms any solution of the equations of motion into another solution (and nonsolutions to nonsolutions). We will call an active transformation with that property a **symmetry** of the dynamics.

That is, a symmetry of a set of equations of motion is an operation that can be applied to any trajectory, but that specifically permutes the solutions. In addition to the overall translation described by Equation 26.4, any isolated, 1D newtonian system also has symmetry under shifts of time by any constant. (There are also discrete symmetries involving reflections in space and in time.)

### 26.4 Passive Viewpoint: Invariance

#### 26.4.1 Alternative representations of the same physical situation

The “active” viewpoint in the preceding section has the advantage of being concrete, but we usually don’t have a catalog of all the solutions to our equations. There is an equivalent viewpoint that, while more abstract, does not require this. Instead of looking for transformations that permute solutions, we focus on a property of the equations themselves.

To see how it works, start with any trajectory and re-express the same trajectory in a new set of coordinates:

\[ x' = x - A \quad t' = t - B. \]  

Because we are not physically changing the trajectory, this transformation is called **passive**: it just changes the representation of a trajectory. Equation 26.4 shifted any trajectory to the right by \( A \), whereas Equation 26.5 shifts the coordinate axes to the right by \( A \).

We now change variables in the equations of motion and ask how they look when expressed in terms of the new coordinates: The usual rules of calculus give \( \frac{d}{dt} = \frac{d}{dt'} \). Everywhere else, we just substitute \( x' + A \) wherever we see \( x \):

\[ \frac{d^2}{dt'^2} (x'_{(1)} + A) = -\frac{k}{m} (x'_{(1)} + A - (x'_{(2)} + A - L)). \]

Cleaning up, we see that the form of the equation of motion, after expressing it in the new variables, is the same as it was in the old variables (Equation 26.2), including the numerical values of constants \((k, L, \text{ and } m)\). We say that the original equations of motion have an **invariance** under the passive transformation Equation 26.5.

#### 26.4.2 Relation between active and passive

Clearly the active and passive viewpoints are closely related. To see the relation, suppose that we know a passive invariance and consider the following operation:

Starting with any trajectory, construct a new, different trajectory by:

- Expressing the original trajectory in unprimed coordinates via some functions;
- Constructing a new trajectory that, in the primed coordinates, is expressed by the same functions.

\[ \text{(26.6)} \]
Figure 26.1: [Mathematical functions.] **Active versus passive.** The initial trajectory (Equation 26.3, solid blue curves) appears different in the original (solid black) and shifted (dashed black) coordinate systems. The corresponding actively transformed trajectory (Equation 26.7, dotted red) appears the same in the shifted coordinate system as the original one in the original system. For example, in the initial trajectory the mass on the left repeatedly crosses the \( t \) axis; in the actively transformed trajectory it repeatedly crosses the \( t' \) axis.

The new trajectory just described will therefore solve the original equations of motion if and only if the old one did, so we conclude that operation that constructed it is an active symmetry.

Thus, active symmetry and passive invariance are complementary viewpoints; in any situation, we can use whichever gives us the best intuition.

Figure 26.1 illustrates this procedure with the solution in Equation 26.3. Applying the recipe in Idea 26.6 with the transformation Equation 26.5 gives

\[
\begin{align*}
\bar{x}'_{(1)} &= C \cos(\omega t') - \bar{x}_{(1)}(t) \\
\bar{x}'_{(2)} &= L - C \cos(\omega t').
\end{align*}
\]

In terms of the original coordinates, we then have

\[
\begin{align*}
\bar{x}_{(1)} &= C \cos(\omega (t - B)) + A \\
\bar{x}_{(2)} &= L - C \cos(\omega(t - B)) + A,
\end{align*}
\]

which is indeed the formula in Equation 26.4, generalized to include time translation.

### 26.5 ROTATIONS AND DILATIONS ARE BOTH LINEAR, BUT ONLY ROTATIONS ARE INVARIANCES

Continuing with the passive viewpoint, we now upgrade to a world with two spatial dimensions. If we set up cartesian axes, we can label every point in the plane by two numbers \([\bar{x}_x, \bar{y}_y]\). Then the same point viewed from a rotated point of view will be labeled by two different numbers \([\bar{x}'_{x'}, \bar{y}'_{y'}]\). We can find the new coordinates by using trigonometry, and the fact that the new coordinate axes are rotated by some angle \(\alpha\) relative to the old ones. There’s a simple formula expressing this:

\[
\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos \alpha & \sin \alpha \\ -\sin \alpha & \cos \alpha \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.
\]

To think about this conceptually, imagine digging up all the streets in Manhattan and laying down a new grid of streets rotated anticlockwise relative to the old one by \(\alpha\). Then
if the Empire State Building is at a point \( P \), it will still be at the same point \( P \) after the new grid is laid down, but the coordinates of that point (nearest street and avenue) will no longer be the same as they were before.

Now, certainly there are many other coordinate systems we could use to label points in the plane, besides the two cartesian systems just described. For example, we could use axes that are not at right angles. But there is something special about a cartesian system: The distance between two points \( P_1 \) and \( P_2 \) is given by the simple formula:

\[
d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}.
\]

If we describe the points using the rotated coordinate system, the formula has exactly the same form:

\[
d = \sqrt{(x_1' - x_2')^2 + (y_1' - y_2')^2}.
\]

Generic coordinate transformations don’t have this property. For example, if we define new coordinates via a dilation transform, \( r' = 2r \), the form of the distance function is not quite the same. In short,

\textit{In euclidean geometry, one class of coordinate systems is special (the cartesian systems). Within that class, however, any system is just as good as any other one.}

When we upgrade the equations of motion for two balls on a spring from 1D to 2D or 3D, they involve the spring potential energy

\[
U = \frac{1}{2}k|\vec{r}_{(1)} - \vec{r}_{(2)}|^2.
\]

Because the distance function takes the same form when expressed in terms of a rotated coordinate system, the equations of motion will have the same property: They are rotation invariant. In contrast,

\section*{Your Turn 26A}

a. Show that the 3D version of Equation 26.2, when expressed in terms of dilated coordinates \( \vec{r}' = 2\vec{r} \), take a new form that look similar, but that have a different value of \( L \) (unless \( L = 0 \)).

b. Even if \( L = 0 \), show that a nonlinear spring will also spoil dilation invariance.

c. Establish the rotation invariance of two point masses bound by gravitational force but otherwise free in space, and their lack of dilation invariance.

d. One may imagine trying to rescue dilation invariance by also allowing rescaling of time: \( \tau' = 2\tau \) and \( \tau' = \zeta \tau \). Show that in a world with both newtonian gravitation and springs, this gambit does not succeed, regardless of what we choose for \( \zeta \).

Newtonian physics does not have any general invariance under dilations.

\section*{26.6 GALILEAN GROUP}

\subsection*{26.6.1 Some coordinate systems on spacetime are preferred}

In math, the assignment of a coordinate system to a space is pretty flexible. Certainly there are lots of choices we could make on our four-dimensional spacetime. But in most of these choices, the equations of physics look different from the usual form. We already

\footnote{See Section 14.2.}
saw one example (dilation). Similarly, most time-dependent transformations, such as $\vec{r}' = \vec{r} + \vec{a}t^2/2$, introduce new “fictitious forces.” That is, the equations are again not form-invariant when re-expressed in terms of this $\vec{r}'$.

Turning that observation around, we can ask which coordinate systems do leave the form of Newton’s laws invariant. In other words, we can let physics select the good systems. We will call them \textbf{G-inertial}, in honor of Galileo. Translations like Equation 26.5 and rotations like Equation 26.8, supplemented by $t = t'$, are invariances of newtonian physics, and hence they take one such G-inertial coordinate system to another.

Confusion may arise over the use of phrases like “frame of reference” (and “observer,” which sounds like it gives an essential role to human consciousness). We will instead usually refer to a “coordinate system,” which may or may not have the property that the equations of motion take their usual form. If they do, then the coordinate system is G-inertial (or simply “good”). In newtonian physics, a human observer always has the option of setting up a G-inertial coordinate system to describe what they measure, but actually doing so may be an elaborate and subtle procedure in practice.

One exception to the terminology just outlined is that we will bow to widespread usage and say \textbf{rest frame} to mean “inertial coordinate system on spacetime in which a particular body momentarily has velocity $\vec{0}$.” It’s unambiguous because we will not have occasion to consider a non-inertial rest frame.

Also, beware that the good coordinates for newtonian physics differ from those in Einstein physics; indeed, Einstein denies that G-inertial coordinate systems even exist, and substitutes a different notion. Yet most authors refer to both concepts indiscriminately as “inertial frames.” When necessary, this book will disambiguate with the prefix “G-” (galilean), and later “E-” for Einstein.

\textit{Section 26.6.1'} (page 417) extends the discussion to include parity and time reversal invariance.

\textbf{26.6.2 Boosts connect coordinate systems in relative motion}

Returning to one dimension, there’s another important class of symmetry transformations, called \textbf{galilean boosts}.

---

6“Centrifugal force” and “Coriolis force” are other examples.
7Some books restrict these terms to refer only to \textit{inertial} (“good”) coordinate systems, but others don’t.
8However, you must use context to tell whether a Galilean or Einstein inertial frame is meant.
26.6 Galilean Group

Figure 26.2: [Mathematical functions.] **Galilean boost.** The initial trajectory (Equation 26.3, solid blue curves) appears different in the original (solid black) and boosted (dashed black) coordinate systems. The corresponding actively transformed trajectory (Your Turn 26B, dotted red) appears the same in the boosted coordinates as the original one in the original system. For example, once again in the actively transformed trajectory the left mass repeatedly crosses the $t'$ axis.

### Your Turn 26B

a. Show that the passive coordinate transformations:

$$x' = x - v_x t, \quad t' = t$$

are invariances of the equations of motion. That is, show that re-expressing Equation 26.2 in terms of the new variables yields equations of identical form.

b. Find the corresponding active transformation of the trajectory Equation 26.3 (see Figure 26.2), and confirm that it does solve Newton’s law (Equation 26.2).

Equation 26.9 describes a new coordinate system, whose time axis is moving to the right at speed $v_x$ relative to the original. The minus sign indicates that this moving axis can overtake an object moving to the right; in that case, the object appears to move leftward in the new coordinate system.

#### 26.6.3 Matrix notation

It will sometimes be convenient to express Equation 26.9 in matrix form:

$$\begin{bmatrix} t' \\ x' \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -v_x & 1 \end{bmatrix} \begin{bmatrix} t \\ x \end{bmatrix}. \quad \text{galilean boost}$$

(26.10)

### Your Turn 26C

Show that if we make a second transformation of this sort, to $t'', x''$, then we just get the product of two matrices, which is again a galilean boost, this time with speed $v_x(1) + v_x(2)$, that is, the matrix $\begin{bmatrix} 1 & 0 \\ -(v_x(1)+v_x(2)) & 1 \end{bmatrix}$. 
The **galilean velocity addition formula** you just found agrees with our everyday experience with baseballs, water waves, and so on.

### 26.6.4 Galilean transformations have a group structure

All together, in one space dimension Newtonian physics has a 3-parameter family of continuous symmetries/invariances (space translation, time translation, boost), as well as discrete reflections\(^9\) in \(x\) and \(t\). We call that family the **galilean group**. Its elements are **galilean transformations**. In the passive viewpoint, they connect the various G-inertial coordinate systems to one another.

Suppose that we define a primed coordinate system by applying a galilean boost (and a translation) to the original one. Next, define a double-primed system by applying a second galilean boost, and another translation, to the primed system. Still working in one dimension,

\[
x'' = x' - v_{s(1)} t' - A_1 = (x - v_{s(2)} t - A_2) - v_{s(1)} t - A_1, \quad t'' = t' = t. \tag{26.11}
\]

Rearranging terms shows that the overall effect is again the combination of a boost (with speed \(v_{s(1)} + v_{s(2)}\) as you found in Your Turn 26(C)) and a shift (by \(A_1 + A_2\)).

In three space dimensions, the galilean group includes a 10-parameter family of invariances (3 space translations, 1 time translation, 3 space rotations, 3 boosts).

**Your Turn 26D**

a. Generalize Equation 26.9 to include 3D rotations, spatial shifts, and also time shifts.

b. Show that if we apply any two of these transformations in succession, the result is a single transformation that is also in this family.

c. Show that any such transformation has an inverse, which is again in the family.

Mathematicians call a set of transformations with properties (b,c) a **group**, hence the name “galilean group.”

### 26.6.5 The physical significance of galilean invariance

By now, certain questions may be bothering you:

1. Why are we spending so much time with balls on springs? Even within Newtonian physics, that’s a specialized, and idealized, system.

2. A coordinate system is just an arbitrary labeling scheme for points of spacetime. So what has all this formalism got to do with physics?

The answer to the first question is that

\[
\text{All of Newtonian physics has the overarching mathematical property of galilean invariance, which transcends details of particular springs, clocks, planets, and so on.} \tag{26.12}
\]

\(^9\)However, in a dissipative system, temporal inversions are not invariances.
Your Turn 26E

For example, confirm that in newtonian gravity, in one dimension, the equations of motion for two point masses attracting each other also have full galilean invariance.

Idea 26.12 partly explains why in physics we get so much mileage out of studying systems that are obviously absurdly oversimplified, for example, linear springs, spherical planets, and other nonexistent things. Often we are just working out the consequences of invariances that continue to apply to realistic versions of those things. For a simple example of why this principle is significant, notice that invariance under spatial translations means there is no distinguished special central point in space.

For question #2 above, we already noted that any proposed symmetry is a physical property that the world may (translation, rotation) or may not (dilation) possess. It’s not just an aesthetic preference. We also noted that in newtonian physics, some of the good ones connect coordinate systems that are in uniform, straight-line motion relative to each other. Because any set of newtonian equations of motion is invariant under such transformations, then those two coordinate systems are indistinguishable by any experiment confined to the system under study. You can do all the experiments you like, and always find the same equations of motion in each such coordinate system. Nothing you can measure says that one such system is at “absolute rest” nor indeed “better” in any way than another. In short:

Newtonian physics hardwires the Principle of Relativity by using equations of motion that are invariant under galilean boosts.

In a more lapidary phrase:

Physicists study invariance because it strips away details and lays bare the structural essentials of a dynamical theory.

We can now see why Idea 26.12 is so important: If part of physics had galilean invariance, but another part did not, then we could devise an experiment using the second part to determine which coordinate systems are at absolute rest. Even if two parts of physics have slightly different boost invariances, we could say that “absolute rest” was the coordinate system in which both simultaneously took their simplest forms. Only if all of physics has the same boost invariance can we say that absolute rest is completely unobservable—the Principle of Relativity.

Many physical problems involving relativity become clearer when seen from this
high-altitude viewpoint: Often, their solution boils down to:

\[
\begin{align*}
  & \text{There's an inertial coordinate system where I understand what's} \\
  & \text{going on.} \\
  & \text{But I want to know what's going on in some other inertial system} \\
  & \text{(perhaps one that I set up in my lab).} \\
  & \text{So I can use the appropriate transformation to go from the first} \\
  & \text{to the second.}
\end{align*}
\]

(26.15)

Applying this strategy to every situation is not always the fastest route to solve a particular problem. But in the long run it's a unified, sure-footed way to cut through the fog.

We will soon see that Einstein retained most of Ideas 26.12–26.15 and merely tweaked some details of how the transformations work (Chapters 29–30). Once we discover the right transformations, we’ll see many examples of the Relativity Strategy at work.

26.6.6 Light cannot be interpreted as a stream of newtonian particles, part 2

Section 20.3 (page 302) argued that Newton's model of light as a stream of material particles was incompatible with the alternative model implied by Maxwell’s theory. Here is a more direct, experimental objection to the newtonian model.

Suppose that we have a catapult that, when at rest, can fire a projectile with initial speed \( v_0 \). Imagine mounting that catapult on a train car, bringing it up to speed \( v \) directed along \( \hat{x} \), and firing the projectile in the \( +\hat{x} \) direction. Intuitively we might expect that on the ground, we'll observe the projectile moving with velocity \( v_0 + v \hat{x} \).

Let’s obtain the result just stated as a consequence of galilean invariance, using Idea 26.15. We know that there’s a G-inertial coordinate system in which the catapult appears to be at rest. Whatever mechanism is inside the catapult, we are assuming it to be galilean-invariant, so the speed of the projectile from the moving catapult, viewed in the moving coordinate system, must again equal \( v_0 \).

Ex. Apply your result in Your Turn 26C to find the speed as seen in the ground-based coordinate system.

Solution: Let’s look for a coordinate system, denoted with double primes, in which the projectile is at rest. We know that this system is moving uniformly at \( v_0 \) with respect to the primed system, in which the catapult is at rest. We also know that the primed system is moving at \( v \) with respect to the ground. Your Turn 26C then says that the the doubly primed system is moving uniformly at \( v_0 + v \hat{x} \) with respect to the ground.

Alternatively, we can write the trajectory of a projectile fired from a catapult \textit{at rest} as the parametric curve \( \begin{bmatrix} x \\ \xi \end{bmatrix} = \begin{bmatrix} \xi \\ v_0(\xi) \end{bmatrix} \), and that of the stationary catapult as \( \begin{bmatrix} x \\ \xi \end{bmatrix} = \begin{bmatrix} \xi \\ 0 \end{bmatrix} \).

Now apply an active boost to conclude that there must be another solution, in which the projectile’s trajectory is \( \begin{bmatrix} x' \\ \xi' \end{bmatrix} = \begin{bmatrix} \xi \\ v_0(\xi) \end{bmatrix} \) and the catapult’s is \( \begin{bmatrix} x' \\ \xi' \end{bmatrix} = \begin{bmatrix} \xi' \\ 0 \end{bmatrix} \). Re-expressing
these new trajectories in the unprimed coordinate system yields

\[
\begin{align*}
\text{catapult: } & \begin{bmatrix} t \\ x \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ u_{s(2)} & 1 \end{bmatrix} \begin{bmatrix} \xi \\ 0 \end{bmatrix} \\
\text{projectile: } & \begin{bmatrix} t \\ x \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ u_{s(2)} & 1 \end{bmatrix} \begin{bmatrix} \xi \\ u_{s(1)} \xi \end{bmatrix}.
\end{align*}
\]

The velocity of the catapult is \( \Delta x/\Delta t = v_{s(2)} \) as desired, and that of the projectile is \( \Delta x/\Delta t = v_{s(1)} + v_{s(2)} \).

W. de Sitter pointed out that this result is bad news for the model of light as tiny material particles emitted from a source and then following newtonian physics. Consider a binary star, that is, two stars orbiting their common center of mass. If light consisted of a stream of newtonian particles, then those particles would move faster when each star was approaching us, and slower when it was receding. When the difference in (velocity)\(^2\) got multiplied by the distance to Earth, it would amount to a big change in arrival times. Sometimes we might even see a double image of one star, because it would emit faster light, then move, then emit slower light from the new position and both would arrive simultaneously at Earth! No such phenomena are observed, so light can’t be a newtonian particle.\(^{10}\)

26.7 **PLUS ULTRA**: 1905 AND ALL THAT

In contrast to particles, waves do move at a speed that is independent of the emitter’s motion. So the preceding argument seems to favor a wave model of light over Newton's particle conception. But Chapter 28 will expose problems with the classical wave model as well. Chapter 31 will show how Einstein evaded both problems, clearing the way for today’s dual particle/wave picture of light. We’ll see that Einstein’s contribution was to say that

Electrodynamics, mechanics, and the rest of physics do hardwire in the Principle of Relativity by using equations of motion that are invariant under a kind of boost transformations, but they’re not quite the galilean transformations.

The correct invariance principle, and hence the correct equations of motion, were missed for centuries because, for mechanical objects moving relative to each other much more slowly than \( 3 \cdot 10^8 \text{ m/s} \), the difference from galilean invariance is quantitatively small. For objects (or waves) that move at or near that large speed, however, the distinction becomes important.

**FURTHER READING**

**Intermediate:**

\(^{10}\)More quantitatively, the newtonian hypothesis also predicts an irregularity in the apparent timing of the eclipses of a binary pulsar that was not observed (Brecher, 1977). Also, light emitted in the forward direction by the decay of a rapidly moving pion travels at \( c \), not at \( \approx 2c \) (Alvänger et al., 1964).
26.1’ Complete isolation
Can a system be truly isolated? You could put it in a Faraday cage to screen out cosmic microwave background radiation (and to trap any radiation given off by the system under study). Then your measurements wouldn’t be affected by the tiny anisotropy that arises because we are moving relative to the CMBR (see Section 30.7.2), nor the energy loss if the system radiates.

But there must be analogous gravitational background radiation, which cannot be so screened, plus relic neutrinos and so on; also, your system may in principle emit gravitational radiation!gravitational. So a truly isolated system may be an unattainable idealization. However, in practice such radiation has not yet been observed experimentally, nor the gravitational radiation from any laboratory system.

26.6.1’a Parity invariance
The main text has so far always supposed that we have arbitrarily designated some cartesian coordinate systems as “right handed.” Then there is a well-defined Levi-Civita tensor, and hence cross products, curl, and $\vec{B}$ field. Suppose, however, that we had selected an oppositely-handed cartesian system. Then $\epsilon_{ijk}$ and other quantities derived from it would all change sign, a passive parity (inversion) transformation. Interestingly, however, the equations of newtonian physics all set two true tensors equal (for example, $md^2\vec{r}/dt^2 = -\nabla U$), or else they set two quantities like $d\vec{L}/dt$ (rate of change of angular momentum) and $\vec{F}$ (torque) equal. Indeed, Problems 15.4 (page 238) and 18.1 (page 290) noted that such “pseudo” quantities may be eliminated altogether; in that formulation, nothing needs to be checked.

Similarly, although we will argue that electrodynamics is not galilean invariant, it too is invariant under passive parity transformations: The equations either set two true tensors equal (for example, Ampère’s law, or the Lorentz force law), or else set two quantities like $d\vec{B}/dt$ and $-\nabla \times \vec{E}$ equal (the Faraday law).

There is a corresponding active viewpoint: If we replace every particle trajectory by another one that runs in the opposite sequence and hence has the opposite velocities, an “active time-reversal transformation”: $\vec{r}(t) = -\vec{r}(t)$ and similarly for every tensor field (for example, $\vec{E}(t, \vec{r}) = -\vec{E}(t, -\vec{r})$ but $\vec{B}(t, \vec{r}) = +\vec{B}(t, -\vec{r})$, then those new functions solve the electrodynamics equations if and only if the original ones did.

It may seem as though we must memorize an arbitrary new bit of information about every physical quantity, its character upon inversion (in addition to its dimensions and tensor rank). Again, however, we may sidestep this if we exclusively use true tensors such as $\vec{\omega}$ instead of $\vec{B}$ (Section 15.2, page 222). This approach is also required when we unify $\vec{E}$ and $\vec{B}$ (Section 33.3.3, page 514 and Section 34.7’b, page 542).

26.6.1’b Time reversal symmetry
Another discrete symmetry seems quite different from parity. In newtonian physics, we can replace every particle trajectory by another one that runs in the opposite sequence and hence has the opposite velocities, an “active time-reversal transformation”: $\vec{r}(t) = \vec{r}(-t)$. In a nondissipative system (no friction nor diffusion), the new trajectory will solve the equations of motion if and only if the original did.

Similarly, although we will argue that electrodynamics is not galilean invariant, it too is invariant under the corresponding transformation. We classify some quantities as even under time reversal (for example, $\vec{E}(t, \vec{r}) = \vec{E}(-t, \vec{r})$) and the others as odd (for example, $\vec{B}(t, \vec{r}) = -\vec{B}(-t, \vec{r})$),
then find that the Maxwell equations all either set $\text{even} = \text{even}$ (for example, the electric Gauss law or Lorentz force law) or else set $\text{odd} = \text{odd}$ (for example, the Faraday law). Both charge and mass are taken to be even under time reversal. Again, this alarming proliferation of tensor types will improve when we unify $\vec{E}$ and $\vec{B}$. However, we will still need a distinction between four-dimensional objects that are even and odd under time reversal (Section 34.7c, page 543). Also, time reversal is not a symmetry of dissipative systems, where the increase of entropy sets an “arrow of time.”
26.1  *Thump*

Newton imagined light as a stream of tiny material particles obeying the same sort of
laws as ordinary matter. Benjamin Franklin objected to this model; in 1752 he wrote in a
letter “I must own I am much in the dark about light…. Must not the smallest particle
conceivable, have with such a motion, a [kinetic energy] exceeding that of a [cannonball]?”

Suppose that a tiny particle, weighing just a picogram, could be brought up to the
speed of light. Evaluate the newtonian kinetic energy formula, \( \frac{1}{2}mv^2 \), for this particle,
and comment on Franklin’s assertion.
Chapter 27

Springs, Strings, and Local Conservation Laws

27.1 Framing: Transport

We continue our little newtonian holiday. This is a book on electrodynamics, but it's also about where theories come from. It's good to see abstract things first in a familiar setting.

The preceding chapter started with a vague Principle (of Relativity), but then it turned into precise algebra and calculus (an invariance property). That's an appealing progression, but in this chapter we'll see that we need to be a bit careful applying it. The payoff is that we'll get a framework applicable to field theories, including eventually relativistic ones including electrodynamics.

We'll also extend a framework relevant to other themes of this book, involving energy and momentum transport by waves.

Electromagnetic phenomenon: Energy and momentum are locally conserved on a vibrating string; they cannot disappear at one point and reappear at a distant point without passing through the intervening region.

Physical idea: Energy density and flux are related by a continuity equation, and similarly for momentum.

27.2 Equation of Motion

27.2.1 Longitudinal vibration

Imagine a coil spring, initially straight and in its zero-tension state, with linear mass density \( \rho^{(1D)}_m \) (\( \sim \text{kg/m} \)). The spring resists either compression or extension by exerting restoring forces.

To analyze this system’s motions, we temporarily break it down into finite elements in series, each with equilibrium separation \( \Delta x \), mass \( \Delta m = \rho^{(1D)}_m \Delta x \) and spring constant \( \kappa / \Delta x \). Here \( \kappa \) is a material parameter describing the spring (the stretch modulus, with units of force). That is, we think of the spring as a chain of point masses \( \Delta m \) joined by massless springs (Figure 27.1). We label each mass by its undisturbed position \( x \).

Consider the mass element whose equilibrium position is \( x = 0 \). Displace it in \( x \) by distance \( u(0) \). The two springs flanking this element exert restoring forces on it: The element gets force \( -(\kappa / \Delta x)(u(0) - u(-\Delta x)) \) from its neighbor to the left, and \( +(\kappa / \Delta x)(u(\Delta x) - u(0)) \) from the right, that is, net force

\[
f(0) = \frac{\kappa}{\Delta x}(u(-\Delta x) - 2u(0) + u(\Delta x)).
\]
27.3 The wave equation seems to lack boost invariance

For small $\Delta x$, Newton’s law then becomes

$$\Delta m \frac{\partial^2 u}{\partial t^2} \bigg|_{x=0} = \frac{k}{\Delta x} \frac{\partial^2 u}{\partial x^2} \bigg|_{x=0} (\Delta x)^2 + \cdots, \quad (27.1)$$

where the ellipsis denotes terms that are higher order in $\Delta x$. We now take the continuum limit $\Delta x \to 0$, or equivalently consider only distortions $u(x)$ that vary on length scales $\gg \Delta x$.

Any other mass element has the same dynamics, so $u$ obeys the **wave equation**

$$\frac{\partial^2 u}{\partial t^2} - c_s^2 \frac{\partial^2 u}{\partial x^2} = 0 \quad \text{where} \quad c_s^2 = \frac{k}{\rho_m^{[1D]}}. \quad (27.2)$$

Because this is a partial differential equation, we call it a field theory in one space and one time dimension. Chapter 18 showed that the Maxwell equations in vacuum also contain the wave equation, so this problem is a warmup for bigger things.

Solutions to the wave equation include the familiar harmonic traveling waves moving at the “sound” speed $c_s$:

$$u_{\pm}(t, x) = \bar{u} \cos(\omega(-t \pm x/c_s)). \quad (27.3)$$

The angular frequency $\omega$ can have any value. The plus sign corresponds to a wave solution moving rightward.

27.2.2 Transverse vibration

You can repeat all the above analysis for disturbances in which a string under tension $F_0$ is plucked transverse to the $x$ axis. This time, the displacement (height) $u(t, x)$ gives rise to a net transverse component of the tension proportional to $F_0(\partial u/\partial x)$, and so on. Again you get Equation 27.2 but with $c_s^2 = F_0/\rho_m^{[1D]}$.

27.3 THE WAVE EQUATION SEEMS TO LACK BOOST INVARIANCE

As in the preceding chapter, we will examine galilean invariance from both the active (Section 26.3, page 406) and passive (Section 26.4) viewpoints. Consider relabeling events via a Galilean boost:

$$x' = x - v_x t, \quad t' = t. \quad [26.9, \text{page} \ 411]$$
Active viewpoint

An active transformation replaces a spring configuration \( u \) by a different one, \( \bar{u} \). Following Idea 26.6 (page 407), the transformed trajectory when expressed in the transformed coordinates is expressed by the same function as the original. Applying this recipe to the solutions in Equation 27.3 yields

\[
\bar{u}_\pm = \bar{u}(\omega(-t' \pm x' / c_x)).
\]

In the original coordinates, then

\[
= \bar{u} \cos(\omega(t \pm (v_s t + x) / c_x)).
\]

Now manipulate a little to bring this expression closer to the same overall form as before:

**Your Turn 27A**

a. Obtain

\[
\bar{u}_\pm(t, x) = \bar{u} \cos(\omega(1 \pm (v_s/c_x))(-t \pm \frac{x}{c_x})).
\] (27.4)

b. Show that \( \bar{u}_\pm \) is a traveling wave moving at speed \( \pm c_x + v_s \).

c. What about the transformed wave’s frequency?

You just showed that the new functions don’t belong to our original family of solutions of Equation 27.2 (Equation 27.3), because they clearly don’t move at speeds \( \pm c_x \)! That may be surprising, so before discussing it let’s first rederive it to be sure.

Passive viewpoint

This time, focus on the wave equation itself, rephrasing it in terms of the new variables. From \( x = x' + v_s t' \) and \( t = t' \) we find

\[
u'(t', x') = u(t', x' + v_s t).
\]

The change of variables formula from vector calculus lets us rephrase the equation of motion in the new variables:

\[
\left[ \left( \frac{\partial x'}{\partial x} \right)^2 - c_x^{-2} \left( \frac{\partial x'}{\partial t} \frac{\partial}{\partial t'} + \frac{\partial}{\partial t'} \frac{\partial}{\partial t} \right)^2 \right] u' = 0.
\]

Simplifying yields

\[
\left[ \frac{\partial^2}{\partial x'^2} - c_x^{-2} \left( -v_s \frac{\partial}{\partial x'} + \frac{\partial}{\partial t'} \right)^2 \right] u' = 0.
\]

The original equation, Equation 27.2, when re-expressed in the new variables, doesn’t maintain its original form. Thus, the wave equation has neither active symmetry, nor passive invariance, under galilean boosts.
In short, the wave equation is not galilean invariant. Is this a crisis in Physics? No, of course not—this is a newtonian system, and newtonian dynamics does have galilean invariance. The problem is that we have neglected a relevant dynamical variable: Before we plucked that string, it could have been in motion with respect to the observer, and hence with respect to any coordinate system in which the observer appears to be at rest. We did not yet account for this possibility.

That is, Equation 27.2 is incomplete: It only applies to the special case where the initial state of the string is at rest with respect to the coordinate system. If that situation holds for the coordinate system \( t, x \), then it won't hold for the boosted \( t', x' \) coordinates, so we shouldn't (and didn't) find the same form for the equation of motion.

Let's start over and formulate a more general situation, a spring initially in uniform motion at arbitrary speed \( v_m \) (the medium's speed) and again subject to transverse displacement. Let \( u(t, x) \) be the displacement of whichever spring segment is located at spatial location \( x \) at time \( t \). Note that observing a fixed coordinate position \( 0 \) at two different times is not the same as following one particular spring segment.

Consider the spring segment that is located at \( x_0 \) at time \( t_0 = 0 \). Imagine painting that one segment red and applying Newton's Second Law to it. At later time \( \Delta t \), the red segment has moved to \( x = x_0 + v_m \Delta t \). Hence, its transverse velocity \( u_y(t_0, x_0) \) is the limit of

\[
v_y(t_0, x_0) = \frac{1}{\Delta t} \left[ u(t_0 + \Delta t, x_0 + v_m \Delta t) - u(t_0, x_0) \right] = \left( \frac{\partial}{\partial t} + v_m \frac{\partial}{\partial x} \right) u \bigg|_{t_0, x_0}.
\]

The net transverse force on this segment is still \( F_0 \delta x (\partial^2 u / \partial x^2) \) as before, so during time \( \Delta t \) its transverse momentum \( p_y \) changes by \( F_0 \delta x (\partial^2 u / \partial x^2) \Delta t \). That is,

\[
p_y(t_0 + \Delta t, x_0 + v_m \Delta t) - p_y(t_0, x_0) = F_0 \delta x (\partial^2 u / \partial x^2) \Delta t,
\]

or

\[
\left( \frac{\partial}{\partial t} + v_m \frac{\partial}{\partial x} \right)^2 - c_f^2 \frac{\partial^2}{\partial x^2} u = 0.
\]

We just found the generalized wave equation for a spring whose undisturbed state is moving uniformly with respect to the coordinate system at speed \( v_m \). When \( v_m = 0 \), it reduces to the familiar form Equation 27.1. But we may expect that if a distant bell is rung you'll hear it slightly sooner if there is a wind blowing toward you than you would in still air. And indeed:
Your Turn 27B

a. Substitute a generic traveling wave as a trial solution into Equation 27.5 and show that it works if the wave moves at speed $v_m \pm c_s$.
b. Show that a traveling wave solution to this equation, viewed in a boosted coordinate system, belongs to the same family of solutions (though with a different $v_m'$). Thus, the system does have symmetry under active galilean transformations.
c. In particular, an observer who flies alongside the spring at speed $v_m = c_s$ will see some waves that appear static. What condition, if any, must be satisfied for a static waveform to solve the wave equation in this case?
d. Show that Equation 27.5 is also invariant under passive galilean transformations, once we understand that both $u(t, x)$ and $v_m$ must transform.

Thus, galilean transformations really are invariances of the spring system, once we include all relevant dynamical variables and attribute appropriate transformations to them. That is, our error in Section 27.3 lay in mistakenly setting the scope of the system too narrow (treated $v_m$ as a fixed constant of the system, rather than as a dynamical variable subject to transformation).

27.5 CONNECTION TO ELECTROMAGNETISM

Chapter 18 showed that the Maxwell equations imply the wave equation, and Section 27.3 showed that the wave equation lacks galilean invariance. Prior to 1905, everyone assumed that the cure would be along the lines described in Section 27.4: “The Maxwell equations are incomplete, valid only in the special case of a coordinate system at rest with respect to the ‘luminiferous æther.’ After we generalize them to account for ‘æther wind,’ then their full galilean invariance will appear.” One thing that bothered Einstein was that, despite great efforts, nobody had succeeded in finding the right generalization that was mathematically consistent and also consistent with experiments. We’ll see soon where he went with that line of thought, but first we pause to think about the transport of energy and momentum in the familiar setting of springs.

27.6 CONTINUITY EQUATIONS FOR ENERGY AND MOMENTUM

27.6.1 Energy and momentum each have local expressions for their density and flux

For future use, let’s see how energy and momentum are locally conserved in the newtonian mechanics of a vibrating string. In this section, we will choose a spacetime coordinate system in which the string is at rest ($v_m = 0$). We continue to look at transverse waves.

We seek continuity equations for energy and momentum, analogous to the ones we
previously found for mass and charge (Section 8.3, page 116). First note that

\[ \text{KE} = \int dx \frac{1}{2} \rho_m^{(1D)} (\partial u / \partial t)^2; \quad \text{PE} = \text{const} + F_0 \int dx \frac{1}{2} (\partial u / \partial x)^2. \] (27.6)

One way to get the second formula is to imagine that an external agent is pulling the string along its length with tension force \( F_0 \). When curved, the string’s end-to-end distance shortens by \( L_{\text{tot}} - F_0 L_{\text{tot}}(dx / \cos \theta(x)) \), where \( \theta \) is the angle relative to straight. Shortening does work against whatever external mechanism is supplying the tension force. Making small-angle approximations gives the work done against the outside force when the string is slightly curved as \( \frac{1}{2} F_0 (\partial u / \partial x)^2 \) per unit length.

Thus, in the continuum limit the total linear density of energy at \( t, x \) (units \( J/m \)) is

\[ \rho_E^{(1D)}(t, x) = \frac{1}{2} \rho_m^{(1D)} ((\partial u / \partial t)^2 + c_0^2 (\partial u / \partial x)^2). \]

If you pluck just one mass in the chain, you’ll create a localized contribution to the second term in this expression, which then partially transforms to kinetic form and spreads. That energy cannot just vanish somewhere and pop up far away! Instead, energy flows with a 1-dimensional flux \( j_E^{(1D)} \) (units \( J/s \)). To find that flux, note that the net rate at which energy gets transported from left to right across a fixed position on the spring is the rate at which work is done on the right side of that point by the left side. This is the product of velocity (which is transverse) times the transverse component of force, so \( j_E^{(1D)} = -F_0 (\partial u / \partial x)(\partial u / \partial t) \).

**Your Turn 27C**

Use Equation 27.2 to show that for any solution of the wave equation,

\[ \frac{\partial \rho_E^{(1D)}}{\partial t} + \frac{\partial j_E^{(1D)}}{\partial x} = 0. \quad \text{continuity equation for energy, newtonian spring} \]

Similarly to the continuity equations for mass and charge, your result expresses the fact that energy is locally conserved: In order to change energy density at a point (first term on the left side), there must be an imbalance in the fluxes on either side of that point (second term on the left side).

Also similarly to the case of charge, integrating the continuity equation over space yields a global conservation law (Equation 8.6, page 118).

**Your Turn 27D**

Now repeat the analysis to find the density and flux of transverse momentum and prove an appropriate continuity equation relating them.

---

1We are assuming an inextensible string, so its contour length does not change.
27.6.2 Energy and momentum densities of a traveling wave

For the solutions given in Equation 27.3, the energy density is

\[ \rho_{E}^{[1D]} = \frac{1}{2} \rho_{m}^{[1D]} \bar{u}^2 (\omega^2 + c_s^2(\omega/c_s)^2) \sin^2(\omega t - (\omega/c_s)x). \]  (27.7)

Note that the kinetic and potential energy terms are in phase. They’re both nonnegative, but both drop to zero twice per cycle, at \( t = \omega x/c_s + n\pi \) for integer \( n \). At these “dead spots,” even the energy flux is zero, because

\[ j_{E}^{[1D]} = -\rho_{m}^{[1D]} c_s^2 (\omega/c_s) \omega \bar{u}^2 \sin(\omega t - (\omega/c_s)x) \]

falls to zero at the same places as Equation 27.7. How can energy flow to the right if there are spots where its flux is zero? To answer, note that at a node, where energy density is zero, the gradient of flux is nonzero. The continuity equation says that energy arriving from the left of that point begins to pile up there. So that point stops being a point of zero energy density, and so on.

27.7 PLUS ULTRA

The preceding section started with expressions for energy density and flux that were nearly obvious, then showed that they obey a continuity equation. Later, we will wish to understand the energy density and flux of electromagnetic fields, which are not so obvious in form. To find the right expressions, we'll work backward, and seek quantities that at least obey continuity equations. Then we'll still need to prove that our proposal is consistent with specialized results that we already obtained.

PROBLEMS

27.1 Slinky

Consider a stretched distributed spring of mass density \( \rho_{m}^{[1D]} (~ \text{kg/m}) \) and stretch modulus \( c_s (~ \text{N}) \). Rederive the results of Section 27.6 for the case of longitudinal (compression) waves.
CHAPTER 28

Einstein’s Version of Relativity: Overview

Failure to appreciate the role of the structure of Indo-European languages in affecting perception has repeatedly led western science into error. The “luminiferous æther” of classical physics was created for the express purpose of standing as a subject of the verb “to wave.”

— Garrett Hardin

28.1 FRAMING: CONSERVATIVE REVOLUTION

Here is an overview of what we’re going to cover, stated without any equations or even diagrams. The ideas won’t be precise, however, until embodied in equations and diagrams. That comes later.

The Principle of Relativity seems experimentally valid for any system that can be isolated from the rest of the world. Newtonian physics has an overarching mathematical property (galilean invariance) that transcends details of particular springs, clocks, and so on and that guarantees that any system fitting the framework will obey the Principle of Relativity. Chapters 26–27 showed that one way to expose that property is to

- See how the equations change their form when expressed in a different coordinate system on spacetime,
- Identify a subfamily of coordinate systems among which the form does not change, and
- Observe that some of those good coordinate systems are in uniform, straight-line motion relative to the others.

The next chapters will outline how Einstein retained much of the preceding framework—indeed bringing it into sharper focus while adjusting some details. Later, we’ll see how this fundamentally conservative approach nevertheless led to revolutionary insights.

Electromagnetic phenomenon: Vacuum is a unique state; it has no measurable descriptors analogous to the density or velocity of a medium that carries sound waves.

Physical idea: Electromagnetic fields require no such material medium.

28.2 THE ÆTHER HYPOTHESIS

Christiaan Huygens proposed a wave theory of light in 1690. This idea was soon sidelined, however, by a particle theory proposed in Newton’s Opticks (1704), only to be revived
in the 19th century, as interference phenomena became more inescapable. This trend culminated with Maxwell’s discovery of a wave equation inherent in electromagnetism; Hertz and many others firm up the evidence that this radiation was the same as light.

Neither Maxwell nor anyone else at that time believed that the equations were fully general: At best, they were regarded as correct in a coordinate system at rest relative to an omnipresent medium called the luminiferous æther. People believed this because of a general sense that waves could only move through a medium. (“How do you have ripples, without the pond?”) Tacitly the words “material medium” implied a substance that itself had states of motion, like air, water, or a string. Obviously the state of motion of the medium would have to enter the fully general equations of electromagnetism, as it does for the equations of sound, water waves, or string vibrations (Section 27.4).

But the æther had to have some weird properties. It had to be completely unaffected by any vacuum pump ever invented, because light travels just fine through vacuum. It had to be present throughout the space between planets, yet exert no frictional drag on them. It had to be rigid, like steel and unlike air, in order to support transverse waves. It had to be incompressible, because if not, there would also be a longitudinal polarization of light (compression waves), as there is for waves in air or steel. Yet the planets had to plow through it effortlessly.

Stepping back from details, a major problem with the æther was that it did no other job than the one for which it was introduced (transmitting light). In contrast, air transmits sound, but it also has other measurable attributes giving rise to other phenomena, for example, its mass density, temperature, pressure, viscosity, and so on; moreover, these attributes can be changed by experimental interventions.

Why were people so desperate to cling to this crazy idea? We can look back and say, a bit more clearly than was said at the time, that people also expected that all laws of Nature must be form-invariant under rotations, translations, and galilean boosts. The Maxwell equations as stated do not have the last of these properties, but it was assumed that after generalizing them to include the possible motion of the æther, they would.\(^1\)

Einstein found too many logical problems with this position, not least his and others’ inability to find an acceptable set of galilean-invariant equations as candidates to generalize Maxwell.\(^2\) Even setting aside this formal objection, modifying the wave equation to account for æther motion did not produce any theory consistent with all experiments. For example:

- When an object moves through an incompressible fluid, it sets the fluid into motion. Lab-based experiments looking specifically for the consequences of such æther entrainment came out negative.
- Moreover, the observed “aberration of starlight” was comprehensible only if Earth did not drag the æther (Chapter 30).

\(^1\)See Chapter 27.

\(^2\)Although the wave equations for sound and light are formally similar, they have quite different origins. If you propose a modification to the electromagnetic wave equation, you can’t stop there: You must also propose a modification to the full set of Maxwell equations that gives rise to your proposed new wave equation and agrees with experiments. This is what Einstein and others could not do.
28.3 The No-æther Hypothesis

28.3.1 The vacuum is a unique state

So Einstein entertained the bizarre suggestion that the Maxwell equations were actually correct and complete as written.\(^5\)

- To the objection that they lacked galilean invariance, he said, perhaps experiments don’t demand such invariance after all; perhaps the equations have some other invariance. In fact, Einstein found that a different invariance, already considered by others, was exact and good enough to satisfy the demands of experiment, including the Principle of Relativity.
- To the objection that replacing galilean invariance with Lorentz invariance had bizarre consequences, Einstein asked, are those consequences actually ruled out by experiment? For example, is there really any feasible method to measure absolute simultaneity? If not, then it’s not so disturbing if theory predicts that different inertial observers will disagree about the simultaneity of two events not located at the same point in space.
- To the objection that Newton’s laws are incompatible with Lorentz invariance, Einstein’s attitude was, we need to reexamine the experimental status of Newton’s laws.

The preceding discussion has carefully avoided saying that “The æther does not exist.” It is not really very scientific to claim the nonexistence of a poorly defined thing. Indeed, one sometimes hears that the quantum vacuum “is” the æther. Einstein would not object. His proposal merely amounts to saying that the vacuum—the state you can approach experimentally by using better and better vacuum pumps, or by going into interstellar space—is unique. Its properties (such as the values of \(\mu_0\) and \(\varepsilon_0\)) are constants. It has no

---

3 See Section 28.2a (page 432).
4 See Your Turn 27B (page 424).
5 These ideas were already in the air; see Section 28.3 (page 437).
further state variables beyond \( \vec{E} \) and \( \vec{B} \) that need to appear in the Maxwell equations, and
in particular no states of motion. (More precisely, it is Lorentz-invariant.) If you want to say it’s filled with an “æther” of virtual particles and antiparticles, fine, but it’s not the material substance that Maxwell and his contemporaries had in mind.

In other words, Einstein convinced himself that there’s no logical need for any æther. Maxwell equations don’t need it. It’s only our brains, trying to make inappropriate analogies to experience, that want it. We can’t intuitively imagine the EM field, nor the vacuum which it disturbs. The birth of the modern viewpoint came when Einstein said (paraphrasing), “That’s OK—I don’t need to imagine it intuitively.”

28.3.2 Follow the symmetry

Instead of attempting to modify the Maxwell equations, Einstein’s clarified a mathematical property (a new invariance) already hiding in them.\(^6\) Then he proposed that all the rest of physics had this same invariance, for example, the mechanisms inside clocks. All his “thought experiments” were mainly attempts to see if his proposal was obviously ruled out by existing knowledge. Over and over, he found that potential objections (paradoxes) were based on assuming some procedure that could not in fact be implemented experimentally (for example, knowing the reading on a distant clock instantaneously).

Then Einstein asked if his proposal made any characteristic, quantitative predictions that were testable. We’ll never know how much he really knew about Michelson–Morley; what he explicitly stated years later was that he relied on the aberration of starlight, and the Fizeau experiment, as sufficient to show he was on the right track. Not coincidentally, both of these concerned—Electromagnetic Phenomena. So we’ll discuss them in detail in the following chapters.

28.4 WHERE WE ARE HEADING

Anyone can open Einstein’s 1905 paper, copy out the transformations of the fields (updating the awful notation), substitute into the Maxwell equations, and show they are indeed an exact invariance. But after that exercise, we are still stumped—how could any human have figured that out?\(^7\) Instead we will take a longer route, following Minkowski and others: We will build a system of thought and notation in which the invariance of the Maxwell equations (and other relativistic field theories), becomes obvious at a glance. That way, even mortals like us can create new relativistic field theories, for example the ones needed to describe the strong and weak nuclear forces.\(^8\)

---

\(^6\) Lorentz had established this approximately in 1904.

\(^7\) Public-key encryption could be an apt metaphor for this situation!

\(^8\) It still required some more of Einstein’s personal genius to adapt the ideas to gravitation. And even Einstein needed the benefits of tensor notation before he could succeed.
FURTHER READING

Semipopular:
This fellow, and his gadget, are brilliant: https://www.youtube.com/watch?v=1rLVZVWfdY.
But we’ll need to flesh the ideas out a bit.

Intermediate:
For the next few chapters: Gray, 2022; Griffiths et al., 2022; Mermin, 2005.
Section 28.2’a (page 432) studied reflection from a moving mirror in a hypothetical, nonrelativistic world. For the relativistic case: Tallents, 2023.

Technical:
Einstein’s article: Einstein, 1905; Einstein, 1998; Kennedy, 2012.
History: Michelson & Morley, 1887; Darrigol, 2022; Pais, 1982.
21st century version of the 1887 Michelson–Morley experiment: Müller et al., 2003.
Modern reanalysis of the original 1887 data: Handschy, 1982.
28.2' \textit{Michelson–Morley 1887}

The main text did not discuss the famous MM experiment in part because it was not on Einstein's list of the two most decisive experiments, but also because it has a number of subtleties. Here is a discussion emphasizing the symmetry viewpoint, which helps to clarify those subtleties.

Before we can claim to have disproved a hypothesis, we must first make it precise. We wish to find a prediction for the Michelson-Morley experiment starting from the assumption that all physics is galilean invariant, then show that the experimental result falsifies that model by disobeying the prediction. Specifically, we will test the hypothesis that light is a vibration in a fluid governed by Newtonian mechanics.

\textbf{Setup}

We consider the usual simplified version of the apparatus sketched in Figure 28.1.\textsuperscript{9} Choose coordinates such that the incoming beam travels along the $-x$ axis. We will assume that both arms of the interferometer have the same length $b$ (a meaningful statement in nonrelativistic physics). The half-silvered mirror (henceforth "half-mirror") is oriented at 45 deg to the $x$ axis.

Before 1905, most physicists implicitly assumed something like the following set of claims:

- There is at least one G-inertial coordinate system in which the æther is at rest (therefore also others, rotated or translated). In particular, the apparatus itself does not drag, entrain, nor otherwise disturb the æther. Call any of these a "wind-free" coordinate system.
- In a wind-free system, any light ray travels on a straight line at constant speed $c$. More precisely, nearly planar wavefronts of a light beam travel on trajectories that are straight lines in spacetime moving with slope $c$. This is certainly the case for sound and other waves in an isotropic medium: Once emitted, they propagate at a speed independent of their direction and of the source's motion.\textsuperscript{10}
- Interference of light beams split from a common precursor beam is governed by total transit time between the splitting and recombination. Because time is unaffected by galilean boosts, we can and will choose to compute transit times in a wind-free system. Instead of thinking about wave phase, we can equivalently imagine a nonperiodic waveform, a single plane wavefront ("blip"); then we ask about the arrival times of the two blips that traverse the two arms of the apparatus.
- The Michelson–Morley apparatus is completely rigid; its geometry is unaffected by æther motion. In particular, the apparatus is not changed when it is rotated, changing the direction of its motion relative to the æther.
- The half-mirror splits the blip in an event we will call $P$; part follows a reflection law to be derived in a moment. The other part of the beam passes through and, by the law of refraction, emerges traveling in the same direction as it entered.\textsuperscript{11}

\textsuperscript{9}Maxwell actually proposed an experiment like Michelson–Morley in 1879, but rejected it as impractical at that time. Michelson did a preliminary experiment in 1881 with an apparatus of the sort described here. For the famous experiment, MM created a more elaborate, folded light path to increase their sensitivity.
\textsuperscript{10}In other G-inertial coordinate systems, light will therefore travel with anisotropic velocity, by the galilean velocity addition formula.
\textsuperscript{11}We will not explicitly mention the time delay from passing through the glass that composes this element, nor the sideways beam displacement. These effects, and the compensator plate present in the actual apparatus, can be added to the discussion without affecting the conclusion.
Figure 28.1: **Light paths in a simplified MM apparatus.** The dashed outline shows the position of the half-mirror at a later time than the initial encounter at $P$. The half-silvered face is shown as a heavy edge line; weaker reflections from the air-glass interface are not shown. (Also, other reflected beams that do not end at $S$ are not shown.) Refractive bends upon entry to, and exit from, the half-mirror are not shown. Other mirrors are shown at their positions when the flash of light hits them. The angle $\alpha$ has been exaggerated for clarity. It would be 0 if the apparatus were at rest with respect to the æther. Finally, the original experiment also included a “compensator,” a fixed glass plate ensuring that both light paths traversed the same total length of glass.

Again, we will analyze the experiment in one of the special (wind-free) $G$-inertial systems. Suppose that the apparatus is moving along $-\hat{x}$ at speed $v = \beta c$ with respect to the stationary æther. For simplicity, suppose that it has been rotated to place one arm parallel to $\hat{x}$ (the “left arm”) and the other parallel to $\hat{y}$ (the “upper arm,” Figure 28.1). We now ask about the total transit times in each arm and how they depend on $\beta$. Although we cannot experimentally change the magnitude of $\beta$, we can reverse its sign by rotating the apparatus by $\pi$.

Everything prior to $P$ is common to both trajectories, so we can neglect that and let time begin at event $P$.

**Left arm transit time**

The “left” part of the beam passes through the half-mirror (at event $P$) and overtakes the left mirror (which is moving away from it), hitting it at event $Q$. It arrives at $0$ deg from perpendicular, so it returns at $0$ deg from perpendicular and eventually hits the half-mirror (which is moving toward it) at event $R$. Later, the part reflected at $R$ hits a projection screen $S$.

Let’s find the transit time for the isolated blip to traverse $PQR$. Notice that $P$, $Q$, and $R$ all sit in the plane $y = 0$. Choose coordinates in which $P$ is the point with $[x^2] = [0]$ (Figure 28.2). Then $Q$ is determined by the following two equations, which involve an unknown parameter $\xi$:

$$Q = b \begin{bmatrix} 0 & \xi \\ \xi & 1 \end{bmatrix} = \begin{bmatrix} ct_Q \\ -\beta c t_Q \end{bmatrix}.$$ 

Solving for $\xi$ and $t_Q$ yields that $Q = \begin{bmatrix} b/(1-\beta) \\ -b\beta/(1-\beta) \end{bmatrix}$.

Next, $R$, the intersection of the half-mirror with the reflected light trajectory, is determined by the following pair of equations:

$$R = b/(1-\beta) + \eta \begin{bmatrix} ct_R \\ -\beta ct_R \end{bmatrix}.$$

Solving for $\eta$ and $t_R$ yields $R$, whose time coordinate is the total transit time for $PQR$: 
Figure 28.2: [Space-time diagram.] **Light path in the left arm of MM apparatus**, for \( b = 1 \text{ m} \) and the exaggerated case \( \beta = 0.15 \). **Solid red**: Path of light wavefront (at \( \pm 45 \degree \)). **Dashed purple**: Trajectory of half-silvered beam splitter. **Dotted blue**: Trajectory of ordinary mirror on the left. Delays from passing through glass after \( P \) and before \( R \) are not shown.

**Your Turn 28A**

Show that

\[
R = \frac{2b}{c} \frac{1}{1 - \beta^2}.
\]

The time at point \( R \) is slightly different from the case of \( \beta = 0 \) (\( cR = 2.05 \text{ m} \) for the example in Figure 28.2, versus 2 m). We will return later to the final segment (after reflection by the half-mirror), from \( R \) to the projection screen at \( S \).

**Reflection from a moving mirror**

Both arms of the light path involve reflection from moving mirrors (Figure 28.1). Here we simplify by studying a fully-silvered mirror; a half-silvered mirror would reflect the same way but with reduced amplitude.

In the upper arm of the apparatus, the blip enters horizontally from the right heading left (from positive to negative \( x \)). As shown in the figure, the angle of incidence on the splitter at \( P \) is 45 deg. After reflection, the blip proceeds at some angle \( \beta \) to its original direction. Thus, its angle of reflection is \( \pi/4 + \alpha \), where \( \alpha \) and \( \beta \) are related by

\[
\pi/4 + \alpha = (\pi - \beta) - \pi/4, \quad \text{or} \quad \beta = \pi/2 - \alpha.
\]

For a stationary mirror, the usual law of reflection gives \( \beta = \pi/2 \), hence \( \alpha = 0 \), but let’s keep an open mind.

Let the incoming wavefront be described by \( f(-x - ct) \), where \( f \) is a function with a peak at 0. Then the reflected wave is described by \( g(-x \cos \beta + y \sin \beta - ct) \), where \( g \) is a function that we must find. The boundary condition at a conducting surface (mirror) requires that the outgoing and

---

12The key idea for the following argument is discussed in Soni, 1988.
incoming waves be related near \( P \) by \( f = -g \) along the moving surface \( \{ y = -x - \beta ct \} \). Hence,

\[
f(-x - ct) = -g(-x \cos \theta + (-x - \beta ct) \sin \theta - ct)
= -g\left((1 + \beta \sin \theta)(-x \cos \theta + \frac{x \cos \theta + \sin \theta}{1 + \beta \sin \theta} - ct)\right).
\]

For this relation to hold at all \( x \) and \( t \), we must have that \( \cos \theta + \sin \theta)/(1 + \beta \sin \theta) = 1 \), or

\[
\sin \alpha + \cos \alpha = 1 + \beta \cos \alpha.
\]

(In addition, \( g \) is Doppler stretched: \( g(u) = -f(u/(1 + \beta \cos \alpha)) \).) Expanding Equation 28.2 for small \( \alpha \) and \( \beta \) gives that \( \alpha \approx \beta \), that is, not zero. Indeed, this extra angle is just what is needed for the outgoing ray to land on the displaced position of the upper mirror (\( T \) in Figure 28.1), at time \( b/c \).

A similar argument with reversed sign applies to the encounter of the left beam with the half-mirror at \( R \), and shows that upon this final reflection it, too, emerges heading toward the point where \( S \) will be located.\(^{13}\)

**Upper arm transit time**

Now that we know the light path in the upper arm, we can find the elapsed time from \( P \) to \( R \) and compare to the left arm. Because we are in a wind-free coordinate system, the time is just path length divided by \( c \), or

\[
\frac{2b}{c} \approx (2b/c)(1 + \frac{1}{2} \beta^2).
\]

Subtract this from Equation 28.1 to find the predicted difference between left and upper arm transit times:

\[
\approx (2b/c)(1 + \beta^2 - (1 + \frac{1}{2} \beta^2)) = b\beta^2/c \neq 0.
\]

**Consequences**

The transit times therefore differ by a \( \beta \)-dependent amount. Michelson and Morley designed their experiment so that the small predicted difference, if present, would be measurable via interferometry. Their observation of no transit time difference, regardless of orientation, implies that \( \beta = 0 \): we must conclude that if this model is correct, we are at rest with respect to the æther. Prior to Einstein, various unpalatable alternatives were entertained:

- By an amazing coincidence, Earth is at rest with respect to the cosmic æther.
- Earth entrains a layer of æther, so it’s at rest with respect to the local æther. Experiments done on mountaintops to reduce this possible effect showed no difference from sea level.
- The apparatus is not rigid but instead gets shrunken by the æther wind, coincidentally by exactly the amount needed to cancel Equation 28.3.

The whole problem goes away in relativistic physics, that is, when we deny the existence of any æther! But as discussed in the main text, that proposal required a new and unfamiliar replacement for the galilean invariance assumed in the preceding discussion.

\(^{13}\)There are no deviations from the usual law of reflection at either of the regular mirrors because, unlike the splitter, each is moving parallel or perpendicular to its surface.
Postscript

The traditional “swimmer” analogy is potentially misleading: Anthropomorphizing risks confusion because a real swimmer is self-propelled, unlike light, and may even have a “goal” to reach a particular point on the shore. Light has no goal and does not “aim” for anything.

28.2'b More about uniqueness of the vacuum state

The main text asserted that the vacuum has no user-adjustable properties. Like any bedrock principle in science, this one is more subtle, and more subject to fine interpretation, than it may seem.

The empty space outside the pole of a magnet in vacuum does have a “property” (the static magnetic field), which is attached to specific points in space in that region. Physics in that region of space is not isotropic and hence not Lorentz-invariant. So it’s more precise to say that only a region of vacuum that is far from or shielded from any matter is universal, including its ability to carry EM fields (or planets), should they be introduced. When charged matter is present, we attribute its effects to a deviation from field-free vacuum (the EM field) whose dynamics is invariant under a group of transformations, and so on with other kinds of interaction (strong, electroweak).

Remarkably, Einstein abandoned even this more limited statement a few years later when he formulated general relativity. He found that it proved fruitful to attribute gravitation directly to... user-modifiable properties of spacetime. Moreover, there is no such thing as “shielding” a region from gravitational fields (Section 26.1, page 405), and no region in space “far enough” from gravitation to be unaffected by it; indeed, the whole expansion of the Universe is controlled by gravitation.

Nevertheless, the statements made in this chapter are still accepted today, in the following sense:

Far enough from any gravitating bodies, Einstein's general theory predicts the existence of special coordinate systems (“locally inertial” or “freely falling”), in which gravitational effects appear to be approximately absent and all the rest of physics, including electrodynamics, has the properties discussed in this chapter.

For example, the locally inertial systems are related to one another by ordinary Lorentz transformations; those transformations are invariances of all the non-gravitational dynamics; and so on. Section 34.10 will return to this train of thought. Ultimately it led to a combined theory of gravitation and other interactions that, although still not integrated fully with quantum mechanics, nevertheless has been successfully extrapolated to make predictions about physics even close to massive objects.

28.2’c In praise of æther

The main text may have sounded scornful of the æther hypothesis. In fact, it played a crucial transitional role in the development of electrodynamics. On the Continent of Europe, most theorists sought explanations based on actions at a distance between charges. Faraday and his successors placed the emphasis on something real in the vacuum between charges. “Maxwell seems to have regarded his main task to have been the transformation of Faraday’s theory into a newtonian mechanical theory” [Chalmers, 1975]. The road to the field viewpoint had to pass through an almost-right waystation, the æther models.

November 14, 2023; Contents Index Notation Glossary
28.3’ Poincaré’s work

In a semipopular book published in 1905, Poincaré wrote

1: “There is no absolute space, and we only grasp relative movement....
2: There is no absolute time; to say that two intervals are equal is a meaningless claim, which can acquire meaning only by convention.
3: Not only have we no direct intuition of the equality of durations, but we have not even one of simultaneity of two events produced at different locations.14
4: Finally, isn’t our euclidean geometry itself only a linguistic convention? ....
And so absolute space, absolute time, even geometry are not conditions imposed on mechanics; none of these preexist in mechanics any more than the French language logically preexists the truths that we express in French.”

Fairly or not, history has emphasized Einstein’s later versions of these statements, in part because he decisively rejected the æther hypothesis. Einstein also offered a clear, axiomatic basis for his claims, which allowed (some) scientists to see that there was a logically consistent alternative to æther. He also realized that Lorentz invariance must be a general, strictly, and exactly valid law of Nature.

14 Poincaré had already published a technical article making this claim in 1898.
28.1  Moving mirror
The steps leading to Equation 28.2 (page 435) helped us to understand reflection at event \( P \) in Figure 28.1. Adapt those steps to understand reflection at event \( R \) and confirm what’s claimed in that diagram.
CHAPTER 29

Provisional Lorentz Transformations and the Fizeau Experiment

When I try to make things clearer by a spacetime diagram, the other participants look at it with polite detachment and, after a pause of embarrassment as if some childish indecency had been exhibited, resume the debate in their own terms.

— J. L. Synge

29.1 FRAMING: DRAGGING LIGHT

We’ve seen that the wave theory of light has scored some successes, giving a detailed account of polarization phenomena (Chapter 18), the transport of energy and momentum (Chapter 20), and so on. But there is still a puzzle, which eventually led Einstein to some disturbing insights into space and time.

Electromagnetic phenomenon: The speed of light in flowing water is different from that in still water (the light is “dragged along”), but in a quantitatively different way from the newtonian expectation.

Physical idea: Nature does have a boost invariance, but it’s not the naïve one.

29.2 REVIEW

29.2.1 Galilean invariance predicts simple addition of velocities

Chapter 26 argued that newtonian physics implements the Principle of Relativity by having an invariance under galilean boost transformations. One way to express this is by using the active view: If we have a system of particles and a solution to the equations of motion given by some functions $\vec{r}_{(1)}(t)$, $\vec{r}_{(2)}(t)$, ..., then the modified trajectories

$$\vec{r}'_{(1)}(t) = \vec{r}_{(1)}(t) + \vec{v}_s t, \quad \vec{r}'_{(2)}(t) = \vec{r}_{(2)}(t) + \vec{v}_s t,$$

will also solve the same equations. Here $\vec{v}_s$ is one overall constant vector.

The equivalent passive view relabels all the events in spacetime according to

$$\begin{bmatrix} ct' \\ x' \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -v_s/c & 1 \end{bmatrix} \begin{bmatrix} ct \\ x \end{bmatrix},$$

[26.10, page 411]

or a similar formula in three spatial dimensions. Chapter 26 showed that if we take some equations of newtonian physics (for example, two masses joined by a spring\(^1\)) and re-express them in terms of the primed coordinates, the new versions have the same algebraic

---

\(^1\) Or two masses with their newtonian gravitational attraction, and so on.
form as the old ones. Section 27.3 also showed that the wave equation does not have this
property, but Section 27.4 gave a resolution of that puzzle appropriate for vibrating strings,
sound waves, and water waves: The wave equation must be generalized to account for
possible motion of the medium relative to the observer.

Finally, Section 26.6.4 found a velocity addition formula, which can be stated in a
rather longwinded way as:

Suppose that we have a coordinate system on spacetime in which a wave or par-
ticle is moving at constant velocity \( v_0 \). Now introduce a new coordinate system
related to the first by a galilean boost with velocity \( \vec{v}_x \). The wave or particle will
be observed in the second system to be moving at constant velocity \( \vec{v}_0 - \vec{v}_x \).

Chapter 26 noted that light from distant objects comes to us at a velocity independent
of the source’s motion, and that this observation, together with Idea 29.1, rules out any
galilean-invariant theory of light as a stream of material particles.

On the other hand, sound or water waves do move at a speed independent of the
source: Imagine running your finger just above the surface of a ripple tank and period-
cally dipping it into the water. Each ripple you cause moves outward at a fixed speed
(independent of how fast your finger is moving). That is, as long as the observer is at rest
relative to the medium, waves in a material substance move at a constant speed indepen-
dent of the motion of the source. Thus, the wave model of light seemed to explain why
each partner in a binary star system never appears doubled.²

But the apparent speed of a wave on water or air certainly does depend on the mo-
tion of the observer.³ In contrast, the speed of light also was found to be unchanged by
uniform, straight-line observer motion. After all, the Earth is hurtling through space, yet
the physics we see in a closed lab does not depend on orientation relative to that motion.⁴
This looks bad for the wave model of light. Einstein was alluding to this problem when
he mentioned the prior “failure of attempts to detect a motion of the Earth relative to the
‘light medium’.”⁵

### 29.2.2 Æther skeptics have some explaining to do

But it’s not enough just to say blithely, “Therefore there’s no æther.” After all,

- Eliminating the medium would also eliminate our rescue of galilean invariance (Sec-
tion 27.4).
- Galilean invariance is what guaranteed the Principle of Relativity, which is exper-
imentally validated.

---

²See Section 26.6.6 (page 414).
³See Section 27.4 (page 423).
⁴See Section 28.2’s (page 432). We do see effects of our motion when we look outside the lab at light from
distant stars (Chapter 30), but even in this case, the speed is fixed at \( c \).
⁵Einstein was never clear whether he was thinking specifically about the now-famous Michelson–Morley ex-
periment, but there were other such experiments at the time, and all (eventually) came out null. (In a letter,
Einstein mentions reading a review article by Wien in 1899 that described thirteen such experiments.)
In this chapter and the next one, we’ll see how Einstein reconciled Maxwell with the Principle of Relativity at the level of a single (scalar) wave equation, temporarily neglecting all the delicious complexity brought by the vector character of electromagnetic fields. As always, we’ll look to some key experiments for guidance.

Let’s pause to dispose of a red herring. Certainly there are bizarre coordinate systems we could choose in which a particular ray of light seems to move at a speed other than c. Simply take \( \hat{r}' = 2\hat{r} \), and leave time unchanged; in the primed coordinate system, light travels at speed \( 2c \). This mathematical fact is physically irrelevant because in the primed system, the equations of physics take nonstandard forms; for example, constants of Nature have different numerical values. We would know right away that something was wrong in the new system, for example, because atoms would have different apparent sizes than in our usual coordinate system. Our puzzle is that in newtonian physics, even the inertial coordinate systems (those in which the equations take their usual form) will disagree about wave speed if there is a material medium, but no such effect is observed for flashes of light in vacuum.

29.3 GRAPHICAL EXPLORATIONS SUGGEST A FORM FOR BOOST TRANSFORMATIONS

If, following Einstein, we suspect that the Maxwell equations are complete and correct as written (no æther), then what invariances do they have? Maybe they have some invariance that, while not galilean, nevertheless connects coordinate systems that

- are in uniform, straight-line motion relative to each other, yet nevertheless
- also agree on the experimental observation that the speed of light is always the constant \( c \approx 3 \cdot 10^8 \text{ m/s} \).

That sounds like a contradiction, but in Einstein’s words, these two requirements are “only seemingly incompatible.” In fact, W. Voigt had already proved that the scalar wave equation was invariant under a family of such transformations in 1887. Einstein took this result seriously, and crucially, extended it from the scalar wave equation to the full Maxwell equations and then to all of Physics.

We can think graphically about the galilean transformation (Equation 26.10, page 411) as introducing a new set of coordinate axes on the \( xt \)-plane. Actually, it’s easier to think about the quantities \( x \) and \( ct \), because these have the same units, and because then a trajectory traveling at speed \( c \) is represented by a line at 45 deg to the horizontal. Such a trajectory is drawn as a solid diagonal line in the figure below:

---

6 This oversimplification will be remedied in Chapters 32 and 33.
7 See also Figure 26.2 (page 411).
The original $x$ and $ct$ axes are also shown as solid lines. The new $x'$ axis is the same as the $x$ axis: It’s the locus of events $\{t' = 0\}$, but $t' = t$ for a galilean boost. However, the new $ct'$ axis is bent over (the locus $\{x' = 0\}$, dashed line). We see graphically that the trajectory shown bisects the right angle between $x$ and $ct$ axes, but doesn’t bisect the acute angle between $x'$ and $ct'$ axes: It changes the apparent speed of light.

We have experience with another sort of linear transformation in the plane: a rotation of the axes. Figure 29.1a shows this option. That transformation also alters the apparent slope of the trajectory shown; again, the trajectory does not bisect the angle between $x'$ and $ct'$ axes. But there is a third possibility (Figure 29.1b): If we bend both axes by opposite angles, then the diagonal line continues to bisect the angle between $x'$ and $ct'$ axes.

Your Turn 29A

Think about the other allowed light trajectory in 1D, which moves at speed $-c$. It bisects the angle between the $-x$ and $+ct$ axes. Convince yourself geometrically that it also bisects the angle between the $-x'$ and $ct'$ axes in the diagram analogous to Figure 29.1b.

29.4 THE WAVE EQUATION IS INARIANT UNDER PROVISIONAL LORENTZ TRANSFORMATIONS

29.4.1 Coordinate transformation

Figure 29.1b represents the following linear transformation of coordinates:

$$
\begin{bmatrix}
    ct' \\
    x'
\end{bmatrix} = \gamma
\begin{bmatrix}
    1 & -\beta \\
    -\beta & 1
\end{bmatrix}
\begin{bmatrix}
    ct \\
    x
\end{bmatrix},
$$

provisional Lorentz boost transformation \hspace{1cm} (29.2)

Here $\beta$ and $\gamma$ are constants, and $\gamma > 0$. Equation 29.2 says “provisional” because, although we’ll find that all transformations of this form are invariances of the vacuum wave equation, we’ll also see that only a subset are invariances of the rest of physics (or even of the full Maxwell equations).\(^8\)

Equation 29.2 has a feature that bothered many people: $t' \neq t$. To many, it seemed necessary that all good coordinate systems would agree on one correct, universal choice for time. Einstein realized that this was a prejudice without experimental justification.\(^9\)

\(^8\)Chapter 30 will argue that the true Lorentz transformations are the special case with $\gamma = (1 - \beta^2)^{-1/2}$, but we don’t need that level of detail yet.

\(^9\)Section 29.7 will expand on this point.
29.4 The Wave Equation is Invariant Under Provisional Lorentz Transformations

29.4.2 Active viewpoint

To see whether Equation 29.2 is at least promising, consider a harmonic traveling wave solution to the wave equation: \( \phi_{\pm}(t, x) = \cos\left(\frac{\omega}{c}(-ct \pm x)\right) \). Following Idea 26.6 (page 407), we apply an active transformation, that is, construct different functions \( \tilde{\phi}_{\pm} \) defined by \( \tilde{\phi}_{\pm} = \cos\left(\frac{\tilde{\omega}}{c}(-ct' \pm x')\right) \), or

\[
\tilde{\phi}_{\pm}(t, x) = \cos\left(\frac{\bar{\omega}}{c}(-ct + \beta x) \pm \gamma(-\beta ct + x)\right),
\]

which can be written as

\[
\cos\left(\frac{\tilde{\omega}}{c}(-ct \pm x)\right), \quad \text{where } \tilde{\omega} = \omega \gamma(1 \pm \beta).
\]

In contrast to the galilean case, these functions are again solutions to the wave equation, with wavecrests still traveling at speed \( \pm c! \) It’s true that each has a different frequency from the original, but we expected that—there should be a Doppler shift. Because any solution to the wave equation can be expanded in Fourier series, we have established active symmetry under provisional Lorentz transformations.

29.4.3 Passive viewpoint

Encouraged by that result on a particular solution, we now start over from the passive viewpoint, that is, we focus on the wave equation itself, not its solutions:

**Your Turn 29B**

Re-express the wave equation in terms of primed coordinates. (Follow the passive viewpoint in Section 27.3, page 421, but with the new transforms Equation 29.2 instead of galilean boosts.) Show that the wave equation maintains its original form after this passive transformation.

In short, the wave equation is invariant under a family of transformations that take a coordinate system and boost it into uniform motion relative to the original one. Hence, the wave equation is still compatible with the Principle of Relativity—just not in the way people had expected.
29.5 EINSTEIN’S VELOCITY ADDITION

Let’s revisit the problem of a particle ejected from a moving catapult (Section 26.6.6), but this time, assume invariance under provisional Lorentz transformations. Following the Example on page 414, we suppose that when the catapult is at rest, it fires a projectile into uniform motion with velocity \( \mathbf{v}_x \). The corresponding trajectory can again be written in parametric form as \([ \xi' \equiv \xi / c, \chi' / \beta c \equiv \chi / c] \), and that of the stationary catapult itself as \([ \xi' \equiv \xi / c, \chi' / \beta c \equiv \chi / c] \).

We then apply an active boost transformation with velocity \( \mathbf{v}_s = \beta \mathbf{c} \) to conclude that there must be another solution, in which the primed coordinates are given by the same functions as appeared in the original solution.\(^{10}\) Thus, the catapult trajectory is \([ \mathbf{v}_s / \gamma / (1 - \beta^2)] \), and the projectile’s is \([ \mathbf{v}_s / \gamma / (1 - \beta^2)] \). Now express these in terms of the original coordinates:

\[
\begin{align*}
\text{catapult:} & \quad \mathbf{y} \begin{bmatrix} 1 & -\beta \mathbf{c} / \gamma (1 - \beta^2) \\ -\beta & 1 \end{bmatrix} \begin{bmatrix} \xi \\ \chi \end{bmatrix} = \begin{bmatrix} \xi' \\ \chi' / \beta c \end{bmatrix} , \\
\text{projectile:} & \quad \mathbf{y} \begin{bmatrix} 1 & -\beta \\ -\beta & 1 \end{bmatrix} \begin{bmatrix} \xi \\ \chi \end{bmatrix} = \begin{bmatrix} \xi' \\ \chi' / \beta c \end{bmatrix} .
\end{align*}
\]

Multiply both sides by the inverse matrix:

\[
\begin{align*}
\text{catapult:} & \quad \begin{bmatrix} \mathbf{c} / \gamma (1 - \beta^2) \\ \mathbf{c} / \gamma (1 - \beta^2) \end{bmatrix} \begin{bmatrix} \xi \\ \chi \end{bmatrix} = \begin{bmatrix} \xi' \\ \chi' / \beta c \end{bmatrix} , \\
\text{projectile:} & \quad \begin{bmatrix} \mathbf{c} / \gamma (1 - \beta^2) \\ \mathbf{c} / \gamma (1 - \beta^2) \end{bmatrix} \begin{bmatrix} \xi \\ \chi \end{bmatrix} = \begin{bmatrix} \xi' \\ \chi' / \beta c \end{bmatrix} .
\end{align*}
\]

Thus, both catapult and projectile are in uniform motion with respect to the ground: The velocity of the catapult is \( \mathbf{v}_c = \beta \mathbf{c} \) as desired, whereas that of the projectile is

\[
\mathbf{v}_p = \mathbf{c} / \Delta t = (\beta \mathbf{c} / (1 + \beta / \gamma)).
\]

Equation 29.4 is a disturbing result. It surely doesn’t look like the galilean formula \( \mathbf{v}_p = \beta \mathbf{c} + \mathbf{v}_s \). But suppose that \( |\beta| \ll 1 \) and \( |v_s| \ll c \); in this limit, we can forget the denominator, and we do recover galilean behavior. In the everyday world of things moving much more slowly than light, Einstein’s kinematics resemble the galilean behavior. This is the world in which we formed our intuitions over millions of years of evolution: Throw a spear while running forward, and the spear’s velocity will be the sum of your arm velocity and how fast you’re running (better able to bring down that gazelle).

We call the low speed world the nonrelativistic limit of the general situation. The word may be puzzling: It does not mean that Principle of Relativity is false in this limit, but rather that the distinction between Einstein’s and galilean relativity becomes unnoticeable.

In the opposite, less familiar, regime where \( \mathbf{v}_s \to \pm \mathbf{c} \), our formula gives \( \mathbf{v}_p \to \pm \mathbf{c} \).

A trajectory that moves at speed \( \mathbf{c} \) in \((t', \mathbf{x}')\) has the same property in \((t, \mathbf{x})\). As we saw in Figure 29.1b, Lorentz invariance reconciles our desire to connect coordinate systems in uniform, relative motion (and hence hardwire the Principle of Relativity), with the universality of the speed of light required by the Maxwell equations.\(^{11}\)

---

\(^{10}\)Section 26.4.2 (page 407).

\(^{11}\)This observation eliminates an objection we made to the particle picture of light in Chapter 26: Regardless of how an astronomical object may be moving relative to us, light leaving it always travels toward us at speed \( \mathbf{c} \).

Although this book mostly focuses on the wave picture, the fact that both viewpoints are experimentally tenable underpins the dual nature of light revealed in quantum field theory.
Finally, you should think about the limit $\beta \rightarrow 1$, holding $v_\alpha$ fixed to some value less than $c$. Figure 29.2 shows this and every other case graphically.

Because every provisional Lorentz transformation preserves the form of the wave equation, the combined effect of two such transformations in succession will have the same property.

**Your Turn 29C**

a. Suppose that a boost with $(\gamma_1, \beta_1)$ is followed by another with $(\gamma_2, \beta_2)$. Show that the combined transformation is again of the form Equation 29.2, with new values for $\gamma$ and $\beta$.

b. Find the inverse of the transformation Equation 29.2 and show that it, too, is a provisional Lorentz transformation.

Thus, our provisional Lorentz transformations form a group, analogous to but distinct from the galilean group (Section 26.6.4). Just as in newtonian physics, we can promote everything to three space dimensions, again obtaining a group of invariances. When we finish specifying the relation between $\gamma$ and $\beta$ in Section 30.3, this group will be called the “Lorentz group.” But already we can find a connection to experiment.

### 29.6 A NONNULL, FALSIFIABLE PREDICTION

#### 29.6.1 Fizeau’s experiment

Later in his life, Einstein said that *just two* experimental observations were all he needed to be convinced he was on the right track. They were (*i*) the aberration of starlight and (*ii*) an experiment first done in 1859 by M. Fizeau, then redone with greater precision by
A. Michelson and E. Morley in 1886.\(^\text{12}\) We’ll discuss the second of these now, and the first in Chapter 30.

It is sometimes said that the later, more famous Michelson–Morley experiment falsified the æther/galilean hypothesis. But one problem with it is that it was a null result; the result was zero dependence of light speed on apparatus velocity, whereas the æther/galilean theory predicted a nonzero result (Section 28.2\(^\text{a}\), page 432). Null experiments are subject to the criticism that zero is a very special value. There may be various explanations for why you got zero (maybe your sensitivity wasn’t as good as you thought).\(^\text{13}\)

It’s more convincing when two theories make different, nonzero predictions for an experimentally observable quantity, and an experiment excludes one but not the other. Fizeau’s experiment had that character. Before doing it, Fizeau first measured the speed of light in air, finding near-agreement with Rømer’s older astronomical measurement.\(^\text{14}\)

Then he measured the speed of light in water, finding it to be \(c/n\), where the refractive index \(n \approx 4/3\) for visible light. That was a comforting result: Huygens had shown that a slowdown of light in water was just what was needed to explain the law of refraction in the wave theory of light.\(^\text{15}\) But crucially, Fizeau proceeded to study the propagation of light in flowing water at various velocities, both along and against the direction of a light beam.\(^\text{16}\) He found that the motion of the water can “drag” light (slow it or speed it up).

### 29.6.2 Galilean versus relativistic predictions

Let’s apply the Relativity Strategy to this problem (Idea 26.15, page 414). Suppose that, as in Fizeau’s experiment, the water velocity is parallel or antiparallel to that of the light. Then:

- Whatever may be the equations governing light in water, we know that they have solutions in which the water is at rest and light flashes move at velocity \(\pm c/n\).
- The galilean hypothesis predicts other solutions in which water is moving at \(\beta c\) and light flashes at \(v_{\text{lab}} = \pm c/n + \beta c\).
- Equation 29.4 says that the hypothesis of provisional Lorentz invariance predicts solutions in which water is moving at \(\beta c\) and light flashes at \(v_{\text{lab}} = (\pm cn^{-1} + \beta c)/(1 + \beta^2(\pm cn^{-1})/c)\).

\(^{12}\)Many more tests of relativity came only after 1905, so were not available to Einstein, including a Fizeau-type experiment with still higher precision by Zeeman.

\(^{13}\)To get a definitive result, the MM experiment should have been, but initially was not, repeated at several times spaced throughout a year. This was not done until 1925 by Michelson’s successor D. Miller, who also placed his apparatus on a mountaintop to minimize entrainment of the æther—and obtained a nonnull result! He won a big prize for this erroneous conclusion. So the MM experiment was hardly a “proof of relativity,” as it is often portrayed, and it certainly was not unanimously viewed as such by contemporaries.

\(^{14}\)Light travels a tiny bit slower in air than in interplanetary space.

\(^{15}\)See Section 21.3.2 (page 309).

\(^{16}\)Later experiments used a chunk of quartz on the rim of a rapidly spinning disk to eliminate turbulence that occurs in water.
If the light flash is moving along the flow, use choose the upper sign in the last formula. Figure 29.3a shows that for positive $\beta$ the water indeed drags the light forward and this quantity is greater than $1/n$, whereas for negative $\beta$ the reverse is true. In experiments, $|\beta| \ll 1$, and we can make a simplified approximate version of the relativistic prediction:\footnote{Max von Laue published this formula in 1907.}

$$v_{\text{lab}}/c \approx (\beta + n^{-1})(1 - \beta / n) \approx n^{-1} + \beta(1 - n^{-2}).$$

(29.5)

As a check, suppose that we remove the water, that is, we set $n \to 1$. Then Equation 29.5 says that $v_{\text{lab}} \to c$. That was after all our starting point: The speed of light in vacuum must always equal $c$.

At last, we have a testable prediction. The hypothesis that the full equations of electromagnetism have galilean invariance predicts $v_{\text{lab}} = c(n^{-1} + \hat{\beta})$, which differs from Equation 29.5. If we plot $v_{\text{lab}}$ (speed of light in water, measured in the lab’s coordinate system) versus $\beta$, then the two competing theories make different predictions for the slope of the data. Thus, both theories make firm, nonnull predictions, with no adjustable fit parameters.\footnote{The value of $n$ can be independently measured from experiments on refraction. So it’s a parameter, but not a fit parameter.} That is, both are highly falsifiable, if you’ve got enough accuracy to measure the effect at all.

Figure 29.4 shows the data from Michelson and Morley’s improved version of Fizeau’s experiment. Following Fizeau, the authors used interferometry to get high enough sensitivity to small velocity changes.

Michelson and Morley made 65 trials of their experiment, varying both the tube length and water velocity. They actually measured the differences in light speed between propagation with and against the water flow. Because the graph shows only a small range of
values for (water speed)/c, the Einstein prediction appears to be nearly a straight line. (At water speed approaching c, Figure 29.2, page 445 predicts that the curve would start to level off.) Figure 29.4 shows significant scatter, but the data certainly rule out the prediction from galilean invariance (slope 2, dashed line),\(^{19}\) and they don’t rule out the prediction from provisional Lorentz invariance.

### 29.7 PLUS ULTRA

1. Our provisional Lorentz boost (Equation 29.2) has the disturbing feature that \(t' \neq t\). Many scientists objected: “How can time itself change?” We could reply on Einstein’s behalf:

   • I said nothing about time itself. I don’t know what time itself means. I have no apparatus to measure time itself. I have no access to any universal time standard.
   • I do have various kinds of devices called clocks. Because they are physical objects, they too are subject to the hypothesis that the equations governing them are invariant under (provisional) Lorentz transformations.
   • I do know ways to attach sets of four numbers to events.\(^{20}\) Some of these coordinate systems are “good” in the sense that in them, physics is described by simpler equations than in the others (and always by equations of the same form). The hypothesis is simply that the “good” coordinate systems are related to each other by

---

\(^{19}\)After Fizeau’s experiment was done, æther theorists tried to wriggle out of this failed prediction with a theory that we now regard as laughably contrived. But it’s best not to laugh—who knows which of today’s theories will also look comical in the future.

\(^{20}\)One way Einstein suggested to set up such a coordinate system is to use an array of identical clocks and synchronize them using light flashes.
transformations that include (at least some of) the ones given in Equation 29.2.

It is true that the new transformations imply that different, equally good, coordinate systems will disagree about whether two distant events are simultaneous \( t_{(1)} = t_{(2)} \) does not imply \( t'_{(1)} = t'_{(2)} \). But what experimental result does that contradict? As mentioned in Section 28.3.1, Einstein couldn’t find any.

We have seen that the hypothesis of invariance under these transformations implies a testable, and verified, prediction for a nontrivial phenomenon, the “dragging of light” by a moving medium. We’ll add more phenomena to this list later.

2. There’s a remarkable feature of the derivation in Section 29.6: Nowhere did we find it necessary to describe the mechanism for the slowing of light in water. That is, details of the dynamics did not enter, apart from the hypothesis that whatever the slowdown mechanism is, it is invariant under provisional Lorentz transformations.\(^{21}\) The kinematic approach followed above is much simpler\(^{22}\) than solving the Maxwell equations for light moving through a medium of water molecules.

3. Although provisional Lorentz invariance looks promising, we are far from being done: We must find invariances of the full Maxwell equations. Rather than attempt that head-on, we will first construct a new kind of tensor language in Chapters 32 and 33. The new language seems elaborate at first, but it makes many derivations of this sort very straightforward.

FURTHER READING

**Semipopular:**

**Intermediate:**
Lahaye et al., 2012; Zhang, 1997; Tallents, 2023.

**Technical:**
Historical: Fizeau, 1859; Michelson & Morley, 1886b. Einstein’s recollection that two key experiments were “enough”: Shankland, 1963; Shankland, 1964. Will, 2006a = arxiv.org/abs/gr-qc/0504085; Will, 2006b.

Fizeau drag in graphene plasmonics: Dong et al., 2021.

---

\(^{21}\)When we complete our specification of the Lorentz transformations in Chapter 30, the derivation will still hold, because \( \gamma \) drops out of this particular prediction.

\(^{22}\)Other scientists came close to relativity before Einstein. Today we regard their work as mostly unreadable, because they got bogged down in detailed dynamical hypotheses.
29.1 Sketchy
Confirm that the provisional Lorentz transformation (Equation 29.2, page 442) really implements the sketch Figure 29.1b and find the angles by which the $ct$ and $x$ axes are bent.

29.2 [[Not ready]]
Aberration of Starlight and Doppler Effects

And then, beside the Thames at Kew,
the house of Samuel Molyneux
supplied the firm foundations needed.
James Bradley, Samuel’s friend, succeeded
in tracking Hooke’s draconic star ...
The trouble was, it moved too far,
too fast, and in the wrong direction!
Despite the most minute inspection,
Bradley found nothing to suggest
his telescope was not at rest;
the star was shifting in the sky,
though maybe God alone knew why!

— James Muirden

30.1 FRAMING: A GREEDY PRINCIPLE

You showed in Your Turn 29B that a family of transformations leave the 1D wave equation invariant. Some of these were unsurprising (translations and reflections in space and time), but a two-parameter family called “provisional Lorentz boosts” were more interesting (Equation 29.2), in part because they relate two coordinate systems in uniform relative motion, and hence are candidates for implementing the Principle of Relativity. We also saw that every coordinate system in the family we are considering agrees about whether or not a trajectory is moving at speed c.

So it’s not true that Einstein said “everything is relative”: Rather, he proposed that

\[ \text{The property of moving at speed } 3 \cdot 10^8 \text{ m/s is absolute (all coordinate systems in an objectively “good” class agree about it).} \]

However, certain other properties then proved to be relative. For example, different “good” systems disagree about whether two events are simultaneous.\(^1\)

Chapter 29 stressed the value of predicting a testable, non-null effect that differs from newtonian physics; the present chapter will develop more predictions of this type. We’ll see that relativity is a greedy principle: once you give it a foothold, it takes over. First,

---

\(^1\)Section 39.2 will discuss this statement in detail, but we already see the idea in Figure 29.1b (page 443): The locus of points simultaneous with the origin in the unprimed system is the x’ axis, which differs from the locus of points simultaneous with the origin in the primed system.
however, we’ll refine our provisional form of our proposed transformations to get their final form.

*Electromagnetic phenomenon:* Each star’s apparent position is shifted relative to others, depending on Earth’s momentary velocity.

*Physical idea:* Lorentz transformation of the wavevector and frequency explains this phenomenon.

### 30.2 AGAIN NO DILATION INVARIANCE

The wave equation in vacuum is just one combination of the Maxwell equations. We’ll now see that some of the “provisional Lorentz” boosts are *not* invariances of all of electrodynamics. So we need to throw some of them out. But we must do so carefully: The subset we keep must form a **subgroup**, that is, the composition of two successive transformations in that subset must also be in it. (Also, the inverse of any one of them must be in the chosen subgroup.)

To see that some of the transformations we found are spurious, we’ll follow an approach like the one in Section 26.5 (page 408). Consider a situation that’s not just fields in empty space, specifically the Coulomb repulsion of two identical, charged particles:

\[
m \frac{\partial^2 \vec{r}_{(\alpha)}}{\partial t^2} = \frac{q^2 \vec{R}}{4\pi \varepsilon_0 R^3}, \quad \text{and so on, where } \vec{R} = \vec{r}_{(1)} - \vec{r}_{(2)}. \tag{30.1}
\]

Here the two point charges, labeled by \( \alpha = 1 \) or \( \alpha = 2 \), are assumed to be identical (the same charge \( q \) and mass \( m \)).

Now consider a class of provisional Lorentz boosts called **dilations**: \( \gamma \neq 1 \) but \( \beta = 0 \), that is,

\[\vec{r}' = \gamma \vec{r} \quad \text{and} \quad t' = \gamma t.\]

Rephrasing Equation 30.1 in terms of \( t' \) and \( \vec{r}' \), we find that in the new coordinates it does not have the same form as initially—there’s a factor of \( \gamma \) that fails to cancel.\(^3\)

Actually, we needn’t have worked so hard. If dilations were an invariance of the laws of Nature, then there would be hydrogen atoms of any size! In the active viewpoint, just apply a dilation to whatever solution corresponds to the usual atom, and find a new solution stretched by an arbitrary amount.\(^4\)

There are several attitudes we could now take:

- We could just try saying, “The charges and/or masses of the particles also change under such transformations.” But if the world had such an invariance, then there’d

---

\(^2\)Although this formula will later need relativistic corrections, it’s certainly valid for slowly-moving particles.

\(^3\)It’s true that the new equation has the same form except for the value of \( q^2/(m\varepsilon_0) \), but that’s not enough to declare that it’s invariant. Note that an equation of this sort also describes two uncharged particles attracting each other gravitationally, so newtonian gravity, too, lacks dilation invariance.

\(^4\)Atomic sizes involve quantum mechanics, but even in classical electrodynamics Chapter 47 will show that an electron’s ability to scatter radiation involves the “classical electron radius,” a length scale with a fixed value for every electron.
be a whole family of different electrons with continuously varying charges and/or masses. Nobody has seen them.\(^5\)

- Or we could try saying, “There is some new dynamical entity, implicitly set equal to a fixed numerical value in the Maxwell equations, which should rather be free and which transforms along with \(x\) and \(t\).” Maybe its transformation rule under dilations could be arranged to be exactly what’s needed to make Coulomb’s law invariant.\(^6\)

Actually, many authors have tried theories with such “dilaton” fields, and correspondingly “spontaneously broken dilation invariance,” but none is widely accepted yet.

- Anyway, this book is dedicated to exploring the hypothesis that the Maxwell equations and Lorentz force law are already correct and complete as written. We just noted that those equations do not have dilation symmetry. Should we therefore restrict to just those provisional Lorentz boosts with \(\gamma = 1\)?

**Your Turn 30A**

Show that doing two transformations in succession, both with \(\gamma = 1\), does not amount to any single boost with \(\gamma = 1\). Thus, this subset of transformations does not “close into a subgroup.”

But Einstein already knew that there was a different subset, which really do close into a group, and are still sufficient to bake in the Principle of Relativity. We will rediscover them in the next section. Einstein then proved that these transformations were also exact invariances of the full Maxwell equations. We’re not ready to do that,\(^7\) but nevertheless we’ll be able to show that the hypothesis that physics is invariant under them makes experimentally testable predictions, for example, for the aberration of starlight and a variety of Doppler effects. Those predictions agree quantitatively with experiment, which will give us the courage to later push through the proof that they are invariances of the full Maxwell equations.

### 30.3 Lorentz Transformations in One Space Dimension

Again, our task is to find a subset of provisional Lorentz transformations that forms a group, excludes the bogus dilations, but still includes (some kind of) boosts. If we succeed on this math quest, then we can explore the physical hypothesis that the reduced set of transformations are invariances of all of Nature.

---

\(^5\) Muons resemble electrons in some ways, but not in others (muons are unstable), and in any case dilation invariance would require continuously variable properties.

\(^6\) The logic here parallels our rescue of galilean invariance by acknowledging a new dynamical variable in Section 27.4 (page 423).

\(^7\) See Chapters 32–34.
30.3.1 A subgroup that excludes dilations

One way to specify a 1-parameter subset of the provisional 1D Lorentz boosts is to require that \( \gamma \) is not independent of the boost velocity \( \beta \), but instead is a scalar function of it. We wish to do this in such a way that the subset closes into a group. We will guess a trial solution, then confirm it. Then we’ll see a deeper meaning for our solution.

The isotropy of space leads us to expect that the scalar \( \gamma \) won’t depend on which direction \( \beta \) points. We also expect that a boost by \( \beta c \), followed by a boost by \(-\beta c\), should amount to no boost at all (think about jogging backwards at speed \( v_r \) inside a train car that itself is moving at \( v_r \) relative to Earth). Thus, we require

\[
\begin{bmatrix}
1 & \beta \\
\beta & 1
\end{bmatrix}
\begin{bmatrix}
1 & -\beta \\
-\beta & 1
\end{bmatrix}
= 
\begin{bmatrix}
1 & 0 \\
0 & 1
\end{bmatrix}.
\]

This condition fixes \( \gamma = (1 - \beta^2)^{-1/2} \), or

\[
\begin{bmatrix}
ct' \\
x'
\end{bmatrix}
= (1 - \beta^2)^{-1/2}
\begin{bmatrix}
1 & -\beta \\
-\beta & 1
\end{bmatrix}
\begin{bmatrix}
ct \\
x
\end{bmatrix}.
\]

Lorentz boost, 1D \hspace{1cm} (30.2)

In particular, if \( \beta = 0 \) then \( \gamma = 1 \), and so pure dilations are not allowed, as desired. From now on, \( \gamma \) will always mean this particular function of \( \beta \).

To see the significance of the Lorentz boosts, consider what happens when we re-express the wave operator in terms of transformed coordinates (Your Turn 29B, page 443):

\[
\left[-\frac{\partial^2}{\partial (ct)^2} + \frac{\partial^2}{\partial x'^2}\right] u \quad \text{becomes} \quad \gamma^2(1 - \beta^2)^2 \left[-\frac{\partial^2}{\partial (ct')^2} + \frac{\partial^2}{\partial x'^2}\right] u. \quad (30.3)
\]

We see that among the provisional Lorentz boosts, the subset Equation 30.2 are those that leave the wave operator completely form-invariant—not just a multiple of itself. Two such transformations in succession will also have that property, so right away we see that the Lorentz transformations must close into a group.

For very small \( \beta \), the transformations Equation 30.2 reduce to

\[ t' \approx t - (v_r/c^2)x \approx t, \quad x' \approx x - v_r t \quad \text{where} \quad v_r = \beta c. \]

These look just like galilean boosts. That’s why Einstein’s correction to the \( t' \) formula was missed for hundreds of years, during which Newton’s laws made accurate predictions about terrestrial and celestial mechanics.

Later chapters will show that indeed, Lorentz transformations are invariances of the full Maxwell/Lorentz system. That is, the coordinate systems in which electrodynamics takes the simplest form are interrelated by Equation 30.2, which is physically different from the situation in newtonian physics. In honor of Einstein, we’ll call any of the “good” systems an E-inertial coordinate system to distinguish them from the corresponding notion in newtonian physics.\footnote{\textsuperscript{9}Lorentz studied these transformations in 1904, but did not regard them as justification to eliminate the æther.}

\footnote{\textsuperscript{9}The latter were called galilean, or “G-inertial” systems in Section 26.6.1.}
30.3.2 Rapidity parameter

The preceding section argued that the combination of two Lorentz transformations must be another Lorentz transformation. It’s algebraically messy to prove that statement explicitly, but Equation 29.4 (page 444) gives the answer in 1D:

\[
\text{A Lorentz boost with velocity } \vec{v}_s, \text{ followed by a boost in the same direction with } \vec{v}_0, \\
\text{amounts to a single boost in that same direction with speed speed } (\vec{v}_s + \vec{v}_0)/(1 + \left(\frac{\vec{v}_s}{c}\right)^2). \tag{30.4}
\]

There is another formulation of this result that will give other insights. Begin with an analogy to ordinary rotations. Why are rotations given by matrices that, in two dimensions, have the form

\[
\begin{bmatrix}
\cos \alpha & \sin \alpha \\
-\sin \alpha & \cos \alpha
\end{bmatrix}
\]?

One way to characterize such matrices \( S \) is by the statements \( S^T S = 1 \) and \( \det S = 1 \) (special orthogonal matrices). But equivalently, rotations are those linear maps of coordinates that leave the algebraic form of the pythagorean formula invariant: \( x^2 + y^2 = (x')^2 + (y')^2 \). Chapter 14 used this property to show that the Laplace operator is form-invariant under rotations. Rotations close into a group: For example, in 2D \( \alpha_1 \) followed by \( \alpha_2 \) is equivalent to \( \alpha_1 + \alpha_2 \).

The wave equation involves something analogous but a bit different:

Your Turn 30B

Show that, in one spatial dimension, the Lorentz boosts are linear maps that preserve the form of the quantity

\[
c^2 \Delta \tau^2 = (c \Delta t)^2 - (\Delta x)^2,
\]

which we’ll call the invariant interval between two events. [Hint: The proof is very similar to the proof of Equation 30.3.]

Reflections in \( x \) and \( t \) also leave the form of Equation 30.5 invariant.

Because Equation 30.5 looks similar to the rotation case (except for the minus sign), we may hope that the appropriate symmetries will also look similar. Indeed,

\[
\begin{bmatrix}
\cosh \Upsilon & -\sinh \Upsilon \\
-\sinh \Upsilon & \cosh \Upsilon
\end{bmatrix}
\tag{30.6}
\]

does the job, for any \( \Upsilon \). Some books call \( \Upsilon \) the rapidity parameter.

Your Turn 30C

a. Confirm that any transformation of the form Equation 30.6 is a special case of the provisional Lorentz boosts, with \( \gamma = \cosh \Upsilon \) and \( \beta = \tanh \Upsilon \)...

b. …and that moreover, these transformations also satisfy the condition to be in the subgroup of true Lorentz transformations.

c. Show that conversely, any Lorentz boost can be written in the form Equation 30.6.

---

10More precisely, the linear maps that leave the pythagorean formula form-invariant consist of the rotations and reflections in \( x \) and/or \( y \); that is, \( \det S \) may also equal \(-1\).
30.4 A TYPICAL PARADOX AND ITS RESOLUTION

People made many objections to Einstein’s theory, and still do. Out of many we could explore, here is one:

Suppose that a cart moves at uniform velocity \( v \) with respect to the lab. The cart is rigid: Its length is always 0.5 meter, when measured in any inertial coordinate system in which it’s at rest. At some moment, we set two beacons at the center of the cart. One of them is then carried toward the rear of the cart at uniform velocity \(-u\) with respect to the cart, while the other is carried toward the front at uniform velocity \(+u\) with respect to the cart. Each beacon emits a flash of light when it arrives at the end of the cart, and we ask whether those flashes arrive simultaneously at the center.
In newtonian physics, it’s clear that they always do arrive simultaneously and the Principle of Relativity is upheld (Figure 30.1a). “But,” says our skeptic, “that result relies on the simple velocity addition formula in newtonian physics. The crazy velocity addition formula will spoil the simultaneity, allowing us to use this apparatus to detect absolute motion and contradicting the Principle of Relativity.”

To evaluate (and then refute) this claim, Figure 30.1b shows an accurate spacetime diagram with \( v = 0.4c \) and \( u = 0.2c \) (the same values as were used in panel (a)). The dotted red line is the trajectory of the center of the cart. The solid green lines are the trajectories of the two beacons on their ways to the ends of the cart. Their slopes are fixed by the relativistic velocity addition formula, and their end points are the intersections with the cart’s front and back ends.\(^{11}\) Our hypothetical skeptic may have forgotten that although each beacon’s time to flash is the same in the cart’s rest frame (because they travel equal distances at equal speed), still they differ in the lab’s coordinate system.

The wavy lines in the figure are the trajectories of the light flashes. Their slopes are \( \pm 45 \text{ deg} \) in any \( E \)-inertial coordinate system, for example, the lab. Contrary to the claim in quotes above, the flashes coincide at the center of the cart, regardless of the values of \( u \) or \( v \). Therefore we cannot use that observed coincidence to claim that \( v \) has any special value and the Principle of Relativity is upheld in Einstein’s picture after all.

### 30.5 Lorentz Transformations in Three Space Dimensions

We can now see how to introduce the other two space dimensions: Any transformation that looks like Equation 30.2 in a 2×2 block that includes \( ct \) and one spatial direction, and is the identity matrix in the other two directions, will preserve the form of the invariant interval, defined by upgrading Equation 30.5:

\[
c^2 \Delta t^2 = (c \Delta t)^2 - ||\Delta \vec{r}||^2,
\]

and hence of the wave operator \( \nabla^2 - \partial^2/\partial (ct)^2 \), and hence of the wave equation itself.

**Your Turn 30D**

The preceding discussion implies that only one of the following matrices is a Lorentz transformation:

\[
\begin{bmatrix}
\gamma & -\gamma \beta & 0 & 0 \\
-\gamma \beta & \gamma & 0 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0
\end{bmatrix}; \quad \gamma \begin{bmatrix}
1-\beta & 0 & 0 \\
-\beta & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}.
\]

Show that explicitly.

There are three independent directions for Lorentz boosts, just as in the galilean case. Combined with the three kinds of rotations and discrete reflections, they amount to a six-parameter group of transformations called the full Lorentz group. Including the four

---

\(^{11}\)This step in turn came from applying the Relativity Strategy (Equation 26.15, page 414): The ends are located at \( x' = \pm 0.25 \text{ m} \).
space and time translations gives a ten-parameter invariance group sometimes called the Poincaré group.

If two successive boost velocities are not parallel, then their combination is not as simple as Equation 30.4. However, the next section will show that in at least one important case it is still straightforward.

Section 30.5’ (page 468) draws a parallel between the Poincaré group and a construction in geometry.

30.6 MORE KEY EXPERIMENTS: ABERRATION OF STARLIGHT AND DOPPLER SHIFT

30.6.1 Light-speed trajectories change apparent direction but not their speed

We are now ready to discuss the second of the two experimental observations that Einstein said convinced him: the aberration of starlight. Each time we look at the night sky, the stars’ positions relative to each other are always almost the same, but not quite. Even when we correct for refraction in our atmosphere, there are some apparent relative shifts, which are periodic with period one year. More precisely, the stars all crowd very slightly toward the momentary direction of our orbital motion around the Sun. At its maximum, the displacement is just 20 arc sec.

Incredibly, this tiny effect was already observed in the late 1600s by astronomers searching for something completely different (stellar parallax in order to confirm the heliocentric model of the Solar System), culminating with careful measurements by James Bradley around 1726. As outlined in the epigraph to this chapter, Bradley was mystified to find that there were indeed tiny annual variations in the relative positions of stars, but with the wrong annual phase to be explained by parallax (Figure 30.2). Bradley even found an explanation for this phenomenon, based on the hypothesis that light was a stream of

---

Figure 30.2: Aberration versus parallax. Dots: [Observational data.] James Bradley’s historic observations of the apparent position of the star γ Draconis throughout a year. Solid curve: The data fit a function that peaks around 9 September each year. Dotted curve: [Simulated data, a.u.] The hypothesis that this apparent motion is due to parallax predicts instead a maximal deviation around 1 July. The vertical scale depends on distance to the star, which was unknown in the 18th century; we now know that parallax is immeasurably small for this star. This curve just makes the point that parallax could not explain the observed data.

---

12 More precisely, the period is one “sidereal year.” Refraction effects can be minimized by observing stars close to the zenith.
newtonian projectiles. After all, he argued, when raindrops are falling straight down but you run to catch a bus, the raindrop trajectories appear to you to be slanted, approaching you from the forward direction. That model explained why, unlike parallax,

- Aberration is independent of the location of the observer relative to the source, and in particular does not become immeasurably small for very distant sources like stars (unlike parallax).
- Aberration does depend on the speed and direction of the observer’s motion. Specifically, aberration “pulls” stellar images toward the direction of the observer’s motion.

Later wave theories of light, however, were unable to explain the aberration without contorted arguments involving the æther, and even then were unable to explain why the effect was unchanged if a water-filled telescope was used.

Einstein found that his kinematics offered a simple, elegant account of aberration, in either the wave or particle pictures. Again apply the Relativity Strategy (Equation 26.15, page 414). Consider a spacetime curve (particle or wavecrest trajectory) specified in parametric form by

\[
\begin{bmatrix}
  ct \\
  x \\
  y
\end{bmatrix}
= \begin{bmatrix}
  \xi \\
  \eta \\
  0
\end{bmatrix}.
\]

This formula specifies a chain of events depending on a parameter \( \xi \), that is, a curve in spacetime. It could describe the progress of a flash of light (a wave packet, or one crest of a wave train) moving at speed \( c \) along the \( x \) axis.\(^{13}\) Applying a Lorentz boost transformation along \( \hat{y} \) yields the same trajectory as viewed in another E-inertial coordinate system:

\[
\begin{bmatrix}
  ct' \\
  x' \\
  y'
\end{bmatrix}
= \begin{bmatrix}
  \gamma & 0 & -\gamma \beta \\
  0 & 1 & 0 \\
  -\gamma \beta & 0 & \gamma
\end{bmatrix}
\begin{bmatrix}
  \xi \\
  \eta \\
  0
\end{bmatrix}
= \begin{bmatrix}
  \gamma \xi \\
  \xi \\
  -\gamma \beta \xi
\end{bmatrix}. \quad (30.8)
\]

**Your Turn 30E**

a. Show that the new trajectory’s speed is \( \sqrt{(\Delta x')^2 + (\Delta y')^2 / (\Delta t')} = \sqrt{(\Delta \xi)^2 + \gamma^2 \beta^2 (\Delta \hat{z})^2 / (\gamma \Delta \hat{z} / c)} \).

b. Confirm that this expression equals \( c \), as it must.

c. But the new trajectory is not directed along \( \hat{x'} \). Show that instead, it makes an angle \( \delta \) with the \( \hat{z'} \) axis, where \( \tan \delta = \Delta y' / \Delta x' = -\gamma \beta \).

We could do a similar calculation for any initial angle between the trajectory and the boost direction (above you did the case where that angle is 90°).\(^{14}\) The new angle depends both on \( \beta \), and on the original angle, so the relative positions of the stars are different according to the boosted (Earth-bound) observer. The effect is small because

---

\(^{13}\)We’ll suppress the \( z \) coordinate to shorten the formulas. It’s present, but it’s not doing anything interesting.

\(^{14}\)You’ll do this in Problem 30.4.
Chapter 30 Aberration of Starlight and Doppler Effects

Earth's velocity change over the course of a year is much smaller than $c$; nevertheless, the effect was measurable in the 18th century.

**Your Turn 30F**

a. Look up the Earth–Sun distance and use it to estimate the velocity of Earth’s center, assuming that a nearly-circular orbit takes a year to complete.

b. From that, estimate the maximum magnitude of stellar aberration.

c. In addition to periodic shifts over the course of a year, there should be other daily shifts due to an Earth-bound observer being carried by Earth’s rotation. Look up the Earth’s radius and use it to estimate this additional velocity. Is it ever a significant addition over what you found in (a)?

d. Wait: The Solar System is also hurtling around the center of our galaxy. Do we need to include this velocity as well when we make predictions?

Einstein’s proposal for the invariances of electrodynamics has made an absolute prediction for the aberration, with no adjustable fit parameters (no parameter at all other than $c$). It either succeeds or fails—it’s falsifiable. And, as he pointed out in his very first paper, it works, without any special pleading, no extra ad hoc hypotheses about how the æther wind is blowing, and so on.

### 30.6.2 Wave frequency transforms in an angle-dependent way

Now you try the derivation again. But instead of transforming a trajectory, this time apply an active transformation to a plane-wave solution to the wave equation. That is, upgrade Section 29.4.2 to start with $\phi = \cos(-\omega t + \vec{k} \cdot \vec{r})$. Here $||\vec{k}|| = \omega/c$, but $\vec{k}$ can have any direction in the $xy$ plane. Again boost along the $y$ direction.

**Your Turn 30G**

a. Show again that the apparent direction of $\vec{k}$ changes, and find the change in its magnitude (as well as the change in $\omega$).

b. The passive Lorentz invariance of the wave equation guarantees that your new function will be a solution, but check it anyway by confirming the expected relation between your two results in (a).

c. Specialize your result to the case where $\vec{k}$ is parallel to the boost. Interpret your result in terms of a “longitudinal Doppler shift.” [Hint: Recall Section 29.4.2.]

d. Specialize again, this time to $\vec{k}$ perpendicular to the boost (as in Your Turn 30E). Interpret your result as an apparent bending of the direction of $\vec{k}$ as well as a “transverse Doppler shift.”

e. Make some observationally testable predictions about the frequency-dependence of your results.

Note that newtonian physics also predicts a longitudinal Doppler shift, but with a different
Figure 30.3: G. Smoot and coauthors’ historic original data documenting the dipole anisotropy of CMBR. To cancel atmospheric emission, two receivers observed in two directions $\hat{n}_1, \hat{n}_2$ on the sky and the apparent temperatures of their black body spectra were subtracted. Problem 30.6 implies that $\Delta T$ is proportional to $\hat{n} \cdot \hat{n}_1 - \hat{n} \cdot \hat{n}_2$, where the unit vector $\hat{n}$ is a fit parameter (the unknown direction of Earth's velocity relative to the coordinate system in which CMBR is isotropic). The data were fit to find both the constant of proportionality and $\hat{n}$, yielding that Earth's speed with respect to the CMBR is $390 \pm 60$ km/s and the direction of maximal temperature was 10.8 hours right ascension and five degrees declination. [From Smoot et al., 1977.]

magnitude from your prediction in (b) above. And newtonian kinematics predicts zero transverse shift, unlike your answer to (c)—a testable prediction.

Quantitative confirmation that the transverse Doppler effect follows the relativistic formula, and hence excludes galilean invariance, had to wait for the Ives–Stilwell experiment (1938). Much more accurate experiments have been done right into the 21st century.

The Doppler effect also predicts that the apparent temperature of the cosmic microwave background radiation appears slightly higher in one direction of the sky, and slightly cooler in the opposite direction (the dipole anisotropy, Figure 30.3). This effect was observed shortly after the discovery of the cosmic microwave background radiation. The tiny shift must be compensated in observations if we want to see the even smaller, and more cosmologically interesting, anisotropy that arises from early Universe fluctuations.

30.7 AN ENORMOUS GENERALIZATION

30.7.1 Lorentz invariance must apply to all of physics

Again, Einstein’s proposed resolution to the problem of Section 29.2.2 is that

Electrodynamics hardwires the Principle of Relativity by using equations of motion that are invariant under Lorentz—not galilean—transformations.

---

15 Compare Equation 27.4, page 422.
17 Some modern experiments used single atom emitters moving at high speeds; see Further Reading.
18 Your result in Your Turn 30G includes the full angular dependence of this shift.
19 You’ll explore the CMBR dipole anisotropy in Problem 30.6.
20 Compare our galilean statement (Equation 26.13, page 413).
Einstein took an extraordinary additional step.\textsuperscript{21} Up till now, Lorentz invariance may have seemed to be a peculiarity of electrodynamics, which we could safely ignore if, say, we were only interested in the motions of planets. But suppose that the Maxwell equations and newtonian mechanics were both correct as written. That is, suppose that there is even one coordinate system in which both of those systems’ equations of motion correctly describe physics. Applying a galilean boost to that system would then spoil the form of Maxwell. Applying a Lorentz boost to it would spoil the form of Newton. In fact, there would be no other coordinate system in uniform, straight-line motion relative to the original one in which all equations of motion have the same form. So in such a world we could define “absolute rest” as that original coordinate system—contradicting the Principle of Relativity:

\textit{If we want to hardwire in the Principle of Relativity via an invariance, then that invariance must apply to all of physics—even to phenomena not yet discovered.}

That’s quite a leap. We can’t have it both ways. Einstein’s hypothesis was that

\textit{Although newtonian physics had looked good for hundreds of years, actually it hadn’t been tested for objects moving at speeds near $c$. So it’s Newton that must be changed, not Maxwell.}

Or, paralleling Idea 26.12 (page 412):

\textit{Physics has an overarching mathematical property that transcends details of particular springs, clocks, planets, and so on. That property is that the specific equations for any situation always have a family of preferred coordinate systems, which are related to each other by Poincaré group transformations.}

(Recall that the Poincaré group contains Lorentz transformations along with translations.)

### 30.7.2 Muon lifetime, galactic redshifts, CMBR dipole, and more

The hypothesis of universal Lorentz invariance now gives us many nontrivial physical predictions (Figure 30.4), all of which start by saying “Suppose that the dynamical laws governing [some process] are invariant under Lorentz transformations…” From there, we can apply the Relativity Strategy (Idea 26.15, page 414). For example, we’ve seen how to understand Fizeau’s experiment, the aberration of starlight, and various Doppler shifts, by using that approach.\textsuperscript{22}

Note that when we hypothesize that “all laws of physics are invariant under Lorentz transformations,” we mean all, including quantum physics. Here are more examples:

---

\textsuperscript{21}Extraordinary yet strangely familiar: Section 26.6.5 (page 412) used the same logic to say that if we want to implement the Principle of Relativity with galilean transforms, then they must be invariances of all of physics.

\textsuperscript{22}For Fizeau: “Suppose that, whatever interactions slow light down in water, they are invariant under Lorentz transformations…” For aberration and Doppler: “Suppose that, whatever dynamics are responsible for the propagation of light in vacuum, they can be expressed in terms of equations invariant under Lorentz transformations…”
Suppose that, whatever process makes the muon disintegrate, that process is invariant under Lorentz transformations. We capture some muons, bringing them to rest with respect to our lab, and find that their mean lifetime is 2.2 μs (Figure 30.5). Then
Figure 30.6: [Experimental data.] **Doppler shift for gamma rays.** The radioactive nuclide $^{60}\text{Co}$ transitions to an excited state of $^{57}\text{Fe}$, which subsequently emits a photon with energy 14.4 keV. A second sample of $^{57}\text{Fe}$ can absorb those photons, or not, depending on whether they are received at the same energy with which they were emitted. Even the fraction of photons that were emitted without recoil energy loss will pass through the absorber, if they have been Doppler shifted off resonance. The absorption line is so narrow in frequency that even a very small relative velocity suffices to create this effect. The curve shows a standard line shape (lorentzian) fit to the data. [Data from Ruby & Bolef, 1960.]

we can predict that a muon moving rapidly relative to the lab’s E-inertial coordinate system will also live 2.2 $\mu$s in any E-inertial coordinate system in which the muon is at rest. Transforming this duration into the laboratory coordinate system via Equation 30.2 (page 454) shows that a fast-moving muon appears, in the lab, to live longer before disintegrating than does a muon at rest, as is observed. Specifically, we predict a lab lifetime $\gamma(2.2 \mu s)$, during which the muon travels lab distance $\gamma\beta c(2.2 \mu s)$, farther than it would have gone under the hypothesis of galilean invariance.

Because muons are created in Earth’s upper atmosphere, few would survive the trip down to a surface-based lab were it not for the time dilation effect. Conversely, measurements of muon flux at various altitudes can be used to quantitatively confirm the prediction made by relativity.

- Suppose that, whatever process is responsible for an excited nucleus of iron to give off a gamma photon by recoilless emission, that process is Lorentz invariant. Then a second iron nucleus that could resonantly absorb such a photon will not do so if it’s in motion relative to the first one, because in its rest frame the photon is Doppler shifted, and hence off resonance, a testable prediction (Figure 30.6).

- The Doppler shift formula also lets us deduce the motion of distant galaxies relative to us: We suppose that, whatever atomic physics is responsible for making hot gas give off light with a pattern of spectral lines, that process is invariant under Lorentz transformations. Then a hydrogen atom moving rapidly relative to us will have the same spectral lines as one in our lab, if it’s measured in the E-inertial coordinate system in which that atom is at rest. Transforming that outgoing wave to our lab’s E-inertial coordinate system gives its apparent frequency when we observe it with a spectrometer, typically shifted to the red because most galaxies are moving away

---

23The Mössbauer effect.” You’ll examine this phenomenon in Problem 31.2.

24In 1868, W. and M. Huggins detected a Doppler shift in the spectrum of Sirius, the birth of this indispensable astronomical method.
from ours.

- Strong and weak nuclear forces, which are not electrodynamic in origin, lead to particle reactions that conserve energy and momentum. But Chapter 31 will show that, in order for energy and momentum to be conserved in every E-inertial coordinate system, we must modify the newtonian definitions of energy and momentum, in ways that have experimentally testable consequences in nuclear and high-energy physics.

The power of relativity lies in the fact that these apparently unrelated phenomena, and many others, are all quantitatively explained with one idea, (30.9). The existence of laws of this sweeping generality is the basic epistemological miracle of physics. It’s what gives physical law a different character from the rules governing other branches of science.

The revolutionary aspect of Einstein’s logic was not just the factual content of his proposal, but also the method: Until then, the general approach had been to propose individual laws of Nature, then test them. Instead Einstein went straight to the next higher level, writing a transformation principle that’s proposed to be an invariance of all laws of Nature, whatever they may turn out to be.

Section 30.7.2' (page 468) discusses the muon lifetime experiment in more detail.

### 30.8 WHAT’S NEXT

1. We now have a proposal for a set of transformations that:
   - Are invariances of the wave equation; and
   - Form a group.

But the wave equation we have studied assumed a scalar field, whereas we know that the electric and magnetic fields are not scalars. Not only do the components of $\vec{E}$ transform among themselves under rotation; we’ll see that $\vec{E}$ mixes with $\vec{B}$ under a boost. So we need to augment our Lorentz transformations on spacetime by making a proposal for what happens to the components of $\vec{E}$ and $\vec{B}$ under them. Only then will we have a firm proposal for what transformations are supposed to leave the Maxwell equations invariant. Then we can do the math to see if it’s true, based on the close relation of Equation 30.6 to rotations.

2. Our logic may still feel a bit ad hoc, but here we were still feeling our way, trying to guess the right hypothesis. Now that we’ve got it, and it looks promising, we are in a position to develop a more streamlined formulation (“high-tech relativity,” Chapters 32–33).

3. First, however, Chapter 31 will explore other kinematic consequences of Lorentz invariance, and their experimental signatures.

---

25Recall Hanging Question #A, page 13
FURTHER READING

*Semipopular:*

*Intermediate:*
Tallents, 2023.

*Technical:*
Observation: Figure 30.3.
30.3’a Light-cone coordinates

Here’s a more elegant derivation of Lorentz transformations than the one in the main text.

Suppress $y, z$ for the moment, and consider only $ct, x$. It is helpful to define light-cone coordinates

$$
\begin{bmatrix}
  u \\
  c \\
\end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix}
  1 & 1 \\
  -1 & 1 \\
\end{bmatrix} \begin{bmatrix}
  ct \\
  x \\
\end{bmatrix},
\begin{bmatrix}
  c \\
  x' \\
\end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix}
  1 & -1 \\
  1 & 1 \\
\end{bmatrix} \begin{bmatrix}
  u \\
  c \\
\end{bmatrix}.
$$

Then the general solution to the wave equation takes the simple form $f(u) + g(c)$ for any two functions $f, g$. The function $f$ describes a waveform moving to the left; $g$ is a waveform moving to the right.

The linear transformations $u' = Au, v' = Bv$ change a solution to $f'(u') + g'(v')$ where $f'(u') = f(Au)$ and so on, which has the same functional form as before. So any such transformation is an invariance of the solution space of the wave equation; that is, waves traveling left or right at velocity $\pm c$ in the original coordinates are again traveling left or right at velocity $\pm c$ in the new system.

In light-cone coordinates, the operator appearing in the wave equation (the wave operator, or dalembertian) has the simple form $\delta^2/\delta u \delta v$. In terms of the transformed coordinates, this is $(AB)(\delta^2/\delta u' \delta v')$. Dividing both sides of the transformed wave equation by $AB$ then shows that such transformations are invariances of the wave equation. They include dilations with $(A = B = 1)$; those are invariances of the vacuum wave equation, although not of the rest of physics. We can eliminate them, and get the expected 1-parameter family of boosts, if we restrict to the case where $A = B^{-1}$. That family of transformations are precisely the Lorentz boosts.

$$
\begin{bmatrix}
  c' \\
  x' \\
\end{bmatrix} = \frac{1}{2} \begin{bmatrix}
  A + A^{-1} & A - A^{-1} \\
  A - A^{-1} & A + A^{-1} \\
\end{bmatrix} \begin{bmatrix}
  c \\
  x \\
\end{bmatrix}.
$$

This can be placed in its more famous form by letting $\gamma = (A + A^{-1})/2$ and $\beta = (A^{-1} - A)/(A^{-1} + A)$, yielding Equations 30.2 or 30.6.

30.3’b Reformulation of the invariant interval

Light-cone coordinates also make it easy to see that the quantity $-2(\Delta u)(\Delta v) = (c^2 \Delta t)^2 - (\Delta x)^2$ is a scalar under Lorentz transformations (it acquires a factor of $A/A = 1$). If two events can be joined by a trajectory moving at $\pm c$, the interval equals zero because either $\Delta u = 0$ or $\Delta v = 0$; if they can be joined by a trajectory moving slower than $c$, then the interval is real and positive.

30.3’c Velocity addition in light-cone coordinates

It’s also easy to find the combined effect of two Lorentz boosts by using light-cone coordinates. Please convince yourself that the combined operation is itself a Lorentz boost with $A_{tot} = A_1 A_2$. To interpret this result, invert the relations between $A$ and $(\beta, \gamma)$ to find

$$
A = \gamma(1 + \beta) \quad \text{or} \quad A^{-1} = \gamma(1 - \beta).
$$

Thus, $A_{tot} = \gamma_1(1 + \beta_1)\gamma_2(1 + \beta_2)$ gives

$$
\beta_{tot} = \frac{(1 + \beta_1)(1 + \beta_2) - (1 - \beta_1)(1 - \beta_2)}{(1 + \beta_1)(1 + \beta_2) + (1 - \beta_1)(1 - \beta_2)} = \frac{\beta_1 + \beta_2}{1 + \beta_1 \beta_2},
$$

and we recover Equation 29.4 (page 444).
Chapter 30  Aberration of Starlight and Doppler Effects

30.3’d Relation to rapidity

Equation 30.11 is the same as Equation 30.6 (page 455) with \( T = \ln A \). This is helpful, because in light-cone coordinates the composition law is simply \( A_{\text{int}} = A_1 A_2 \) (show that). So \( T_{\text{int}} = \ln(A_1 A_2) = T_1 + T_2 \), which agrees with your result in Your Turn 30B (page 455).

30.5’ The euclidean group

The Poincaré group is analogous to a construction from ordinary geometry: If we take the three-parameter group of spatial rotations and add the three rigid spatial translations, the resulting six-parameter group is called the “euclidean group.” Congruent shapes are those related by euclidean transformations.

30.6.2’ Another view of the longitudinal Doppler shift

There is a more geometrical (less algebraic) way to think about the longitudinal Doppler shift:

The diagram above shows the loci of a chain of wavefronts, each moving along \( \hat{x} \) at speed \( +c \) and separated in time \( t \) by period \( T \). The dashed lines are coordinate axes for an E-inertial coordinate system moving with respect to the unprimed system. The period \( T' \) of the same wave observed in this system depends on the intersection of the \( t' \) axis with a wavefront, as shown.

Your Turn 30H

Work out the relation between \( T' \) and \( T \), and again recover the longitudinal Doppler formula.

30.7.2’ More about muon lifetime

The muon had not yet been discovered in 1905, so the result in Figure 30.5 (page 463) was not available to Einstein. Today we call the physics responsible for its disintegration “the weak interaction,” part of the more general “electroweak theory.”

The figure shows that the muon lifetime is actually a random variable: It has an exponential distribution with expectation 2.2 \( \mu s \). It is this expectation that gets transformed when the muon is moving relative to the lab. This sounds like an annoying extra complication, but actually, it explains how we are able to measure muon lifetime even though we don’t know the exact creation times of individual cosmic-ray muons. We measure the probability per unit time of disintegration for a sample of muons in flight, and compare it to the corresponding quantity for a sample of muons that have been captured, and hence slowed down, by atomic nuclei.
30.1 *Rapidity*
Continue Your Turn 30C (page 455):

a. Section 30.3.1 argued that because the transformations Equation 30.6 can be characterized as those that leave something invariant (in this case, the wave operator, Equation 30.3), they must close into a group. Now confirm this expectation directly: Use a hyperbolic trig identity and Equation 30.6 to show that a boost with $\gamma_1$, followed by one with $\gamma_2$, is equivalent to a single boost with $\gamma_{\text{tot}} = \gamma_1 + \gamma_2$.

b. Confirm that this combination rule amounts to the same thing as a boost by the velocity $v'$ obtained from the formula we found earlier, Equation 29.4 (page 444).

30.2 *Cart before the horse*
Figure 30.1 (page 456) showed a particular case of the thought experiment described in Section 30.4. Maybe the result shown was accidental. Make a similar figure showing the case in which the cart’s velocity relative to the lab is $v = 0.2c$ and the clocks move apart from its center at speeds $v = \pm 0.4c$ in the cart’s rest frame. For concreteness, suppose that in its rest frame, the cart’s total length is $L$ and the clocks reach the ends of the cart simultaneously. At that moment, each emits a flash of light. [*Hint:* Use a computer to make an accurate figure. Make sure to use equal scaling for the $x$ and $ct$ axes.]

30.3 *Length contraction*
Relativistic length contraction is harder to observe directly than is time dilation. Here is an indirect approach.

A long, straight, thin wire lies along the $x$ axis. The wire is electrically neutral but carries current $I$. We idealize this situation by supposing that the wire consists of charges $+\Delta q$ that are at rest in the lab coordinate system (the “nuclei and immobile electrons”), as well as charges $-\Delta q$ that are moving at speed $-v\hat{x}$ with respect to the lab (the “mobile electrons”). Each species has the same spacing $\Delta x$ in the lab coordinate system, because the wire is neutral. The quantities $I$, $\Delta q$ and $v$ are all positive. Thus, there is current in the $+\hat{x}$ direction. We are imagining a continuum limit where $\Delta q \to 0$ holding fixed the linear charge density $\Delta q/\Delta x$.

a. Write an expression for $v$ in terms of $I$, $\Delta q$, and $\Delta x$.

A test charge $q$ moves alongside the wire; its speed relative to the “nuclei” is also $-v\hat{x}$ (that is, parallel to the wire’s axis in the opposite direction to the flow of current). The test charge stays a fixed distance $r$ from the axis of the wire.

b. The wire is net neutral, so it creates no electric field. You know how to compute the magnetic field from the current, and the resulting force on the test charge, so write an expression for that force. Which way does it point?

c. Now think about how the system looks in a Lorentz-boosted coordinate system moving at $-v\hat{x}$ relative to the lab system. In the lab coordinates, the trajectories of the “nuclei” are the lines $(t, n\Delta x)$ for various constant integer values of $n$. Transform those
trajectories to the moving coordinates and for fixed $t' = 0$ find the spacing $\Delta x'$ of these charges in the boosted coordinate system.

d. In the lab system, the trajectories of the “mobile electrons” are the lines $(t, n\Delta x - vt)$ for various constant integer values of $n$. Transform those trajectories to the boosted coordinates and for fixed $t' = 0$ find the spacing of these charges.

e. What, then, is the net linear charge density of the wire in the boosted coordinates? Assume that electric charge itself is a scalar under Lorentz transformation (the charge of an object is the same in any E-inertial coordinate system).

f. What electric field do you expect from the charge arrangement in (e)?

g. In the boosted coordinate system, the test charge is at rest, so the magnetic field if any is irrelevant. Nevertheless, there is a force. What is the origin of this force? How is it related to the one in (b)?

[About science: Implicitly this problem asks you to assume that electrodynamics is fully Lorentz-invariant, which is something we haven’t proved yet. If you get a prediction using some unproved step and it seems reasonable, then that can give you the confidence needed to justify the hard work of trying to show the full result later (Chapter 34).]

30.4 Time course of aberration

In this problem, you’ll explain the data in Figure 30.2.

a. Consider a trajectory that, when viewed in an inertial coordinate system in which the Sun is at rest, travels at speed $c$ in some direction $\hat{m}$. We view this trajectory from a new inertial system moving at speed $c\hat{\beta}$ relative to the old one. Generalize Equation 30.8 (page 459) to this case and show that, to first order in $\beta$, the trajectory’s direction in the new system is

$$\hat{m}' = \hat{m} - \hat{\beta} + \hat{m}(\hat{m} \cdot \hat{\beta}).$$

(30.12)

[Hint: Start with arbitrary $\hat{m}$ but $\hat{\beta}$ still directed along $\hat{y}$. Then find a rotation-invariant generalization of your result.] Discuss the cases where $\hat{\beta}$ is parallel, antiparallel, and perpendicular to $\hat{m}$.

b. Light from a star travels to Earth along a trajectory with $\hat{m}$ equal to minus the unit vector from Earth to the star. So the apparent direction to the star is $-\hat{m}'$. Show that your result Equation 30.12 implies that

The apparent location of the star in the sky is slightly more aligned with $\hat{\beta}$ than the true direction.

c. Figure 30.7a shows a general situation. Before tackling it, consider the special case of a star located perpendicular to the plane of Earth’s nearly circular orbit ($\theta_0 = \pi/2$), and describe qualitatively how its apparent position moves in the sky with the seasons.

d. Next, consider the special case of a star located parallel to the plane of Earth’s nearly circular orbit ($\theta_0 = 0$), and describe qualitatively how its apparent position moves with the seasons.

e. Bradley observed the star $\gamma$ Draconis, whose position in the sky is tilted by $\theta_0 \approx 75 \text{ deg}$
Problems 471

Figure 30.7: Stellar aberration geometry. (a) A star at altitude $\theta_0$ above the plane of Earth’s nearly circular orbit. (b) Apparent positions of the star as seen from the moving Earth at various times of the year. The vector $-\hat{m}$ points from Earth toward various points on the dotted curve. The drawings are not to scale; actually, the star is so distant that we may neglect the Earth’s displacement from the Sun. Also, the actual angular deviations are much smaller than those shown in (b).

up from the plane of Earth’s orbit.26 Explain qualitatively which of the apparent sky positions $P$–$S$ correspond to the orbital positions 1–4.

f. Bradley could only measure one component of angular displacement, corresponding to the vertical wobble in panel (b) of Figure 30.7b. The overall displacement thus corresponded to the angular distance between $P$ and $R$ in the figure. Estimate this quantitatively, then derive how this component of the displacement should depend on time over the course of a year.

30.5  $T_2$ Optical Foucault pendulum

A lab that is anchored to Earth’s surface sets up a non-inertial coordinate system, due to Earth’s rotation. We can detect this small acceleration without looking at the stars, for example, by setting up a Foucault pendulum. In this problem you will explore an optical analog, which is the basis of an important technology.

Imagine a flat table with mirrors, such that light will traverse a roughly square path in vacuum and return to its starting point. More precisely, the light path is a trapezoid: One edge is oriented North–South and has length $L$ in its rest frame. The next edge (called $b$) is oriented East–West and has length $L$ in its rest frame. The third edge is oriented North–South and has length $L$ in its rest frame. The last edge (called $a$) is oriented East–West and has length slightly longer than $L$ in its rest frame, because lines of latitude on Earth are not of equal length.

You will be working out the round-trip transit time for light in the rotating apparatus, and specifically the difference in transit time depending on whether the light goes round

---

26Because this star has right ascension of about 18 hours, we may simply add its declination to the tilt of Earth’s axis from the ecliptic plane, obtaining the stated angle.
clockwise or anticlockwise (when viewed on a line directed toward the center of Earth).\textsuperscript{27} The apparatus is much smaller than Earth: $L = 1 \text{ m}$. It is located at north latitude $\alpha$, that is, the polar angle measured from the North Pole is $\theta = \pi/2 - \alpha$.

a. You know the angular frequency $\omega$ of Earth’s rotation (and which way it is rotating). From that you can make a dimensionless parameter $\epsilon = \omega R_{\text{earth}}/c$. Evaluate this numerically.

There would be no difference in transit times if Earth were not rotating. But perhaps there will be an effect at order $\epsilon$. So work the following steps keeping only first-order contributions. (If the answer is zero, you can go back and look at higher-order terms.) Use the fact that $L \ll R_{\text{earth}}$.

Let unprimed variables $ct$ and $\vec{r}$ refer to an inertial (hence nonrotating) coordinate system in which the center of Earth is at rest. The key facts about rotation are that (i) edges $a$ and $b$ move at different speeds relative to the unprimed system, because they are located at slightly different polar angles $\theta_a > \theta_b$, and that (ii) each is directed nearly parallel to its velocity. (Actual lines of latitude and longitude are curved, and so do not coincide perfectly with the straight edges of the apparatus, but this difference is unimportant in the problem.)

You can forget about the other two edges, which are oriented perpendicular to their velocities.

You know the length of each edge in its own rest frame. Begin by studying a light beam that proceeds in a clockwise direction. Thus, it starts at the southeast corner, traverses $a$ heading West, reflects off a mirror, proceeds North, and reflects again. Then it traverses $b$ heading East, reflects one more time, and proceeds South to its starting point.

b. Find the transit times in the unprimed coordinate system for edges $a$ and $b$ and add them. [Hint: The Relativity Strategy may be helpful (Idea 26.15).]

c. Repeat for a light beam circulating anticlockwise.

d. Subtract the two preceding results and express your answer in terms of $\theta, L, \omega, R_{\text{earth}}$, and constants of Nature. Although you have computed time in the unprimed system, explain why the round-trip transit time difference will have the same value according to a clock fixed to the instrument.

e. Evaluate your answer for an apparatus located at north latitude $\alpha = \pi/4$. Which transit time is faster: the clockwise or the anticlockwise route?

f. Compare your answer to the period of visible light. Is this a measurable effect?

\section*{CMBR anisotropy}

a. Generalize Your Turn 30G to three spatial dimensions. That is, start with a plane wave with angular frequency $\omega$ and wavevector $\vec{k}$, re-express it in a Lorentz-boosted coordinate system, show that it remains a plane wave, and identify the new frequency and wavevector as seen in the new system.

Let’s model the cosmic microwave background radiation as a classical EM field consisting of a superposition of many plane waves with various different wavevectors. We assume

\textsuperscript{27}You are following in the footsteps of Max von Laue (1920).
that there’s an E-inertial coordinate system \((ct, \vec{r})\) in which the CMBR is isotropic. That is, when viewed in this system the waves have random phases and polarizations, and wavevectors drawn from the isotropic probability distribution\(^{28}\)

\[
\mathcal{Q}(k) d^3k = C f_0(||k||/\tau) d^3k.
\]  
(30.13)

In this formula, \(\mathcal{Q}\) is a probability density function. Recall what that means: In a little box of \(d^3k\) space with volume \(d^3k\), we have \(M \mathcal{Q}(k) d^3k\) component plane waves, where \(M\) is some big constant. The constant \(\tau\) is related to the temperature of the radiation (that is, \((2.7 K)k_B/(hc)\)), \(C\) is a normalization constant, and \(f_0(x) = (e^x - 1)^{-1}\) is the Planck function. But we won’t need any quantum mechanics for this problem.

We want to know what this EM field looks like in our terrestrial coordinate system \((ct', \vec{r}')\), which is E-inertial but moving at speed \(v_\mu\) along the \(-\hat{z}\) direction relative to the original coordinate system. Certainly it will still be a superposition of plane waves, each with wavevector \(\vec{k}'\) related to the original system as you found in (a). We are interested in the density of those \(\vec{k}'\)’s in wavevector space.

b. Find the probability density \(\mathcal{Q}'(\vec{k}')\) of \(\vec{k}'\) vectors. (You can forget about polarization.)

Show that the distribution in the primed system, restricted to any particular direction \(\vec{k}\), again has the Planck form, but with a direction-dependent effective temperature \(\tau_{\text{eff}}(\vec{k})\), which you are to find.

\([\text{Hint:} \ You \ will \ need \ to \ find \ the \ volume \ in \ \vec{k}' \ space \ corresponding \ to \ a \ small \ volume \ in \ \vec{k} \ space \ and \ divide \ those \ two \ volumes. \ By \ axial \ symmetry, \ that \ answer \ will \ depend \ only \ on \ the \ angle \ between \ \vec{k}' \ and \ \hat{z}.]\)

30.7  \ Disco discovery

In this problem, use classical (not quantum) physics.

An electromagnetic plane wave has wavelength \(\lambda\) and moves along the positive \(\hat{z}\) direction when measured in one E-inertial coordinate system (the “lab system”). The wave is reflected by a spherical mirror, which is moving relative to the lab system, also in the \(+\hat{z}\) direction, but with velocity \(v\).

a. Some of the light will be reflected directly backward, along the \(-\hat{z}\) direction. Find its wavelength \(\lambda'\) as measured in the lab system, in terms of \(\lambda\), \(v\), and physical constants.

\([\text{Hint:} \ Apply \ the \ Relativity \ Strategy \ (Equation \ 26.15, \ page \ 414). \ There’s \ another \ E-inertial \ coordinate \ system \ (not \ the \ lab \ system) \ in \ which \ you \ certainly \ know \ the \ relation \ between \ incident \ and \ reflected \ frequencies. \ Convert \ that \ knowledge \ into \ a \ statement \ relating \ the \ wavelengths \ as \ seen \ in \ the \ lab.]\)

b. Generalize your answer to the case where the scattered light is observed in an arbitrary direction, making an angle \(\theta\) with the \(\hat{z}\) axis in the lab system. \([\text{Hint:} \ Same \ hint \ as \ in \ (a). \ You \ may \ find \ it \ easier \ to \ express \ your \ answer \ in \ terms \ of \ the \ scattering \ angle \ as \ seen \ in \ the \ other \ coordinate \ system, \ then \ express \ that \ angle \ in \ terms \ of \ \theta.]\)

\([\text{Notes:} \ (i) \ If \ you \ know \ the \ Compton \ formula, \ and \ it \ disagrees \ with \ your \ answer, \ don’t \ worry. \ Historically \ this \ disagreement \ led \ to \ the \ acceptance \ of \ Einstein’s \ light-quantum\)
theory—a modification to classical electrodynamics. In the classical domain (for example, coherent states of many photons bouncing radio off a satellite) your result is accurate. (ii) Bouncing a radar beam off a speeding car and measuring the beat frequency between outgoing and returning signals is another real-world application.]
Oh, that Einstein, always cutting lectures—I really would not have believed him capable of it.
— Einstein’s former teacher Minkowski, upon reading the relativity paper.

31.1 FRAMING: INSEPARABLE ASPECTS

Another outstanding kinematic consequence of the hypothesis that all of physics, not just electrodynamics, is Lorentz invariant concerns energy and momentum. Although the experiments confirming it involved nuclear physics, Einstein’s initial discovery was based on Electromagnetic Phenomena.

**Phenomenon:** The mass of a deuterium nucleus is less than the sum of the proton and neutron masses.

**Physical idea:** Relativistic energy–momentum is conserved, but there are no separate conservation laws for mass and energy.

31.2 CONSERVATION OF NEWTONIAN ENERGY AND MOMENTUM IS NOT COMPATIBLE WITH LORENTZ INVARiance

Section 30.7 mentioned that Lorentz invariance is all-or-nothing: We can’t have some of physics invariant under Lorentz transformations while some other part is invariant under galilean transformations. Accordingly, let’s think beyond the Maxwell equations, to consider any sort of interaction that could be called a “collision” among “particles.” For our purposes, a “particle” is a region of space containing something that is initially isolated from the rest of the world (no relevant interactions). We imagine several of these, initially all mutually noninteracting, which come together and interact during a finite time interval (a “collision”). Afterward, some other “particles” emerge that are again noninteracting. In some contexts it may even be appropriate to treat an entire galaxy as a “particle,” or a planet, . . . , on down to atomic nuclei and beyond.

Suppose that there are two incoming particles with masses $m_{1,2}$ and velocities $\vec{v}_{(1,2)}$, and two outgoing particles with $m_{3,4}$ and $\vec{v}_{(3,4)}$. In first-year physics, we start with Newton’s laws and prove that

$$\vec{p}^N_{(1)} + \vec{p}^N_{(2)} = \vec{p}^N_{(3)} + \vec{p}^N_{(4)}, \quad \text{where } \vec{p}^N_{(\ell)} = m_{\ell}\vec{v}_{(\ell)} \text{ (newtonian).} \quad (31.1)$$
The quantity \( \vec{p}^N_{(\ell)} \) is called the **newtonian momentum** of particle \( \ell \).

But even if we didn’t yet know Newton’s laws, and had merely guessed the conservation law Equation 31.1, we could nevertheless state confidently that it is consistent with the rotational invariance of the world. That’s because under rotations the components of velocity (and hence those of \( \vec{p} \)) transform in a simple way, as a 3-vector. Moreover, mass is a scalar (a rotation-invariant quantity), so the \( m_\ell \vec{v}_{(\ell)} \) are also 3-vectors:

\[
[\vec{p}^N_{(\ell)}] = S[\vec{p}^N_{(\ell)}].
\]  
(31.2)

When we express each term of Equation 31.1 in terms of a rotated coordinate system, then the matrix \( S^{-1} \) is a common factor:

\[
S^{-1}[(\vec{p}^N_{(1)} + \vec{p}^N_{(2)} - \vec{p}^N_{(3)} - \vec{p}^N_{(4)})] = 0.
\]  
(31.3)

Multiplying both sides of this equation by \( S \) gives an equation of the same form as Equation 31.1, so the newtonian conservation law is invariant under rotations.

**Your Turn 31A**

In newtonian physics, mass can be exchanged among the participants in a collision, but *total* mass is conserved:

\[
m_1 + m_2 = m_3 + m_4 \quad \text{(newtonian).}
\]  
(31.4)

From this, show directly (without appeal to Newton’s laws) that Equation 31.1 is also invariant under galilean boosts.

In short,

*Even if we didn’t know Newton’s laws, or the details of what’s inside our “particles,” we could nevertheless say that Equation 31.1 is at least compatible with the overarching principle of invariance under the galilean group.*

However, we cannot adapt the simple argument in Equation 31.3 to show that Equation 31.1 is consistent with Lorentz invariance, because \( \vec{v}^N \) is a complicated, nonlinear function of \( \vec{v} \) (Equation 29.4, page 444). Indeed, given a set of four momenta \( \vec{p}^N_{(\ell)} \) whose components obey Equation 31.1, then their values in another \( E \)-inertial coordinate system will *not* in general obey it. So Equation 31.1 cannot be a valid law of Nature in the Lorentz-invariant world that we are proposing. Nor can Newton’s laws be valid, because Equation 31.1 is a consequence of them.

There is another famous conservation law in first-year physics:

\[
\mathcal{E}^N_{(1)} + \mathcal{E}^N_{(2)} = \mathcal{E}^N_{(3)} + \mathcal{E}^N_{(4)}, \quad \text{where } \mathcal{E}^N_{(\ell)} = \frac{1}{2} m_\ell ||\vec{v}_{(\ell)}||^2. \quad \text{(newtonian)}
\]  
(31.5)

---

1Chapter 14 introduced the rotation properties of 3-vectors (Equation 14.1, page 210).

2Newton himself didn’t use conservation of energy. Although G. Leibnitz noted a form of conservation as an algebraic property of Newton’s laws in a special case, Émilie du Châtelet seems to have been responsible for conceptualization of mechanical energy as a distinct concept, and she disseminated that view in her translation and commentaries on Newton.
This formula is rotation invariant by an even easier argument than before: Each term is separately invariant.

**Ex.** Check that Equation 31.5 is galilean invariant.

**Solution:** Sum over all (two) of the incoming particles:

\[
\mathcal{E}_{\text{tot}}^N = \sum_{\ell} \mathcal{E}_{\ell}^N = \sum_{\ell} \frac{1}{2} m_\ell \| \vec{v}_e - \vec{v}_s \|^2 = \sum_{\ell} (\mathcal{E}_{\ell}^N - m_\ell \vec{v}_e \cdot \vec{u}_s + \frac{1}{2} m_\ell \| \vec{v}_s \|^2)
\]

\[
= (\sum_{\ell} \mathcal{E}_{\ell}^N) + \vec{u}_s \cdot (\sum_{\ell} \vec{p}_{\ell}^N) + \| \vec{v}_s \|^2 \sum_{\ell} m_\ell.
\]

The first term on the right is conserved by Equation 31.5; the second is conserved by Equation 31.1; the third is conserved by Equation 31.4; so \(\mathcal{E}_{\text{tot}}^N\) is conserved.

However, Equation 31.5 also turns out not to be Lorentz invariant. Therefore it, too, cannot be a valid law of Nature in any Lorentz-invariant world.

So are energy and momentum not conserved?

### 31.3 CONSERVATION LAWS RECOVERED

#### 31.3.1 “Einstein thinking” places symmetry first

The fact that the newtonian formulas are not acceptable does not rule out the possibility that we may find some other quantities and propose conservation laws for them that would be Lorentz-invariant.

But where should we look for such formulas? Einstein’s approach was so radically different from his contemporaries’ that it really deserves to be called **Einstein thinking**.³

Faced with this sort of question, the obvious approach seems to be to guess or deduce the right equations of motion, then prove a theorem about a mathematical property they possess.⁴ By 1905, this approach had led to a lot of unreadable papers, and moreover, scientists didn’t even realize how hopeless it was, because many phenomena now described by particle physics hadn’t even been discovered.

We will stand the approach just described on its head:

- Start with a proposal for a symmetry of physics, in this case Lorentz.
- Discard hypotheses incompatible with the proposed symmetries, in this case conservation of newtonian momentum and energy.
- Find replacement hypotheses that are compatible, without attempting yet to deduce them from any equations of motion.
- Seek experimentally falsifiable consequences of the proposal.
- If the proposal survives enough nontrivial challenges, use it as a guide to find the right equations of motion.

³In fact, Einstein originally called his ideas the “theory of invariants.” The phrase “theory of relativity” was coined by somebody else, and Einstein only adopted it reluctantly some time later.

⁴For example, we might guess the correct lagrangian function, then apply Noether’s theorem to it (Chapter 40).
To get started on this program, recall again the root of our problem: Velocity is $d\vec{r}/dt$, and both the numerator and denominator of this expression transform under Lorentz boosts (unlike the case with galilean boosts). If only we could replace the denominator by something that didn’t transform, then we’d be in a simple situation like that for rotations.

Here is a view that, while not Einstein’s historical route, follows the sort of logic that he eventually applied to many problems. First, note that the invariant interval between two events in spacetime (Equation 30.7, page 457) is unchanged under Lorentz transformations: $\Delta \tau = \Delta \tau'$. Thinking of a particle’s trajectory as a chain of events in spacetime, $\Delta \tau^2$ between any two of its constituent events is always nonnegative, because particle trajectories cannot move faster than speed $c$. In fact, we’ll see that an ordinary material particle cannot even reach speed $c$, so $d\tau^2$ is actually positive for any two distinct events on its trajectory. Hence, we can integrate $d\tau$ forward along the trajectory to obtain a parameter for the trajectory, called proper time $\tau$. That is, we may consider the time and the spatial position along the particle’s trajectory both to be functions of $\tau$.

Working in one space dimension for simplicity, define

$$\dot{p} = m \frac{dx}{d\tau}, \quad \text{relativistic momentum}$$

(31.6)

which is a new function defined along the trajectory. In this formula, $m$ is a constant with dimensions of mass, an intrinsic property of the particle. We’ll call it “the mass” of the particle.

We also introduce an analogous quantity

$$K = m \frac{d(ct)}{d\tau}.$$  

(31.7)

The point of these definitions is that then the pair

$$\begin{bmatrix} \kappa \\ \dot{p} \end{bmatrix} = m \frac{d}{d\tau} \begin{bmatrix} ct \\ x \end{bmatrix}$$

has the same linear transformation under Lorentz boosts as do $ct$ and $x$:

$$\begin{bmatrix} \kappa' \\ \dot{p}' \end{bmatrix} = m \frac{d}{d\tau'} \Lambda \begin{bmatrix} ct \\ x \end{bmatrix} = m \frac{d}{d\tau} \Lambda \begin{bmatrix} ct \\ x \end{bmatrix} = \Lambda \left( m \frac{d}{d\tau} \begin{bmatrix} ct \\ x \end{bmatrix} \right) = \Lambda \begin{bmatrix} K \\ \dot{p} \end{bmatrix}. \quad (31.8)$$

Here $\Lambda$ is a 2x2 Lorentz transformation matrix and we used the fact that $d\tau' = d\tau$. Note that we are allowed to pull the Lorentz transformation matrix outside the derivative because its entries are constants. Even if the particle is itself accelerating, nevertheless we are boosting to a coordinate system with some constant velocity $\beta c$ relative to the original one. In short, $K$ and $\dot{p}$ form a pair with a simple, linear transformation rule.

---

5See also Problem 31.1. We have previously used $\xi$ to denote a generic parameter along a trajectory; $\tau$ is specifically proper time. For the trajectory describing a light flash, however, $d\tau = 0$, so we must use some other parameterization, for example the one used in Section 30.6.1 (page 458).

6Old books introduce the term “rest mass.” That quantity is now simply called “mass,” because the alternative concept once called “relativistic mass” is no longer deemed worthy of any name at all.
We now propose two new conservation laws:

\[ \dot{p}(1) + \dot{p}(2) - \dot{p}(3) - \dot{p}(4) = 0 \quad \text{and} \quad K(1) + K(2) - K(3) - K(4) = 0, \tag{31.9} \tag{31.10} \]

which differ from the discredited newtonian versions.

**Ex.** Show that Equations 31.9–31.10 are automatically Lorentz invariant.

**Solution:** Equation 31.8 is analogous to Equation 31.2, and we can repeat the argument based on Equation 31.3.

Although we still haven’t proposed any detailed dynamical laws for collisions (possibly involving nuclear forces and so on), nevertheless we found a proposal for the corrected form of the momentum that leads to an acceptable conservation law. Indeed, for a slowly moving particle \( \bar{p} \) becomes equal to Newton’s momentum. To see this, note that

\[ d\tau = \sqrt{dt^2 - (vdt/c)^2} = dt\sqrt{1 - (v/c)^2} = \gamma^{-1}dt, \tag{31.11} \]

and \( \gamma \to 1 \) for a slowly moving particle. Thus, \( \bar{p} = m(dx/d\tau) \to m(dx/dt) = p^N. \)

What about the new quantity \( K \)? To find its meaning, note that Equation 31.11 gives \( K = mc\gamma \approx mc(1 + \frac{1}{2}(v/c)^2 + \cdots) \). So for a slowly moving particle \( cK \to mc^2 + \mathcal{E}^N \), that is, a constant plus the newtonian energy. Equation 31.10 then says that for small velocity,

\[ (m_1 + m_2 - m_3 - m_4)c^2 + \mathcal{E}_1^N + \mathcal{E}_2^N - \mathcal{E}_3^N - \mathcal{E}_4^N \approx 0. \tag{31.12} \]

Equation 31.12 is indeed compatible with the newtonian Equations 31.4 (which says the first four terms sum to zero) and 31.1 (which says that the next four also sum to zero).

More generally, we define

\[ \mathcal{E} = cK = mc \frac{d(ct)}{d\tau}. \quad \text{relativistic energy} \tag{31.13} \]

How can we dare to change the meaning of “momentum” and “energy”? The newtonian quantities are just not *useful*, because they cannot be conserved quantities in any Lorentz-invariant world. We found different quantities that *could* be conserved, and named them after the things they resemble. In fact, from now on we’ll follow other authors and drop the checks: \( p \), and its 3D generalization \( \bar{p} = m\overline{d\tau}/d\tau \), will henceforth refer only to the relativistic formula, and \( \mathcal{E} \) will always mean \( (mc)\overline{d(\tau)} / d\tau \). There won’t be any ambiguity, because from now on we won’t use the newtonian quantities at all. Generalizing to three spatial dimensions gives our proposed conservation law as an equality of 4D vectors:

\[ \left[ \frac{\mathcal{E}_{(1)}}{\bar{p}_{(1)}} \right] + \left[ \frac{\mathcal{E}_{(2)}}{\bar{p}_{(2)}} \right] - \left[ \frac{\mathcal{E}_{(3)}}{\bar{p}_{(3)}} \right] - \left[ \frac{\mathcal{E}_{(4)}}{\bar{p}_{(4)}} \right] = 0. \tag{31.14} \]

Chapters 32–33 will christen such quantities **four-vectors**.
Chapter 31  Relativistic Momentum and Energy of Particles

31.3.2 What has/has not been shown

We have shown that proposed conservation laws involving replacements for newtonian formulas, Equations 31.6 and 31.13, are at least compatible with the physical hypothesis that all of physics is Lorentz invariant. We would eventually like these formulas to emerge from some complete theory, but in 1905 it was too early for that. Instead, following “Einstein thinking,” we will postpone that project and instead look for direct experimental tests of the proposed conservation laws, Equations 31.14.

Later chapters will then develop the dynamical details, in the context of electrodynamics. Specifically, Chapter 35 will look for appropriate formulas for the energy and momentum of fields, then prove a conservation theorem about the total energy and momentum of particles and fields starting from the Maxwell equations and the Lorentz force law.

31.3.3 A geopolitical application

Having once committed ourselves to look at nature on its own terms, it is something like a point of honor not to flinch at what we see.

— Steven Weinberg

Newtonian physics assumes conservation of mass, and proves the conservation of energy, Equation 31.5, from the laws of motion. But Section 31.3.1 only obtained a single combined law, Equation 31.12, in the newtonian limit. Einstein realized that there was no fundamental reason why total mass must be unchanged in collisions. He concluded that a mass defect (change in total mass) must, if present, appear as nonconservation of kinetic energy in a collision reaction, and that the enormous energies involved in radioactive decay would be accompanied by a fraction of a percent change in mass. Thus, in 1907, he wrote that “Bodies whose energy content is variable to a high degree, for example radium salts,” might be used to test his prediction.

Experiments performed much later with the first particle accelerator confirmed Einstein’s prediction quantitatively: J. Cockcroft and E. Walton sent a beam of protons into $^7$Li and showed that the reaction products were two helium nuclei. With values available at that time, they estimated the mass lost in this reaction as about $2.56 \times 10^{-29}$ kg, consistent with their measurement of the total gain in kinetic energy of 17.2 MeV.

But already in 1907, in a laconic, eerily prescient remark, Einstein wrote, “It is possible that radioactive processes will be detected in which a significantly higher percentage of the

---

7Remarkably, today’s Standard Model’s interactions all look a lot like electrodynamics.
8The first complete, general derivation appears to be due to Max von Laue in 1911.
9Einstein was up to date: Rutherford/Barnes and Soddy/Ramsey had measured the energy of a single decay of radon just two years earlier (1903), finding it to be over a million times the energy released when hydrogen and oxygen combine to form a molecule of water.
10For a modern measurement with precision $4 \cdot 10^{-7}$, see Further Reading. F. Aston had previously shown that the mass of a helium nucleus is slightly smaller than the sum of the masses of two isolated deuterium nuclei, but was not able to induce the corresponding reaction.
mass of the original atom will be converted into the energy of a variety of radiations than in the case of radium.” Such reactions, if they existed, would therefore liberate even more energy per kilogram of fuel than natural radioactivity. Many years later, such reactions were found, including nuclear fission and fusion.

That was definitely a practical prospect. Eventually, everybody realized that if you could slowly release the energy equivalent of a gram of matter, you’d get $10^{14}$ J, plenty to run a big city for a long time. Everybody also realized that if you could do the same conversion in a few microseconds, you could burn that city to the ground.

Nobody knew in 1907 whether these possibilities could be implemented in practice. But within a few decades the outlines began to form. The belligerents in the second World War embarked on urgent crash programs to develop such weapons. An entire world vanished forever on 16 July, 1945.

### 31.4 PARTICLES WITH SPEED AT OR NEAR $c$

#### 31.4.1 Any particle interpretation of light must involve the limit of zero mass

Suppose that a particle’s speed approaches $c$, that is, suppose that $\beta \to 1$. In this limit, we expect Newton’s formulas to be badly inaccurate. Equations 31.6, 31.7 and 31.13 give

$$\frac{p}{\mathcal{E}} = \frac{m \frac{dx}{dt}}{cm \frac{d(\mathcal{E})}{dt}} = c^2 \frac{dx}{dt} \to \frac{1}{c^2},$$

or

$$\mathcal{E} \approx pc.$$ (31.15)

This is precisely the relation that we found earlier for energy and momentum fluxes of a classical plane wave solution! So a dual, quantum-mechanical interpretation of light seems possible after all: The “missing” factor of $1/2$ that we noticed earlier is actually just as it should be. What was wrong was the expectation that newtonian formulas should apply to things moving at speed $c$.

One may object that as $\beta \to 1$, our formula for $\gamma \to \infty$, and hence also the momentum becomes infinite! Indeed, there is no way to push an ordinary particle (one with nonzero mass) up to speed $c$. However, we can imagine a limit in which $\beta \to 1$ and $m \to 0$ in just such a way that $p \to$ constant:

*The only way for a particle to move at speed $c$ is for it to be massless. The only way for a massless particle to have nonzero energy and momentum (and hence to exist at all) is for it to be moving at $c$. We can take the limit in various ways, so any values of $p$ and $\mathcal{E}$ are allowed, as long as $\mathcal{E} = pc$.*

So that’s another viewpoint on why light always moves at a universal speed. The dual particle and wave viewpoints are compatible, at least insofar as kinematics is concerned.

---

11Equation 20.9 in Section 20.3 (page 302). Section 55.4 will derive this relation for photons by quantizing the field.
It’s no accident that when Einstein was working on his light-quantum hypothesis, he was also working out special relativity.\(^\text{12}\)

### 31.4.2 Interactions involving massless particles

The newtonian conservation laws allow us to predict the results of collisions among, say, two balls that collide and stick. Similarly, our proposed Lorentz-invariant conservation laws allow us to make a falsifiable, quantitative prediction for the result when, say, an x-ray photon collides with an electron at rest. The successful test of this Compton scattering process lent credence not only to the photon hypothesis, but also to relativity itself.

You may still be bothered, however: \textit{How can a “real thing” have no mass?} Maybe the following thought experiment will help. Imagine a box whose interior walls are perfect mirrors. Initially there’s no light inside. The box will have some resistance to acceleration (inertia), which we describe by a mass \(m_{\text{box}}\). Now imagine filling the box with lots of light, but changing nothing else. The light carries energy, but its net momentum is zero. The relation between energy and mass implies that the mass of the light-filled box is greater than the empty box, even though they differ only by the presence of particles that, taken individually, obey \(E = pc\).

### 31.5 PLUS ULTRA

This concludes our study of “low-tech relativity.” Although the structure is logically satisfying, the discussion has emphasized that Einstein’s version of relativity is justified only because it makes predictions for real experiments (not just thought-experiments). Those predictions were confirmed, and they differed from the corresponding newtonian predictions.

We are starting to see something remarkable: The four coordinates (\(ct\) and spatial position \(\vec{r}\)) undergo a peculiar kind of linear transformation, a little like rotations.\(^\text{13}\) And now we see that \(E/c\) and \(\vec{p}\) undergo the same peculiar but linear transformation (Equation 31.8). This observation suggests that there may be a tensor formalism describing such quantities, and other more elaborate ones. Just as 3-tensor notation helped us to classify quantities and formulate rotationally-invariant laws of Nature, so we will find in Part V that “4-tensor” notation will help us to deal systematically with the consequences of the hypothesis that Nature is Lorentz-invariant. Briefly, we will set up a parallel between the newtonian framework:

\textit{3D Euclidean geometry}: Cartesian coordinates are the ones in which the pythagorean formula takes its usual form. All cartesian coordinate systems are related to one another by euclidean group transformations (translations and rotations, plus reflections). The components of a three-tensor have definite, linear transformations under rotations. Every physical quantity in newtonian physics belongs to (is a component of) some

\(^{12}\text{Section 29.5 already disposed of another objection.}\)

\(^{13}\text{Section 30.3.2 (page 455) pointed this out.}\)
3-tensor. Any law of physics that sets a 3-tensor equal to zero, such as Equation 31.1, is automatically rotation-invariant.

and its proposed replacement:

4D spacetime geometry: E-inertial coordinate systems are the ones in which the invariant interval has its usual form. All E-inertial coordinate systems are related by Poincaré group transformations (translations, rotations, and Lorentz boosts, plus reflections). The components of a four-tensor have definite, linear transformations under Lorentz transformations. Every physical quantity in true (Lorentz-invariant) physics belongs to (is a component of) some 4-tensor. Any law of physics that sets a 4-tensor equal to zero, such as Equation 31.14, is automatically Lorentz-invariant.

The second of these frameworks will prove helpful as we get to work proving that the full Maxwell equations are Lorentz-invariant, and it will also have practical benefits for solving harder problems than the ones we’ve done so far.

**FURTHER READING**

*Semipopular:*

*Technical:*
Historic confirmation of $E = mc^2$: Cockcroft et al., 1932.
Modern version: Rainville et al., 2005.
Einstein 1907: Einstein, 1907.
Chapter 31 Relativistic Momentum and Energy of Particles

PROBLEMS

31.1 Proper time
Section 31.3.1 claimed that the trajectory of any material particle (that is, not a photon) is a curve in spacetime that can conveniently be parameterized by its proper time, which is the same in any E-inertial coordinate system. This claim is supposed to hold even in the full three spatial dimensions, and even for particles that are not free, that is, particles that are being accelerated by some force. You can establish it as follows.

Suppose that we are given a trajectory specified by four functions \( t(\xi) \) and \( \vec{r}(\xi) \). The parameterization is arbitrary, except that time \( t \) is strictly increasing as a function of \( \xi \). To be physical, the trajectory must always be moving with speed less than \( c \), or in other words \( ||d\vec{r}/d\xi||^2 < (c dt/d\xi)^2 \) everywhere. Show how to obtain a new parameter \( \tau \) (an increasing function of \( \xi \)) with the property

\[-(c d\tau)^2 + ||d\vec{r}||^2 = -c^2 (dr)^2.\]

31.2 Recoil
The unstable nucleus \(^{60}\text{Co}\) decays in two steps to an excited state \(^{57}\text{Fe}^+\), which then drops to the ground state emitting a photon. The last step releases \( \Delta \mathcal{E} = 14.4 \text{ keV} \). The half-life of this transition is long, so the natural width of the spectral line, set by the Uncertainty Relation, is fantastically narrow: The fractional width \( \Delta \mathcal{E}/\mathcal{E} \) is \( \approx 10^{-12} \). Conversely, the absorption spectrum for \(^{57}\text{Fe}\) to get excited by an incoming photon is equally narrow (Figure 30.6).

An isolated \(^{57}\text{Fe}^+\) nucleus will give off a photon with slightly reduced energy, because \( \Delta \mathcal{E} \) must be shared between the photon and the kinetic energy of the recoil of the nucleus.

a. Find the recoil kinetic energy if the iron nucleus is isolated. The mass of an \(^{57}\text{Fe}\) nucleus is 56.9 Da. A convenient definition of the dalton is 1 Da = 931.5 MeV/\(c^2\).

b. What is the corresponding fractional reduction of the energy of the photon?

c. Could the photon emitted by a free nucleus initially at rest in the lab be reabsorbed by another such nucleus?

Remarkably, for atoms in a crystal lattice there is a significant probability that the final state will involve rigid recoil motion of the entire crystal, not just the one nucleus that made the transition. The mass of the entire crystal is essentially infinite, so the kinetic energy of its final state is essentially zero. This “recoilless emission” is called the Mössbauer effect. Because no energy is lost to recoil, some of the outgoing photons from a bulk sample get the entire \( \Delta \mathcal{E} \), and hence can be resonantly absorbed by a ground-state iron nucleus in a target. A tiny frequency shift can push the photons off resonance, making Mössbauer spectroscopy the basis for extremely accurate measurements.

\[1^{4}\text{In quantum language, there is a significant probability of creating zero phonons (no lattice vibrations) in the transition.}\]
d. Suppose that the emitter and absorber are in uniform motion with relative velocity $v$ that is parallel to their separation. Then even those photons that were emitted without recoil will be Doppler shifted. How large may $v$ be before those photons can no longer be resonantly absorbed without recoil?

[Remark: Almost immediately after the discovery of the Mössbauer effect, R. Pound and G. Rebka applied it to confirm Einstein’s prediction of a tiny gravitational redshift effect.]

31.3 Pair production on the CMB

[[Not ready]]
CHAPTER 32

Four-Vectors

[Einstein was asked] to give the manuscript of the June [1905] paper to the Book and Authors War Bond Committee. Einstein replied that he had discarded the original manuscript, but added that he was prepared to write out a copy of its text in his own hand. Helen Dukas sat next to Einstein and dictated the text to him. At one point, Einstein laid down his pen, turned to Helen and asked her whether he had really said what she had just dictated to him. When assured that he had, Einstein said, ‘Das hatte ich einfacher sagen können.’ ['I could have said that more simply.]

— Abraham Pais

32.1 FRAMING: UNIFICATION

This chapter begins developing what one might call “high-tech relativity.” We will rediscover some results already seen in the preceding Parts III–IV. Why repeat?

• The high-tech approach is abstract and framed in the language of linear algebra. Physical intuition was better served by seeing first what could be seen from the older viewpoint, and by building that viewpoint based on a few key experiments.
• But the high-tech approach unifies various ideas that may have seemed disconnected previously. Before we press on to new results, it is important to see how compactly we can regenerate the old ones.

Electromagnetic phenomenon: Electrons, protons, and other “material” particles also show wavelike behavior.
Physical idea: The relation between wavelength and momentum is uniquely determined by relativistic invariance.

32.2 HOW TO AVOID READING THIS CHAPTER

We are studying the Maxwell equations for fields, plus the Lorentz force law for charged particles. We have seen that these equations correctly describe many phenomena.

We abstracted Lorentz invariance from just a subset of these equations (the scalar wave equation). We got some experimentally testable predictions (Fizeau experiment, aberration of starlight, mass–energy equivalence.) But so far we neglected the vector characteristic of the fields, and hence also polarization of light. We now want to build a bridge
between the full Maxwell equations and the hypothesis of Lorentz invariance. To do this, we’ll construct a grammar of Lorentz-invariant constructs, which we can then stick together (following some grammatical Rules) to build equations that are guaranteed to be Lorentz invariant. Then we’ll see that the Maxwell equations can be expressed in that language.

You should read and work through this chapter. However, nothing stops us from considering the hypothetical student who wants the plot spoilers up front:

Up and down indices

This chapter, and the next two, will develop the modifications to tensor analysis needed to make relativistic invariance obvious at a glance in equations of motion, just as ordinary vector/tensor notation makes rotational invariance obvious at a glance. A key complication is that we will need to keep track of two kinds of coordinate index, which will be called “up” and “down” indices. Why, when all your life one kind has been sufficient?

The answer will turn out to be that derivatives \( \frac{d}{d(ct)} \) and \( \vec{V} \) transform differently from coordinates \( (ct, \vec{r}) \). In euclidean 3-space, if we use cartesian coordinates, then we can forget about the distinction. In the non-euclidean space that we’ll develop for relativity, we do have to keep track of it.¹

Luckily, we’ll find a set of notational Rules that will make it unnecessary to think much about this complication. Once we’ve justified The Rules, we’ll see they are easy to follow. You could, hypothetically, just jump to Section 34.4.

Chicken and egg

We have accumulated some evidence that a new group of transformations may be symmetries of electrodynamics, and indeed of all of Physics. But now we seem to face a chicken-and-egg problem: How can we prove that the Maxwell equations are invariant under these transformations, when we don’t know how the \( \vec{E} \) and \( \vec{B} \) fields should transform? The thought-experiment about the coil and magnet has suggested that, under a boost transformation, the components of electric and magnetic fields should mix.² It sounds complicated. Once we make the right guess we can confirm it by mathematical operations… but how do we make the right guess?

Thinking back, the structure of electrodynamics as presented so far is that we took the Lorentz force law as a starting point; it gave an operational meaning to \( \vec{E} \) and \( \vec{B} \). Once those vector fields were defined, then the Maxwell equations make falsifiable predictions about their relations to each other and to charges and currents. So Section 33.3 will again begin with the Lorentz force law, asking:

1 Can it be formulated (perhaps with modifications that are small in the world of slowly-moving objects) in a way that is Lorentz-invariant?
2 If so, what does that say about the transformation properties of \( \vec{E} \) and \( \vec{B} \)?

¹ Even in euclidean space, if we use curvilinear coordinates, the distinction matters (Section 32.4.2), which is why we have done nearly all our tensor analysis in cartesian coordinates.
² Hanging Question #A (page 13). See also Problem 30.3.
Are the Maxwell equations also invariant under those transformations? The plot spoiler is that the answers are:

1. Yes. The only needed correction is unsurprising: Substitute relativistic momentum for newtonian momentum.
2. **The electric and magnetic fields together form a single 4-tensor field.** When we transform to a new inertial coordinate system, the components of \( \vec{E} \) and \( \vec{B} \) scramble among themselves, just as the components of the quadrupole tensor in electrostatics mix under rotations. We are going to make this analogy precise.
3. Yes. No further corrections (beyond Maxwell’s modification of the Ampère law) will be needed at all.

You could, hypothetically, jump ahead to Equations 33.3 and 33.6 to see how it works.

### 32.3 3D REVIEW

Every physical quantity carries some discrete information about its status: its dimensions, which are powers of a few basic symbols (L, M, and so on). Keeping track of dimensions helps us to formulate correct equations and spot incorrect ones. Earlier chapters mentioned an equally powerful principle: Physical quantities carry an independent sort of discrete status, because each one belongs to some class of tensors. This section will review and extend some material introduced in Chapters 13–14.

#### 32.3.1 Rotations preserve the form of the metric

The components of a 3-vector \( \vec{r} \), in a particular cartesian coordinate system, are three numbers \( \vec{r}_i, i = 1, 2, 3 \). These numbers represent the vector, which is itself a geometrical object (magnitude and direction).

When we change to another right-handed, cartesian coordinate system, the same vector is represented by three different numbers \( \vec{r}'_a \), where

\[
\vec{r}'_a = S_{ai}\vec{r}_i \quad (\text{and} \quad t' = t).
\] (32.1)

The matrix \( S \) is a set of nine constants. Again, prime denotes a new coordinate system. For extra clarity, we will often use coordinate indices \( i, j, ... \) from the middle of the alphabet for one coordinate system, but \( a, b, ... \) from the start of the alphabet for the alternative coordinate system.

The matrix \( S \) is not arbitrary, because cartesian coordinate systems have the property that the pythagorean formula always has the same form:

\[
||\vec{r}'||^2 = \vec{r}'_a \vec{r}'_a = (||\vec{r}||')^2 = \vec{r}'_a \vec{r}'_a = S_{ai}\vec{r}_a S_{aj}\vec{r}_j.
\] (32.2)

---

3 This is also Equation 14.1. \( S \) is set in sans-serif to remind us it’s a matrix. But it doesn’t get any arrow because it’s not a tensor: Instead of having a tensorial transformation rule under change of coordinates, it specifies a change of coordinates.

4 This is also Equation 14.3. The pythagorean formula doesn’t have this form in curvilinear coordinates, but we will stick to representing tensors in cartesian coordinates.
It will sometimes be convenient to use the mathematician’s matrix notation. We indicate the components of vectors and matrices with square brackets, omit explicit indices, and imply summations with the usual rules of matrix multiplication. Thus, Equation 32.2 involves

\[ \begin{bmatrix} \vec{r}' \end{bmatrix} = \begin{bmatrix} S & \vec{t} \end{bmatrix} \begin{bmatrix} \vec{r} \end{bmatrix}. \] (32.3)

Matrix notation is very compact, but you have to be careful about the order in which you write things.

The expression in Equation 32.3 will equal \( \begin{bmatrix} \vec{r} \end{bmatrix} \begin{bmatrix} S & \vec{t} \end{bmatrix} \), for any \( \vec{r} \), only if \( S \) has the property\(^5\)

\[ S^T S = 1. \] transformation between cartesian systems (32.4)

Both sides of Equation 32.4 are symmetric matrices, so some of the nine components of this equation are redundant: It amounts to just six independent constraints on the entries of \( S \). Therefore we expect a family of solutions with \( 9 - 6 = 3 \) parameters—for example, the three Euler angles used to specify a rotation.

Here is some mathematical terminology. A real matrix that satisfies Equation 32.4 is called orthogonal\(^6\). If two matrices \( S \) and \( T \) are both orthogonal, then so is the product \( ST \) (and also \( S^{-1} \)). Thus, orthogonal matrices close into a group, a notion previously introduced in Section 30.2. The group of orthogonal \( 3 \times 3 \) matrices is sometimes called \( O(3) \). Taking the determinant of both sides of Equation 32.4 shows\(^7\) that \( \det S = \pm 1 \).

Rotation matrices have the additional property that \( \det S = +1 \); orthogonal matrices without this property include inversion, \( \vec{r} = -\vec{r} \), and reflection through any plane\(^8\). Again, if \( S \) and \( T \) both meet this extra condition, then so will their product, and so will \( S^{-1} \), so rotations also close into a group. We say they form a subgroup of \( O(3) \) called the special orthogonal matrices, or \( SO(3) \). Any two right-handed, cartesian coordinate systems are related by an \( SO(3) \) matrix.

### 32.3.2 Equations of the form (3-vector) = 0 are rotationally invariant

Any three-component quantity whose entries transform in the same way as \( \vec{r} \) when we change from one cartesian coordinate system to another can represent a 3-vector, or 3-tensor of rank 1. For example, the time derivatives \( d\vec{r}/dt \) and \( d^2\vec{r}/dt^2 \) are also 3-vectors, because rotations and spatial reflections don’t affect time. The vector sum of two 3-vectors is itself a 3-vector, because

\[ S_{ai} \vec{u}_i + S_{aj} \vec{w}_j = S_{ai} (\vec{u}_i + \vec{w}_j). \]

Similarly, if we multiply a 3-vector by, say, 2, the result is again a 3-vector.

---

\(^5\)This is also Equation 14.4.

\(^6\)See Section 14.2.2 (page 210).

\(^7\)This is also Equation 14.5.

\(^8\)Any matrix in this sector can be expressed as a rotation combined with spatial inversion.
Now consider Newton’s law for an isotropic, harmonic oscillator with viscous friction:

\[ m \left( \frac{d^2 \vec{r}}{dt^2} \right) = -k \vec{r} - \zeta \left( \frac{d \vec{r}}{dt} \right). \]  

(32.5)

Multiply everything from the left by \( S \):

\[ S \left[ \left( m \left( \frac{d^2 \vec{r}}{dt^2} \right) \right) + k[\vec{r}] + \zeta \left( \frac{d \vec{r}}{dt} \right) \right] = 0. \]

Now push the constant matrix \( S \) inside the derivatives:

\[ m \left( \frac{d^2 \vec{r}'}{dt^2} \right) + k \vec{r}' + \zeta \left( \frac{d \vec{r}'}{dt} \right) = 0. \]

This shows that Equation 32.5, re-expressed in the primed coordinate system, retains its original form: It’s invariant under rotations.

A bit more precisely, we got rotational invariance under the assumption that \( t, m, k, \) and \( \zeta \) were all unaffected by the rotation: They are scalars, also called 3-tensors of rank zero. Of these, \( m, k, \) and \( \zeta \) are scalar constants, whereas \( t \) is a scalar variable.

### 32.3.3 The 3-tensor transformation rule

Next, consider an anisotropic, but still linear, system of springs (this time without friction). There is a coordinate system for which every allowed motion is a solution to the equation

\[ m \frac{d^2 \vec{r}}{dt^2} = -\vec{K}(\vec{r}), \]

(32.6)

where \( \vec{r} \) is a vector-valued linear function of \( \vec{r} \). Chapter 13 called such a function a tensor, and Section 13.3 (page 196) explained how to represent it via an array of components: Each component of the right side of Equation 32.6 may then be written \( \vec{K}_{ij} \vec{r}_j \), or simply

\[ m \frac{d^2 \vec{r}}{dt^2} = -\vec{K} \cdot \vec{r}. \]

(32.7)

To see whether and in what sense Equation 32.6 is rotation invariant, let’s try the same approach used in Section 32.3.2: Again multiply everything from the left by a rotation matrix:

\[ S[m \left( \frac{d^2 \vec{r}}{dt^2} \right)] = -S[\vec{K} \cdot \vec{r}]. \]

It may seem that we have hit a wall, because \( S \) will not in general commute with the matrix of \( \vec{K} \)’s components. But consider the following restatement:

\[ m \left[ \frac{d^2 \vec{r}'}{dt^2} \right] = -S[\vec{K} S^{-1}[\vec{r}']] = -S[\vec{K} S^{-1}[\vec{r}']}. \]

This new version has the same form as the original equation, albeit with a modified spring tensor:

\[ \vec{K}'_{ab} = S_{ai}(S^{'})^{-1}_{bj} \vec{K}_{ij}. \]

\(^9\)Point 3 on page 199 introduced this example.
We have reverted to explicit-index notation, so that we can write the factors in any order we please. This formula simplifies when we recall that \( S^{-1} = S^\dagger \) (Equation 32.4):

\[
\vec{K}'_{ab} = S_{ai} S_{bj} \vec{K}_{ij}.
\]  

(32.8)

Any nine-component quantity whose entries transform in this way can represent a 3-tensor of rank 2. We say that one copy of \( S \) “acts on” each index of \( \vec{K} \).

Thus, we find that expressing Equation 32.6 in terms of rotated coordinates indeed yields an equation of the same form,

\[
m \frac{d^2 \vec{r}'}{dt^2} = -\vec{K}' \vec{r}' = m \vec{F}',
\]

if we acknowledge that the spring system must be rotated as well as the trajectory, or in other words that the components of \( \vec{K} \) must transform in the way appropriate for a tensor of rank two.

The dyad product \( \vec{r} \otimes \vec{r} \) is another example of a 3-tensor of rank 2, because each factor separately contributes an \( S \). More generally, we can define 3-tensors of any rank \( p \): They are represented by collections of \( 3^p \) components, with a transformation law involving \( p \) copies of \( S \). The matrix sum of the components of two 3-tensors itself specifies a 3-tensor, because

\[
S_{ai} S_{bj} \vec{K}_{ij} + S_{ai} S_{bj} \vec{L}_{ij} = S_{ai} S_{bj} (\vec{K}_{ij} + \vec{L}_{ij}),
\]

and similarly for scalar multiplication.

Returning to the spring system, suppose that our mass is suspended between three springs stretched along the original \( x \), \( y \), and \( z \) axes respectively. Then

\[
\vec{K} = A \vec{x} \otimes \vec{x} + B \vec{y} \otimes \vec{y} + C \vec{z} \otimes \vec{z},
\]

which indeed is explicitly a 3-tensor, because each of its terms is separately a 3-tensor.

If we have two tensors of rank \( p \) and \( q \) respectively, then we can generalize the dyad product by forming all products of their elements, a total of \( 3^{p+q} \) numbers carrying \( p + q \) indices. That suggests that these numbers form the components of a rank-(\( p + q \)) tensor, called the tensor product, and indeed it’s true by the same argument used earlier for the dyad product (which is the case \( p = q = 1 \)).

### 32.3.4 Symmetric and antisymmetric 3-tensors

A spring tensor has the property that \( \vec{K}_{ij} = \vec{K}_{ji} \), or in matrix language \( [\vec{K}] = [\vec{K}]^\dagger \). The quadrupole moment tensor from Chapter 3, and the moment of inertia tensor, also have this “symmetric” property.

---

10This is also Equation 14.2.
11This logic is similar to that in Section 27.4 (page 423).
12Thus, the electric quadrupole tensor and the moment of inertia tensor are physical quantities specified by 3-tensors of rank 2.
Your Turn 32A

a. Show that if a tensor is symmetric in one cartesian coordinate system, the same will be true after transformation via Equation 32.8.

b. Show that the sum of two symmetric tensors of the same rank is symmetric.

Thus, the property of being symmetric is itself a rotationally-invariant property of a tensor, and hence something that we may legitimately specify without spoiling rotational invariance.

Similar remarks apply to antisymmetric tensors, for example, the magnetic field tensor $\tilde{\omega}$ or the magnetic dipole moment tensor $\tilde{1}^\mathbf{I}$.

If a three-tensor $\tilde{T}$ is not symmetric, then its transpose represents a new tensor of the same rank. That tensor can then be added/subtracted from the original version to produce the “symmetric/antisymmetric parts” of $\tilde{T}$,

$$\tilde{T}^\mathbf{S} = \frac{1}{2} (\tilde{T} + \tilde{T}^t), \quad \tilde{T}^\mathbf{A} = \frac{1}{2} (\tilde{T} - \tilde{T}^t)$$

respectively. Then $\tilde{T} = \tilde{T}^\mathbf{S} + \tilde{T}^\mathbf{A}$.

32.3.5 3D contraction is another invariantly defined operation

The dot product of two vectors may be thought of as the trace of their dyad product, a 3-scalar. More generally, the trace of any rank-two tensor is a scalar (Section 13.3.1). Similarly, the dot product, or “contraction,” on the right side of Equation 32.7 “absorbs” two indices, leaving just one uncontracted index. That is, contraction reduces the rank of the right side from three to one, whereupon it matches the left side.

More generally still, suppose that $T_{i_1 \cdots i_p}$ are the components of a rank-$p$ tensor. We choose two positions $K$ and $L$ in the index list, set the indices equal, and sum them, leaving the remaining $p - 2$ indices loose. The result is a set of $3^{p-2}$ numbers, which we can write as

$$\Omega_{i_1 \cdots \hat{K} \cdots \hat{L} \cdots i_p}$$

where the notation $\hat{K}$ means “omit this index.” The notation suggests that these numbers form the components of a rank-$p - 2$ tensor, called the contraction of $T$ on the chosen indices. Indeed,

$$\Omega'_{a_1 \cdots \hat{a}_K \cdots \hat{a}_L \cdots a_p} = S_{a_1 l_1} \cdots S_{b l_K} \cdots S_{b l_L} \cdots T_{l_1 \cdots l_p}$$

The factors in braces, summed over $b$, yield $[S^i S]_{l_k l_p}$, which is the identity matrix. The result is therefore equal to the contraction of $T_{i_1 \cdots i_p}$, transformed in the usual way on the remaining $p - 2$ indices. In short, contraction of a tensor again yields a tensor, with rank lowered by two.

---

13See Sections 15.2 (page 222) and 17.2 (page 252).
14There are corresponding operations on 3-tensors of rank 3 as well, involving sums over all six permutations of the indices (Equation 15.10, page 226). However, the totally antisymmetric and totally symmetric parts do not suffice to reconstruct the original tensor (as they do for rank two).
32.4 OTHER ROTATIONALLY INVARIANT SYSTEMS IN MECHANICS

32.4.1 Newtonian gravitation

Here is another example, mentioned in Section 26.5 (page 408): To study celestial mechanics, we combine Newton’s Second Law with his law of gravitation for a mass $M$ that is anchored at the origin:

$$m\left(\frac{d^2\vec{r}}{dt^2}\right) = -\frac{GMm}{r^3}\vec{r}. \quad (32.10)$$

To analyze this equation’s symmetry, begin with the denominator. It involves $r = \sqrt{||\vec{r}||^2}$, which we saw in Section 32.3 is unchanged upon rotations about the origin. So the right-hand side of Equation 32.10 is a scalar constant $-GMm$, times a scalar function $r^{-3}$, times the 3-vector $\vec{r}$. All together, it’s therefore a 3-tensor of rank $0 + 0 + 1$: a 3-vector. Setting it equal to the left side then yields a rotationally-invariant equation, just as in the harmonic oscillator.

Your Turn 32B

a. Equation 32.10 assumes that the Sun is fixed in space. Write the more general form in which two gravitating bodies (“Sun” and “Jupiter”) are both free in space, and show that the equations are now rotationally-invariant about any point.

b. Show that expanding the scope of the system in this way (acknowledging that $\vec{r}_{\text{Sun}}$ is a dynamical variable) also restores explicit translation invariance. This property was hidden in Equation 32.10, which appears to have a special point at $\vec{r} = \vec{0}$.

32.4.2 Field equations: the gradient operator

We can also discuss field equations in this language, for example, Newton’s gravitational field equation:

$$\nabla^2 \phi_N = 4\pi G_N \rho_m. \quad [1.1, \text{page 17}]$$

First notice that the Chain Rule from calculus gives

$$\vec{\nabla}_i \equiv \frac{\partial}{\partial r_i} = \frac{\partial \vec{r}_a'}{\partial r_i} \frac{\partial}{\partial \vec{r}_a'} = S_{ai} \frac{\partial}{\partial \vec{r}_a'}, \text{ or} \quad (32.11)$$

$$\vec{\nabla}_a' = S_{ai} \vec{\nabla}_i. \quad (32.12)$$

We again used the fact that $(S^t)^{-1} = S$.

Equation 32.12 is of the same form as Equation 32.1: $\vec{\nabla}$ itself transforms as a vector. More precisely, the gradient of a scalar function (like temperature), is a vector field (telling us locally which direction to go if we seek higher temperature). This is the step that will fail in 4D, requiring us to introduce two kinds of index.\(^{15}\)

\(^{15}\)Section 34.2.1 (page 526) will pick up this thread. This is the step that fails if we use noncartesian (curvilinear) coordinates, even in 3D euclidean space. However, the approach of Chapter 34 (doubling index type) applies there as well. (In the curved spacetime of general relativity, there may be no cartesian coordinate systems.)
From Equation 32.12, you can prove that $\tilde{V}_j \tilde{V}^j = \tilde{V}_a \tilde{V}^a$, and hence that Equation 1.1 is rotationally invariant if we take $G_N$ to be a scalar constant, and the mass density $\rho_m$ and the gravitational potential $\phi_N$ to be scalar fields.

To practice and extend the concepts, consider the velocity vector field of a fluid, $\tilde{v}(\tilde{r})$. We may be interested in whether the velocity is uniform in space.

**Your Turn 32C**

Show that
a. $\tilde{V} \circ \tilde{v}$ is a rank-two tensor field; and
b. $\tilde{V} \cdot \tilde{v}$ is a scalar field, that is, an ordinary function.

c. What can we say about the strain rate tensor, whose components are $\tilde{V}_j \tilde{v}_j + \tilde{V}_j \tilde{v}_j$?

In short, $\tilde{V}$ increases the rank of any tensor field by one.

Section 32.4.2 (page 507) gives more details on the transformation of vector and tensor fields.

### 32.5 SUMMARY: THE RULES IN 3D

The foregoing observations suggest a vast generalization, which we’ll call the **Tensor Principle**:16

\[
\text{Physical quantities arrange themselves into 3-tensors (or 3-tensor fields), in some cases constrained by symmetry or antisymmetry. Physical laws are rotationally invariant, and moreover can be written in manifestly invariant form by exploiting simple Rules.}
\]

(32.13)

Idea 32.13 introduces the term “manifest invariance”: A property is manifest if it can be verified at a glance, in this case by checking that some Rules have been followed.

Let us collect those familiar Rules, which you have been using all your life. Taking a moment to state them out loud will help us to generalize them. Some were proved earlier in this chapter; others are easy (but worthwhile) to prove now:

a. A 3-tensor of rank $p$ can be represented in a particular cartesian coordinate system by a collection of $3^p$ numbers, with a transformation law involving $p$ copies of an orthogonal matrix $S$, each “acting on” an index. For rotations, $S$ must also satisfy $\det S = +1$ [Section 32.3.3].

b. A 3-tensor field of rank $p$ is the same idea, but each entry is a function of $\tilde{r}$.

c. Permuting the indices on the components of a tensor yields another tensor of the same rank [Section 32.3.3].

---

16 Quantum mechanics amends this claim slightly to allow an additional class of quantities called “3-spinors” (Section 34.11, page 545), but with the proviso that spinor fields are not directly observable.
d. The sum of two tensors with the same rank yields a new tensor of that same rank. The new tensor has components that are sums of corresponding components of the two summands; for example, $\mathbf{A}_{ij} + \mathbf{B}_{ij}$ are the components of a new tensor of rank 2 [Section 32.3.3]. Combining with (c) shows that $\mathbf{A}_{ij} + \mathbf{B}_{ji}$ also makes sense. However, every term must have exactly the same set of loose indices; $\mathbf{A}_{i} + \mathbf{B}_{j}$ does not specify the components of any tensor.

e. The collection of all products of the components of a rank-$p$ and a rank-$p'$ tensor itself constitutes a rank-$(p + p')$ tensor called the tensor product. For example, the dyad product of vectors is a tensor of rank 2 with components $(\mathbf{a} \otimes \mathbf{b})_{ij} = a_i b_j$ [Section 32.3.3].

f. Contraction (dot product, trace) is an invariant operation that converts a tensor, or tensor field, to another one with rank decreased by 2 [Section 32.3.3].

g. The derivative operator $\mathbf{\nabla}$ increases the rank of a tensor field by 1 [Section 32.4.2].

h. A physics equation of the form $A_{i1,j2,\ldots} = 0$, where $A$ is a tensor, is rotationally invariant. Hence, the same is true for an equation of the form $A = B$, where both $A$ and $B$ are tensors (or tensor fields) of the same rank. Examples include Equations 32.5, 32.6, and 32.10.

i. The volume element $d^3r$ transforms to $d^3r'$ under rotations because the jacobian matrix has $|\det S| = 1$. Thus, we may convert any tensor field to an ordinary tensor of the same rank by integrating over all space.

It may have seemed that “pseudo” quantities such as magnetic field, torque, and so on were exceptions to Idea 32.13, but we saw how they can be repackaged as true tensors that are directly defined from physical measurement. For example, Section 15.2 (page 222) re-expressed magnetic field as an antisymmetric rank-2 tensor field $\mathbf{\omega}$, and Equation 15.1 (page 222) showed how to determine it by measuring force and velocity without choosing any right-hand convention.

Note that galilean boosts are not as simple as rotations. Diagnosing whether an equation has this important invariance is not just a matter of glancing at its index structure. We won’t need to deal with this, however, because we’re pursuing the hypothesis that the world is not galilean invariant after all.

32.6 FOUR DIMENSIONS

32.6.1 Packaging

We want to construct an upgraded tensor analysis in which the inertial coordinate systems in Einstein’s version of relativity play a role analogous to the cartesian coordinate systems in 3D. That is, we want a formulation of physics in which invariance under the Lorentz transformations, which take us from one $\mathbf{E}$-inertial coordinate system to another, is an obvious property of the equations of motion. The Lorentz transformations modify both

---

17 We place no arrow over the $r$ because this object is a scalar (not a vector).
the space and time coordinates describing events. So we introduce a new four-component object that, in a particular inertial coordinate system, is represented by the components:

\[ X^\mu = \begin{bmatrix} c \gamma \nabla \phi \varepsilon \nabla \phi \end{bmatrix} = \begin{bmatrix} c \gamma \nabla \phi \varepsilon \nabla \phi \end{bmatrix}. \] (32.14)

Here \( \mu \) is an index that runs over the four values 0, \ldots, 3. Note the conventions:

- *Time is regarded as coordinate number zero*, or more precisely, \( x^0 = ct \).
- The index indicating which coordinate we’re discussing is placed in the upper position, not lower as we always do in three dimensions. Thus, \( x^1 \) is the quantity we’ve been calling \( x \) or \( \vec{r} \), up till now, and so on.\(^{18}\)
- Instead of over-arrows, we’ll flag 4-tensor quantities with an underscore. Similarly to 3D, and unlike some other books, we retain the underscore even when talking about a particular component, to emphasize that although \( x^1 \) is a single number, still it’s *not* a scalar; it is a component of a 4-vector.\(^{19}\)

As with 3-vectors (Equation 32.3), we will sometimes write \[ X \] as an abbreviation for the components \( x^\mu \) regarded as a column (that is, we suppress the explicit index \( \mu \)) and use the rules of matrix multiplication to imply summations. Then \( [X] \) is the corresponding row vector.

**32.6.2 The Lorentz group and its main subgroups**

We are exploring certain linear transformations on the coordinates representing an event (that is, a point in spacetime):

\[ X'^\mu = \Lambda^\mu_\nu X^\nu, \quad \text{or} \quad [X'] = \Lambda[X]. \] (32.15)

More generally, any four-component quantity whose entries transform in the same way as \( X \) when we change from one E-inertial coordinate system to another can represent a **4-vector**.

As in 3D, summation over repeated indices (here \( \mu \)) is implied. As with 3D rotations, the entries of \( \Lambda \) are all *constants*, and so may be pushed past derivatives.\(^{20}\) For extra clarity, we will often use coordinate indices \( \mu, \nu, \ldots \) from the middle of the Greek alphabet for one coordinate system, but \( \alpha, \beta, \ldots \) from the start of the alphabet for the alternative coordinate system. Notice a typographic convention: The second index \( \mu \) on \( \Lambda \) appears to the right of the \( \alpha \) index. We must keep track of which index labels row (the first one) and which labels column (the second one), even though one of them is written as a superscript and the other as a subscript.

\(^{18}\)Chapter 34 will give lower 4D indices a different meaning. How do we avoid confusion between a vector component index and an exponent? Sadly, sometimes even experts do get confused. In this book, when a symbol is underscored, that’s a visual cue that a superscript suffix denotes a component (and also that an exponent would not make sense).

\(^{19}\)Unlike 3D, when we introduce 4-tensors of higher rank we will use the same underscore for all of them.

\(^{20}\)Also as in 3D, \( \Lambda \) has no underscore because it’s not a tensor: Instead of having a tensorial transformation rule under change of coordinates, it specifies a change of coordinates.
It’s convenient to introduce an abbreviation: The **metric** $g_{\mu\nu}$ is the matrix of constants\(^{21}\)

$$ [g] = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$  \hspace{1cm} (32.16)

Instead of the condition for a rotation, restrict to those matrices $\Lambda$ with the property that

$$[\Lambda^t \Lambda] = [g], \quad \text{defining property of Lorentz transformation} \quad (32.17)$$

This property is different from Equation 32.4 (page 490), because $[g]$ is not the identity matrix. Nevertheless, it encompasses all the operations we have previously studied piecemeal:

**Your Turn 32D**

Confirm that the Lorentz transformations found in Chapter 30 obey Equation 32.17, for example,

- **Boost along $\hat{x}$**: $[\Lambda] = \begin{bmatrix} \gamma & -\gamma \beta \\ -\gamma \beta & \gamma \end{bmatrix}$. **Rotation about $\hat{z}$**: $[\Lambda] = \begin{bmatrix} 1 & \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$.  \hspace{1cm} (32.18)

Later, we’ll connect our original method of discovering Lorentz transformations to Equation 32.17 (Section 34.2.3).

More generally, any rotation matrix $S$ will give rise to a Lorentz transformation with the special form

$$[\Lambda] = \begin{bmatrix} 1 & \vec{0} \\ \vec{0} & S \end{bmatrix}. \quad (32.19)$$

**Your Turn 32E**

For future use, check that Equation 32.17 implies the useful identity

$$\Lambda^t = [g] \Lambda^{-1} [g]. \quad (32.20)$$

**[Hint: First notice that $[g]^2 = \mathbb{1}$.]**

Lorentz transformations close into a group, sometimes called O(3,1):

---

\(^{21}\) $[g]$ is the same set of numerical values in any inertial coordinate system. It may not be obvious that $g$ defined in this way is a 4-tensor, but Section 34.3.3 will show that it is true, just as in three dimensions the collection of nine constants $\delta_{ij}$ is a “tensor from Heaven” (Section 14.2.3).
Section 32.3.1 pointed out that every orthogonal matrix has determinant $+1$ or $-1$. Because this is a discrete choice, the orthogonal group $O(3)$ can be divided into two separate, 3-parameter subsets. One of those—the one containing the identity matrix—is the subgroup $SO(3)$ of rotations. In fact, $SO(3)$ has exactly two connected components; no further discrete reduction is possible.

Similarly, we may take the determinant of both sides of Equation 32.17 to find that every Lorentz transformation must have determinant $+1$ (proper) or $-1$ (improper), and the proper transformations form a subgroup, sometimes called $SO(3,1)$.

Unlike the case for rotations, however, a further reduction is possible: There is no succession of small boosts and rotations that can completely reverse the $t$ axis. To see this, consider the 00 component of Equation 32.17:

$$g_{00} = -1 = \Lambda_{00} + \Lambda_{0}^{\nu}\Lambda_{0}^{\nu}$$

so

$$(\Lambda_{0}^{0})^2 = 1 + \sum_{i=1}^{3}(\Lambda_{0}^{i})^2 \geq 1.$$ 

Hence, $|\Lambda_{0}^{0}| \geq 1$. That is, the Lorentz group splits into orthochronous (time-sense preserving) transformations with $\Lambda_{0}^{0} \geq 1$, and the rest (time-sense inverting) with $\Lambda_{0}^{0} \leq -1$. In fact, the orthochronous transformations form a subgroup, which we’ll call $O^{+}(3,1)$. It in turn contains the still smaller subgroup of transformations that are both proper and orthochronous, sometimes called restricted Lorentz or $SO^{+}(3,1)$.

Any Lorentz transformation can be generated from matrix products of the boosts and rotations (plus reflections). This should not be too surprising: Equation 32.17 sets two symmetric $4 \times 4$ matrices equal, so it’s ten independent constraints on the sixteen numbers $[\Lambda]$ (because $4(4 + 1)/2 = 10$). So there is a six-parameter family of solutions (because $16 - 10 = 6$). That corresponds to our expectations for rotation and boost (three Euler angles plus three velocity components).

Because dissipative processes give statistical physics a definite arrow of time, we do not expect invariance in general under non-orthochronous transformations. Accordingly we will suppose that spacetime is endowed with a metric and an overall time direction. Unlike in 3D, however, will not designate any coordinate system as “right-handed”; we’ll investigate symmetry under the full orthochronous Lorentz group.

Section 32.6.2′ (page 507) uses ideas in this section to make precise the notion of an arrow of time.

32.6.3 The invariant interval

Lorentz transformations are nearly as simple as the 3D transformations in Section 32.3.2. For example, Section 31.3.1 found a quantity that is related to time but is unchanged by Lorentz transformation. Consider a particle trajectory as a curve in spacetime. For any
two nearby points on that curve, the invariant interval (Equation 30.7, page 457) can be rewritten as
\[ c^2 \Delta \tau^2 = -(\Delta X)_{\mu} g_{\mu\nu} (\Delta X)^\nu. \]  
(32.21)

To show that the invariant interval really is form-invariant under Lorentz transformations, write
\[ (c\Delta \tau)^2 = -(\Delta X)_{\mu} \Lambda_{\mu}^{\nu} \Lambda_{\nu}^{\lambda} (\Delta X)^\lambda = -(\Delta X)^{\mu} g^\mu_{\nu} (\Delta X)^\nu = (c\Delta \tau)^2. \]  
(32.22)

Analogously to 3D, a single quantity that is invariant under Lorentz transformations (for example, \( \Delta \tau \)) is called a 4-scalar.

The invariant interval has dimensions \( L^2 \). \( \Delta \tau \) equals the time that elapses between two events in an E-inertial coordinate system in which both occur at the same position \( \vec{r} = \vec{0} \). Chapter 31 called its integral along a trajectory the proper time, which is apt23 because that coordinate system would also be the rest frame of an observer who runs at constant velocity from one event to the other and carries a clock to measure the time between the two events.

32.6.4 4D contraction

The idea of invariant interval is so useful that we generalize it. If \( \underline{Y} \) is any 4-vector (not necessarily a displacement in spacetime), we might be tempted to form a single number via \( \sum_{\mu} Y_{\mu} Y^{\mu} \), but that’s not invariant: The 3D derivation in Section 32.3.1 fails because not all Lorentz matrices are orthogonal. Instead, define the notation \( ||\underline{Y}||^2 \) by the formula\(^24\)
\[ ||\underline{Y}||^2 = Y_{\mu} g_{\mu\nu} Y^{\nu}. \]  
(32.23)

This quantity really does equal \( (\underline{Y'})^\mu g_{\mu\nu} (\underline{Y'})^\nu \), so it’s a 4-scalar. (The proof is the same as in Equation 32.22.) In the special case of a spacetime displacement, setting \( \underline{Y} = \Delta \underline{X} \) gives that \( (c\Delta \tau)^2 = ||\Delta \underline{X}||^2 \).

Similarly, for any two 4-vectors the invariant inner product is defined as \( Y_{\mu} \Lambda_{\mu}^{\nu} Z^{\nu} \). It’s also a 4-scalar, analogous to the dot product in 3D, so we sometimes abbreviate it \( \underline{Y} \cdot \underline{Z} \).

A big difference with ordinary geometry, however, is that we can have \( ||\underline{Y}||^2 = 0 \) even if \( \underline{Y} \) itself is not zero.\(^25\) A 4-vector with this property is called lightlike, because any two points on a light ray’s trajectory have such a separation.\(^26\) More generally, if \( \underline{\Delta X} \) is the spacetime separation between two events, then we call the three cases \( ||\Delta \underline{X}||^2 < 0, = 0, \) and \( > 0 \) by the names timelike, lightlike, and spacelike separation, respectively. A material particle always moves slower than \( c \), so it will always move to a new spacetime point that is separated by a timelike displacement vector from its original point.

\(^{22}\)This step is analogous to Equation 32.2. This analogy is the reason that \( g \) is again called the “metric.”
\(^{23}\)In some languages, “propre” can mean “one’s own.”
\(^{24}\)Note that the notation \( ||\vec{r}||^2 \) denotes the ordinary length-squared of a 3-vector, whereas \( ||\underline{Y}||^2 \) denotes the 4D invariant product of \( \underline{Y} \) with itself.
\(^{25}\)More precisely, this cannot happen for a real 3-vector. The helicity basis vectors do have the property that \( \hat{\epsilon}^{(\pm)} \cdot \hat{\epsilon}^{(\pm)} = 0 \) (Section 18.10.2, page 285).
\(^{26}\)Some books use the synonym “null” for lightlike.
Figure 32.1: Classes of 4-vectors. Events Q and S are lightlike separated from event P: All inertial coordinate systems agree that they lie on P’s forward and backward light cones, respectively (solid and dotted red lines). Events T and U are timelike separated from P: All inertial coordinate systems agree that they lie inside the forward and backward light cones, respectively. Event R is spacelike separated from P: All inertial coordinate systems agree that it lies outside both light cones.

The locus of all events that are lightlike-separated from P is called P’s light cone. It is three-dimensional, but on a diagram with one space dimension suppressed it will look like a cone. (If two space dimensions are suppressed, the light cone looks like two crossed lines.) The part with \( t < t_P \) is the “past light cone of P”; the other part is the “future light cone of P.”

With this terminology, we can partially clear up an ambiguity concerning the sign of \( \Delta \tau \) in Equation 32.21: If two events are timelike separated, we may take the \( \Delta \tau \) from the earlier to the later to be the positive square root of \( \Delta \tau^2 \). (If they are lightlike separated, then \( \Delta \tau = 0 \) is already unambiguous. If they are spacelike separated, then \( \Delta \tau \) is both ambiguous in sign and pure imaginary.)

32.6.5 The four-velocity invariantly characterizes a particle trajectory

We can describe the trajectory of a material particle as a parametric curve in spacetime (a chain of events) by using proper time as the parameter:\(^{28}\) \( \vec{X} = \Gamma(\tau) \). Because the invariant interval is a 4-scalar (Equation 32.22), the operation \( \frac{d}{d\tau} \) does not alter the transformation properties of whatever it hits. Thus, the four functions

\[
U^\mu(\tau) = \frac{d\Gamma^\mu}{d\tau}
\]

also form a 4-vector, called the trajectory’s 4-velocity at whatever point we evaluate the derivative. One way to evaluate it is to write the curve with an arbitrary parameter \( \xi \), then compute \( \vec{U} = (d\Gamma^\mu/d\xi)/(d\tau/d\xi) \). Our sign choice for \( \tau \) implies that \( U^0 > 0 \).

---

\(^{27}\) Section 34.7c will point out that this sign choice complicates the notion of time reversal invariance.

\(^{28}\) Recall Section 31.3.1 (page 477). We can’t use this strategy for the trajectory of a light pulse, because \( d\tau \equiv 0 \) everywhere along a lightlike curve, so Section 30.6.1 used a different parameterization.
Your Turn 32G

a. Show that the 4-velocity always obeys the identity

\[ \|U(\tau)\|^2 = -c^2. \]  \hspace{1cm} (32.25)

b. Construct a 3D analogy: A curve may be parameterized by arc length s. Then \( \frac{d\vec{T}}{ds} \) is a unit 3-vector defined along the curve (its unit tangent vector field).

Here is an example: Consider a particle in uniform straight-line motion with speed \( v = \beta c \) directed along \( \hat{x} \):

\[ \Gamma(\xi) = \begin{bmatrix} \xi \\ \beta \xi \\ 0 \\ 0 \end{bmatrix}; \quad \frac{d}{d\xi} \Gamma = \begin{bmatrix} 1 \\ \beta \\ 0 \\ 0 \end{bmatrix}. \]

Equation 32.21 gives

\[ U = c^{-1}\sqrt{1 - \beta^2} \frac{d\xi}{d\tau} = (c\gamma)^{-1} \frac{d\xi}{d\tau}, \]

where \( \gamma = (1 - \beta^2)^{-1/2} \), and so

\[ \[U\] = (\frac{d\Gamma}{d\xi})/(d\tau/d\xi) = \begin{bmatrix} \frac{\gamma c}{\gamma} \\ \frac{\gamma^2 c}{\gamma} \\ 0 \\ 0 \end{bmatrix}. \]  \hspace{1cm} (32.26)

Your Turn 32H

Confirm that Equation 32.25 holds, starting from Equation 32.26.

32.6.6 Payoff: the 4-wavevector and its transformation rule

This material has been pretty abstract. But unlike a lot of subjects, where “in theory it’s easy but not in practice,” in this case it’s the other way round! For many purposes, all you need to remember is

The location \( X \) of an event has components \( X^\mu \) with upper index, and hence so does its derivative \( U \). The constant matrix \( g \) as we have used it so far has two lower indices. Keep calm and only contract upper with lower indices. If you feel an urge to contract upper with upper, you may be missing a \( g \) matrix.

(32.27)

For example, if you forget the \( \frac{\gamma}{\gamma} \) factor in Equation 32.21, the rule (32.27) will quickly alert you.

Here is another example. When we discussed plane waves in Chapter 30, we found ourselves manipulating the phase expression \(-\omega t + \vec{k} \cdot \vec{r}\). Notice that this expression can be compactly written as \( k^\mu \eta_{\mu\nu} X^\nu \), where the 4-wavevector is defined as

\[ k^\mu = \begin{bmatrix} \omega/c \\ \vec{k} \end{bmatrix} = \begin{bmatrix} \omega/c \\ \vec{k}_x \\ \vec{k}_y \\ \vec{k}_z \end{bmatrix}. \]  \hspace{1cm} (32.28)

---

29We previously obtained this in Equation 31.11 (page 479).
The virtue of this reformulation is that it tells how \( k \) must transform. The invariance of 4D contraction says that

\[
k^{\alpha} g_{\alpha \beta} \bar{X}^{\beta} = k^{\mu} g_{\mu \nu} \bar{X}^{\nu}, \quad \text{where } k^{\alpha} = \Lambda^\alpha_\mu k^\mu. \tag{32.29}
\]

The phase of a wave is physical; for example, it determines the planes on which the electric and magnetic fields equal zero, and so on. Hence the phase should be the same when expressed in any coordinate system. Equation 32.29 says that indeed, the phase function remains linear in \( \bar{X}' \) if \( k \) transforms as a 4-vector.

**Your Turn 32I**

a. Show that this compact statement implies our previous low-tech results about the aberration of starlight and both kinds of Doppler shift (Your Turn 30G and Problem 30.6).

b. Also show that for a wave to travel at speed \( c \), we must have \( ||k||^2 = 0 \).

Besides being pretty, that last formula is manifestly Lorentz invariant, as it must be—we designed Lorentz transformations precisely to maintain the speed of light in every inertial coordinate system.

**32.7 MOMENTUM AND ENERGY REVISITED**

With the framework we have developed, we can elegantly restate our earlier proposal for relativistic energy and momentum\(^{30} \) as

\[
p = mU. \quad \text{four-momentum} \tag{32.30}
\]

The mass \( m \) is a 4-scalar, a single number characterizing the particle. Because \( U \) transforms as a 4-vector, and the mass is a 4-scalar, therefore the proposed formula for four-momentum is also a 4-vector. That is, *unlike Newton’s formula, it has a linear transformation law under Lorentz boosts.* \( \underline{p} \) has the same units as the newtonian momentum; indeed, combining Equations 32.26 and 32.30 gives that \( p^0 = mc^2 \) and \( p^i = m\bar{v}^i \). Thus, \( \underline{p} \) is a particle’s relativistic (energy)/c and \( \bar{p} \) is its relativistic momentum.

With this definition, Einstein’s proposed conservation law says

\[
\sum_{\ell} 
\underline{p}_{(\ell, \text{in})} = \sum_{\ell} 
\underline{p}_{(\ell, \text{out})}. \tag{32.31}
\]

Certainly if that formula is true in any one inertial coordinate system, it will take the same form in any other one, by an argument like the one we applied to Equation 32.5: Both

\(^{30}\text{See Section 31.3.1 (page 477). We drop the check over } \underline{p} \text{ because we no longer need to distinguish relativistic and newtonian momentum.}\)
sides transform the same way (as 4-vectors), so Equation 32.31 is Lorentz-invariant at a glance.

In short, the distinction between energy and momentum has now melted away (apart from the constant factor of $c$). They are parts of a single 4-vector, so their conservation laws are not independent facts.

### 32.7.1 Beyond $E = mc^2$

Equations 32.30 and 32.25 imply a relationship between the momentum, energy, and mass of any particle:

$$
||\vec{p}||^2 = -(mc)^2 \quad \text{or} \quad -(\vec{p}^0)^2 + \vec{p}^i \vec{p}_i = -(mc)^2.
$$

Our identifications of $\vec{p}^0$ as a particle’s total $E/c$, and the spatial components $\vec{p}^i$ as its momentum, $\vec{p}_i$, yield the relation

$$
E^2 = (||\vec{p}||c)^2 + (mc^2)^2.
$$

For a particle at rest, this reduces to the famous and dangerous result discussed in Section 31.3.3.

For a particle moving slowly, so that $pc \ll mc^2$, we can apply a Taylor expansion to Equation 32.33 to get $E \approx mc^2 + \frac{\vec{p}^2}{2m} + \cdots$, approximately a constant plus the newtonian formula, recovering Equation 31.12 (page 479).

### 32.7.2 The 4-momentum of a massless particle is a lightlike 4-vector

There is another interesting limiting case. For a particle moving fast, so that $\vec{p}c \gg mc^2$, we recover $E \approx ||\vec{p}||c$ (Equation 31.15, page 481). For the case $m = 0$, this relation is true regardless of the value of momentum:

$$
E = ||\vec{p}||c. \quad \text{massless particle}
$$

### 32.7.3 de Broglie’s prediction for electron wavelength was dictated by Lorentz invariance

Einstein also proposed in 1905 that monochromatic light should have a dual interpretation as a stream of particles, each with energy $E_{\text{photon}} = \hbar \omega$. But neither side of this formula is a 4-vector. Can we reconcile this law with Lorentz invariance?

Notice that the Einstein relation can be written as $p^0_{\text{photon}} = \hbar \xi$. In order for it to be part of a Lorentz-invariant physical theory, then, there must be a bigger law:

$$
p^0_{\text{photon}} = \hbar \xi. \quad \text{Einstein/de Broglie relations}
$$

31Section 55.4 will rederive this relation for photons by quantizing the field.
Perhaps this is no surprise: We already knew that light has a wavevector \( \mathbf{k} \), and that the Maxwell equations require \( ||\mathbf{k}||^2 = 0 \). Hence Equation 32.35 implies \( ||p||^2 = 0 \) as well. But that is just a concise version of Equation 32.34.

So when L. de Broglie proposed that electrons and other “material” particles also had a dual nature,\(^{32} \) he didn’t need to look hard for a Lorentz-invariant rule describing the wave: \emph{Equation 32.35 is the only suitable proposal.} Indeed, after substituting the electron mass into Equation 32.33, then Equation 32.35 gives an experimentally falsifiable prediction for the relation between electron energy and wavelength, later confirmed by electron diffraction experiments. de Broglie’s insight is all the more impressive because at that time, there was no known candidate for a relativistic wave equation for electrons. It’s another example of “Einstein thinking.”\(^{33} \)

### 32.7.4 Particle creation and destruction

Prior to 1897, those scientists who believed in the atomic theory of matter (by no means everyone) had a vision in which everything was constructed from about a hundred species of little, hard marbles that had not been created nor destroyed, only rearranged, since the Beginning. The birth of atomic and then nuclear physics shook that edifice to its foundations, only to replace it by something rather similar: Atoms had constituents (electrons and nuclei), and even the nuclei themselves had constituents (protons and eventually neutrons), but those particles were deemed to be little, hard marbles that had not been created nor destroyed, only rearranged, since the Beginning.

Just as Einstein had found no scientific necessity for the masses of atomic nuclei to be unchanged in a collision, however, so too there proved to be no reason why their numbers and types should not change. If the incoming participants in a collision have sufficient energy, then more participants can exit than entered, created from nothing but that energy.\(^{34} \) The barrier is especially small for creation of massless particles. Indeed, everybody knew that an excited hydrogen atom can give off light without ceasing to be a hydrogen atom, but initially that process had seemed difficult to imagine from a light-particle point of view. The idea of particle creation from energy solved that puzzle, and the much more perplexing puzzle of where the electrons emitted in nuclear beta decay were located prior to the reaction.\(^{35} \) Later, as particle accelerators became more powerful, particle creation from energy was observed even for massive particles, first electrons and then everything else. Even without constructing an accelerator, a whole shower of cosmic rays can be created in the upper atmosphere from a single energetic incoming proton.

Conversely, an electron and positron can mutually \emph{annihilate}, the key process underlying positron emission tomography (PET). The energy equivalent of their combined

---

\(^{32} \) See also Section 21.7 (page 331).

\(^{33} \) See Section 31.3.1 (page 477).

\(^{34} \) Chapter 55 will develop a framework to describe this phenomenon.

\(^{35} \) Confinement of a preexisting electron in a nucleus would violate the Uncertainty Relation. Enrico Fermi broke this impasse in 1933, proposing that the electron or positron does not exist prior to emission from the nucleus. (His article was also the first to use quantized spin-1/2 fields in particle physics, predating Heisenberg by several months.)
masses emerges as light.

**FURTHER READING**

Note that some authors use a different convention that takes $[g]$ to be *minus* the matrix in Equation 32.15. This convention leads to correct results if it is applied consistently. Be sure you know which convention is in force before you take formulas from a book or article.

An older tradition, now deprecated, introduces an *imaginary* quantity equal to $ict$ and calls it $X_4$. This desperate, unphysical attempt to make the metric look euclidean leads to endless confusion with quantum mechanics, where complex variables enter in an unrelated way. (It also must be unlearned when it’s time to move onward to general relativity.)

*Technical:
Historic: de Broglie, 1923a; De Broglie, 1923; de Broglie, 1923b.*
More about field transformation

Let us be more explicit about the field transformation rule, by stating where derivatives are evaluated. If $\phi$ is a function of $\vec{r}$, then:

$$\vec{V}'_\alpha \phi|_{\vec{r}'} = S_{\alpha\mu} \vec{V}_\mu \phi|_{S^{-1}(\vec{r}')}. $$

To confirm that this formula makes sense, including that tricky matrix inverse, consider two such transformations in succession:

$$\vec{V}''_\mu \phi|_{\vec{r}''} = S^{(2)}_{\mu\nu} \vec{V}'_\nu \phi|_{S^{(2)-1}(\vec{r}'')} = S^{(2)}_{\mu\nu} S^{(1)}_{\nu\rho} \vec{V}_\rho \phi|_{S^{(1)-1}(\vec{r}'')} = (S^{(2)} S^{(1)})_{\mu\nu} \vec{V}_\nu \phi|_{S^{(1)-1}(\vec{r}'')} = (S^{(2)} S^{(1)})_{\mu\nu} \vec{V}_\nu \phi|_{S^{(1)-1}(\vec{r}'')}.$$ 

The last expression is just what we would have obtained had we applied $S^{(2)} S^{(1)}$ directly to the field. Similar results hold for 3-tensor fields of any rank.

Signs of the times

Section 32.6.2 (page 497) said that a choice for the direction of time is physical information, unlike the choice of a right hand, which is purely conventional in classical physics. To make this more precise, we can choose one set of orthonormal basis 4-vectors for spacetime, then declare that it and every other one obtained from it by orthochronous Lorentz transformations are all “forward directed.” Any of those may then be used to fix the sign of proper time along a timelike trajectory.
32.1  *Time for the stars*

Suppose that you receive an invitation to a birthday party on a planet of a distant star. The star is located along the $X^1$ axis of an inertial coordinate system $X^\mu$ in which Earth is at rest.

You get in your spaceship and accelerate along the $\hat{x}$ direction. Your trajectory is a curve in spacetime. Take a minute to sketch how you think this curve should look in the $x(ct)$ plane (and also the trajectory corresponding to the friends and loved ones you left at home.)

Your trajectory can be written in parametric form: $\Gamma(\tau)$, where $\tau$ is the time you perceive on the ship. Section 32.6.5 (page 501) defined four-velocity as $U = \frac{d\Gamma}{d\tau}$. It will be convenient to define the dimensionless variable $w = \frac{U^1}{c}$ and substitute $cw$ for $U^1$. Equation 32.25 gave a relation that lets us express $U^0$ in terms of $w$.

To travel without too much discomfort, you adjust the rockets so that you feel pushed against the rear wall of your ship with a constant force equal to 1.5 times your normal Earth weight. In first-year physics, we might express that condition as

$$\frac{dx}{dt} = a_0 = 1.5(9.8 \text{ m/s}^2),$$

that is, one (differential) equation in one unknown function. But after a while your velocity may become comparable to light, so we need to find how to modify this equation before solving it. Specifically, take the following steps to translate the requirement of constant artificial gravity into a differential equation for $dU/d\tau$.

Consider one moment $\tau_*$ along your journey. There is an inertial coordinate system $X^{*\alpha}$ in which you are momentarily at rest at $\tau_*$. This is the system obtained by boosting the unprimed system by $\beta_*c$ where

$$\beta_* = \frac{U^1(\tau_*)}{U^0(\tau_*)}.$$

In it, your velocity at $\tau_*$ equals zero, and hence your velocity near $\tau_*$ is increasing from slightly negative to slightly positive.

Even if we don’t know the relativistic modification of Newton’s law, we do know that physics should reduce to newtonian form when things are moving slowly. So we know that the acceleration at $\tau_*$, measured in the primed coordinate system, should equal your weight on Earth, times 1.5, divided by your mass. Thus, we demand of the trajectory that

$$\frac{d}{dt'} \left[ \frac{U'^1}{c^2 U'^0} \right]_{\tau_*} = a_0.$$

(32.36)

Now apply the Relativity Strategy, that is, translate Equation 32.36 to the Earth-bound inertial coordinate system. Remember that (i) the Lorentz boost connecting the primed

---

36 Of course, you won’t see the Sun rise and set, but you could measure $\tau$ by the growth of your fingernails, or the number of heartbeats, or a clock you carry with you.

37 Don’t worry about how the ship is propelled, fuel requirements, and so on!

and unprimed systems depends on \( \tau_u \), but not on \( \tau \) (it’s not an accelerating system). (ii) Factors like \( U^{\mu} \) that do depend on \( \tau \) may be evaluated at \( \tau_u \), but only after the time derivatives (if any) have been evaluated.

a. Equation 32.36 appears to be one differential equation in two unknown functions, but remember that the components of 4-velocity obey the constraint Equation 32.25 (page 502). Use that constraint to express Equation 32.36 in terms of just one unknown function \( \omega(\tau) \) and its derivative(s). Specifically, show that

\[
\frac{d\omega}{d\tau} \bigg|_u = \frac{a_0}{c} \sqrt{1 + \omega^2}.
\]  

(b) Equation 32.36, and hence also 32.37, must hold at every \( \tau_u \) along the acceleration part of the trip; that is, it is a differential equation. Solve it for \( \omega(\tau) \) with appropriate initial condition.

c. Integrate your answer to (b) to find the actual trajectory \( \Gamma(\tau) \).

Of course, you don’t want to arrive at your destination and crash into it! You must also decelerate prior to arrival. So after proper time \( \tau_{rev} \), you reverse the engines and accelerate along the \(-\hat{x}\) direction, again maintaining a constant force of 1.5 times your normal Earth weight (this time from the front wall of the spaceship), until you come to rest.

d. Revise your sketch to show the entire journey.

e. Suppose that your total elapsed proper time is \( 2\tau_{rev} = 1 \) year. Find the total distance \( \Delta X^1 \) you’ve traveled from Earth after carrying out both steps of the outbound journey. Express your answers in light years.

f. You spend a couple of hours at the party, then reverse your trip to come home. Thus, upon your return you have aged two years. How much have your friends aged since you last saw them?

g. Convinced that Earth will soon be rendered uninhabitable by its inhabitants, you organize expeditions to scout other planets, then return home and report. Each spaceship takes a trip like the one above, but this time the round-trip duration is such that the crew ages by 30 years (not 2 years). How big a chunk of our galaxy can you explore in this way? When should those who stayed behind expect the scouts to return home?

h. Following (g), take the total distance \( \Delta X^1 \) to the destination and divide by 15 years, obtaining a quantity with dimensions of speed. Make an Insightful Comment about your answer, then find and calculate some other, more meaningful, quantity with the same dimensions.

[Remark: Atomic clocks are so accurate that the “twin paradox” behavior you found in this problem can be measured even for clocks carried over terrestrial distances at ordinary speeds.]
The logically simple does not, of course, have to be physically true; but the physically true is logically simple, that is, it has unity at the foundation.

— Albert Einstein

33.1 FRAMING: SYMMETRIES AS DRIVERS

Prior to Einstein, physicists thought of Physics as a search for the right equations of motion. When they attempted to marry the mechanics of charged particles with electromagnetic fields, they got bogged down. Einstein and his successors realized that symmetries of Nature should be the primary drivers for model building; once the right symmetry principle was found, dynamics could then follow along.

Electromagnetic phenomenon: The orbital period of a charged particle in uniform magnetic field starts to depend on its energy, when that energy is increased beyond $mc^2$.

Physical idea: The Lorentz force law must be reinterpreted as giving the rate of change of relativistic, not newtonian, momentum.

33.2 4-TENSORS

33.2.1 Tensor products of Lorentz transformations

Based on our experience in 3D, we now generalize 4-vectors: Any quantity with 4 components that transform analogously to Equation 32.8 (page 492) as we change between E-inertial coordinate systems on spacetime can represent a 4-tensor of rank\(^1\) \(\left(\frac{1}{6}\right)\). For example, if

\[
F^{\tau\alpha\beta} = N_\mu N^\nu F_{\mu
u},
\]  

then \(F\) has rank \(\left(\frac{1}{6}\right)\).

Similarly to three dimensions, we will see how a 4-tensor can be thought of geometrically as specifying a linear 4-vector-valued function of another 4-vector; or a scalar function that is linear in each of two 4-vectors; and so on.\(^2\)

---

\(^1\)Chapter 34 will justify the elaborate notation by extending the definition to rank \(\left(\frac{1}{6}\right)\).

\(^2\)Recall Section 13.3. This viewpoint works even on curved spacetimes that have no E-inertial coordinates.
33.2.2 An extended Tensor Principle

The preceding discussion suggests an upgraded Tensor Principle: 3

\[ \text{Physical quantities all arrange themselves into 4-tensors (or 4-tensor fields), in some cases constrained by symmetry or antisymmetry. Physical laws are Lorentz invariant, and moreover can be written in manifestly invariant form by exploiting simple Rules about tensors.} \]

If we restrict to rotations only, 4 then every 4-tensor falls into blocks that are themselves 3-tensors; thus, Idea 33.2 includes and extends our earlier 3D principle.

So far, our evidence in favor of Idea 33.2 is that indeed some quantities obey it:

- The mass \( m \) of a point particle is a single quantity intrinsic to an object—a 4-scalar, often abbreviated “scalar.” (Such a quantity could also be called a “4-tensor of rank \((0,0)\),” because it has no indices of any type.)
- The speed of light \( c \) is a single, Lorentz-invariant constant of Nature—also a 4-scalar.
- The interval \( ||\Delta X||^2 \) between neighboring events is a 4-scalar as well.
- The time and location of an event have been fused into \( X \), which we have called a 4-vector. We can also call it a 4-tensor of rank \((1,0)\) because its component representation has one index in the upper position (and none in the lower position).
- The energy and momentum of a point particle have been fused into \( p \), which we saw transforms the same way as \( X \) and hence is also a 4-vector. 6
- The angular frequency and wavevector of a plane wave have been fused into \( k \), which again is a 4-vector. 6

The next section will explore whether the electric and magnetic fields also follow the Tensor Principle. First, let’s review how some laws of Nature can usefully be written as relations among 4-vectors. Here are some examples:

**Wave equation**

For the special case of plane waves, you showed in Your Turn 32I (page 503) that the wave equation boils down to \( ||k||^2 = 0 \), a condition on \( k \). (Section 34.2.3 will return to the wave equation itself.)

**Momentum conservation**

Chapter 31 gave us a taste of “Einstein thinking” as a chain of logic:

---

3 Compare Section 32.5 (page 495). Quantum mechanics amends this claim slightly to allow an additional class of quantities called “4-spinors” (Section 34.11, page 545), but with the proviso that spinor fields are not directly observable.

4 See Equation 32.19 (page 498).

5 See Equation 32.30 (page 503).

6 See Equation 32.29 (page 503).
Chapter 33  The Faraday Tensor

1. We still expect four conservation laws, even if they’re not exactly Newton’s.

2. What could they be? Instead of trying to tinker with Newton’s formulas, start from scratch. The statement that a four-vector quantity is the same before and after a collision is an invariant statement.

3. What could that four-vector be? Newton says that both energy and momentum are proportional to a constant, \( m \), intrinsic to the body in question. The expression \( p = m U \) is a four-vector related to velocity with appropriate units, and hence so is \( \sum \epsilon P(\epsilon) \).

4. The statement \( \overrightarrow{p_{\text{tot,\text{in}}} - p_{\text{tot,\text{out}}} = 0} \) is therefore a candidate law.

**Einstein/de Broglie relations**

These physical laws also took the compact form \((\text{4-vector}) - (\text{4-vector}) = 0\) (Section 32.7.3, page 504).

**Next steps**

“Einstein thinking” proved to be powerful, and quickly came to dominate in the search for other new laws. Next, we’ll see that applying it to particle trajectories leads to a new kind of object, a 4-tensor of rank higher than one.

### 33.3 LORENTZ FORCE LAW

#### 33.3.1 The Faraday tensor unifies electric and magnetic force laws

Let’s abstract some structural features of the Lorentz force law, try to guess a reformulation in terms of 4-tensors, and then compare to the pre-Einstein version. It has the general structure: \( \frac{\text{d}p}{\text{d}\tau} = qF(U(\tau)) \), where \( q \) is a constant of proportionality intrinsic to a test body.

We can write a specific formula of this sort involving 4-tensors:

\[
\frac{\text{d}p}{\text{d}\tau} = qF(U(\tau)). \quad \text{Lorentz force law, 4-vector} \tag{33.3}
\]

On the left side, \( \tau \) is proper time, so \( \frac{\text{d}}{\text{d}\tau} \) preserves the 4-vector character of \( p \), a gambit we already used in Equation 32.24 (page 501).\(^8\) On the right, \( q \) is a 4-scalar constant. \( F \) is a linear function that takes a 4-vector and returns a 4-vector. In three dimensions, such machines are specified by tensors of rank two: For example,\(^9\) the anisotropic spring system

---

\(^7\)Equations 0.5, page 3 and 15.1 page 222.

\(^8\) More precisely, \( \tau \) is forward-directed proper time, and so \( p \) and \( U \) are odd under time reversing Lorentz transformations (Section 34.7c, page 543).

\(^9\) Other examples we studied included viscous drag, the moment of inertia, and electric polarizability, which are 3-tensors defining linear, vector-valued functions of 3-vectors (Section 13.3.1, page 196).
studied in Section 32.3.3 had a restoring force given by \(-\vec{r} \cdot \vec{F}\). Similarly, a 4-tensor of rank \(\delta\) can be used to specify a linear function via

\[ U \rightarrow F(U) \]  

where \(F(U)^\mu = F^\mu_\nu g^\nu_\alpha U^\alpha\).  

(33.4)

We will call \(F^\mu_\nu\) the **Faraday tensor**. More precisely, the components \(F^\mu_\nu\) are a collection of functions of space and time, which are to be evaluated along the particle’s trajectory in Equation 33.3. That is, \(F\) is a 4-tensor **field of rank** \(\delta\).

**Your Turn 33A**

Show that including the \(g\) factor in Equation 33.4 guarantees that \(F(U)\) is a 4-vector.  

[**Hint:** Adapt the derivation that led from Equation 32.17 to 32.22.]

Hence, multiplying \(F(U)\) by the 4-scalar \(q\) and setting the result equal to the 4-vector \(dp/d\tau\) constructs an equation of motion (Equation 33.3) of the form (4-vector) – (4-vector) = 0, which is invariant.

The result just obtained, and the method used, are so important that we should pause to rederive them more explicitly.

**Ex.**

Show directly that Equation 33.3 is Lorentz invariant.  

**Solution:** We adapt the same argument used in Section 32.3.3 (page 491) for rotation invariance of the anisotropic harmonic oscillator. Multiply both sides of Equation 33.3 from the left by a Lorentz transform matrix:

\[ \Lambda \frac{d}{d\tau} [p] = q[\Lambda] \frac{d}{d\tau} [F] \frac{d}{d\tau} [U]. \]

The left side is \(d[p]/d\tau\). The right side looks bad, because \(\Lambda\) will not in general commute with the matrix of \(F\)’s components. To make progress, now insert the identity matrix \(\Lambda^{-1}\Lambda\) just as with the harmonic oscillator. This time, however, we must also insert another identity matrix, \([g]\)²:

\[ \frac{d}{d\tau} [F'] = q[\Lambda] [g] \Lambda^{-1} \Lambda [U] = q\left(\Lambda [F] [g] \Lambda^{-1} [g]\right) \left([g] [U']\right). \]

This new version has the same form as the original, albeit with a modified Faraday tensor:

\[ [F'] = \Lambda [F] [g] \Lambda^{-1} [g]. \]

But Equation 32.20 (page 498) lets us restate this as \([F'] = \Lambda [F]\) or

\[ F'^{\alpha\beta} = \Lambda^\alpha_\nu \Lambda^\beta_\mu F^{\mu\nu}, \]  

(33.5)

which is precisely the transformation rule Equation 33.1. Thus we find that expressing Equation 33.3 in terms of Lorentz-transformed coordinates indeed yields an equation of the same form,

\[ \frac{d}{d\tau} [\dot{p}] = qF'^{\alpha\beta} g^\beta_\nu U^\nu. \]
if we acknowledge that the components of $F$ transform in the way appropriate for a tensor of rank $(\mathcal{S})$.

### 33.3.2 The Faraday tensor is antisymmetric

At first, Equation 33.3 may not seem promising as a reformulation of the Lorentz force law. We wanted a 4-tensor to accommodate the electric and magnetic fields, which have a total of six components, but the object $F$ appearing in Equation 33.3 seems to have $4 \times 4 = 16$ entries!

To make progress, note that $F$ is not entirely free. Equation 33.3 says that it specifies the rate of change in $U$, but $U$ cannot change in an arbitrary way: Section 32.6.5 pointed out that always $||U||^2 = -c^2$, a constant. Thus, the derivative of this quantity is fixed to be zero:10

$$\frac{d}{dt} (U^\mu g_{\mu\nu} U^\nu) = 0.$$ Using the product rule gives

$$2U^\mu g_{\mu\nu} \left(\frac{dU^\nu}{d\tau}\right) = 0.$$ Equations 33.3 and 33.4 then imply

$$(U^\mu g_{\mu\nu})F^{\nu\lambda}(g_{\lambda\xi}U^\xi) = 0 \quad \text{for any } U.$$ That is, $F$ must always give us zero when contracted on each of its indices with the same thing. To guarantee that, we must demand that $F$ be an antisymmetric 4-tensor of rank $(\mathcal{S})$. This extra condition is itself Lorentz-invariant.12

An antisymmetric $4 \times 4$ matrix has just six independent entries—just what we need to contain the electric and magnetic fields.

### 33.3.3 Relate to traditional form

We can give those six entries any names we like. Here are some suggestive names for them:

$$F^{\mu\nu} = \begin{pmatrix} 0 & \frac{E^l}{c} \mu\nu \\ -E^l/c \end{pmatrix} = \frac{1}{c} \begin{pmatrix} 0 & E_x & E_y & E_z \\ -E_x & 0 & B_z & -B_y \\ -E_y & -B_z & 0 & B_x \\ -E_z & B_y & -B_x & 0 \end{pmatrix}^{\mu\nu}.$$ Here the magnetic field tensor $\vec{\omega}$ is defined by Equations 15.2 or 15.3 (page 222) and $\vec{B} = c\vec{\omega}$. Equation 33.6 can be summarized by13

$$F^0l = -F^{l0} = E_l/c \quad \text{and} \quad F^{ij} = \epsilon_{ijk} B_k, \quad i, j, k = 1, 2, 3. \quad \text{Faraday tensor}$$ (33.7)

---

10See Equation 32.23 (page 500).
11The logic is the same as when we interpreted the magnetic field as an antisymmetric 3-tensor (Section 15.2, page 222), because that machine eats a particle’s velocity and always yields a force perpendicular to $\vec{v}$.
12The logic is the same as when we argued that the condition that a 3-tensor be antisymmetric is itself rotation-invariant (Section 32.3.4, page 492). Section 34.3.2 will argue that for 4-tensors, it only makes sense to impose this condition on indices that are all in the same position (in this case, both up).
13The identifications in Equation 33.7 are only valid in a right-handed, inertial coordinate system. As mentioned in Section 15.2, that restriction is the drawback to describing Nature using $\vec{B}$. Equation 33.3 (and the first version of Equation 33.6) don’t suffer from this restriction.
With these names, the 1-component of the proposed reformulation of the Lorentz force law (Equation 33.3) says
\[
\frac{d}{dt}p^1 = q\left( P^{10} g_{00} U^0 + P^{12} g_{22} U^2 + P^{13} g_{33} U^3 \right).
\]
Equations 31.11, 32.26, and 33.6 give
\[
\gamma \frac{d}{dt}p^1 = q\left( -\left( E_1/c \right)(-1)(c\gamma) + B_3(1)(c\gamma) - B_2(1)\gamma \vec{v}_2 \right).
\]
Canceling the \( \gamma \) factors shows that this is just the 1-component of the Lorentz force law in its traditional form (Equation 0.5, page 3), modified only by using the relativistic formula for momentum. The other two spatial components work similarly.

In short,

\textit{Equation 33.3 contains the Lorentz force law. The electric and magnetic fields enter as components of an antisymmetric rank-(2 0) tensor via Equations 33.6 or 33.7.}

You’ll investigate the remaining component of Equation 33.3 in Problem 33.1.

\section*{33.3.4 More on the marriage of \( \vec{E} \) and \( \vec{B} \)}

Like any equation of physics, Equation 33.3 is packed with implicit meaning—a framework established in the preceding chapters. Let’s pause to unpack it.

We imagine some apparatus, with coils, charged plates, whatever, that creates conditions in a region of vacuum (possibly time-dependent). We imagine interrogating those conditions by shooting in charged test particles and observing their trajectories in some coordinate system. Equation 33.3 claims that those trajectories are always solutions to a set of ordinary differential equations. More precisely, it claims that we can find:

- a coordinate system \( t, \vec{r} \) independent of what kind of test particles we use, or their initial conditions, or the apparatus,
- two fixed numbers \( m, q \) characterizing each test particle, independent of what apparatus we choose and the initial conditions on the test particle trajectory, and
- six functions \( F^{\mu\nu} \) on spacetime, depending on the apparatus and coordinate choice but independent of the test particle type or initial conditions,

such that every physical trajectory, in every apparatus, for every test particle type, is a solution of Equation 33.3. Although there are many ways to make these choices, there are even more possible apparatuses, trajectories, and test particle types, so the claim has falsifiable content, while at the same time also telling us in principle how to \textit{measure} the Faraday tensor.

What gave us the right to just declare that \( E_0 = \vec{E}_1/c \) and so on? Remember, \textit{names are arbitrary}. We could give all six entries different letters of the alphabet if we wished (as indeed Einstein did). Equation 33.6 just assigns names that clarify the connection to our previous form of the Lorentz force law. What’s important is that we \textit{consistently} use the...
same names everywhere (for example, rename $\vec{E}_1$ as $cF_{01}^e$ both in the Lorentz force law and later in the Maxwell equations).

Note that every entry of the Faraday tensor participates in Equation 33.3 in the same way. The asymmetry that bothered us between electric and magnetic fields (Hanging Question #C) was more a matter of unfortunate language than real physics.

### 33.3.5 On beauty

Alif: Painting brings to life what the mind sees, as a feast for the eyes. Lam: What the eye sees in the world enters the painting to the degree that it serves the mind. Mim: Consequently, beauty is the eye discovering in our world what the mind already knows.

— Orhan Pamuz

Any physicist will tell you that Equation 33.3 is “beautiful.” What is beauty?

The epigraph describes one aspect, which in a scientific context can express the sensation we get when “pure thought” predicts a phenomenon later observed. Beyond this, however, many scientists would say that it’s the combination of surprise and inevitability. We asked for an invariant force law with a particular overall structure, and there was only one reasonable choice for its detailed form. The surprise for many physicists was the required change in the left side of Equation 33.3.

Soon we’ll extend this observation to the Maxwell equations themselves. Those ad hoc-looking features (like the minus sign that’s hard to remember) are just artifacts of awkward traditional notation. In good notation, not only is the Lorentz invariance manifest; also the structure of the equations will turn out to be rigidly dictated, with no ad hoc features.

“Beauty” also can involve getting something for nothing, because physicists are so stingy (we prefer to say “parsimonious”). Without consciously trying, we wrote a formula (Equation 33.3) that is automatically also invariant under spatial inversions! That is, if you observe the world with a left-handed inertial coordinate system, and deduce the six functions $F_{\mu\nu}$, and your friend observes with a right-handed system, and deduces $F^{\tau\alpha\beta}$, then your measured values will be related by Equation 33.1 (page 510). You will both agree that particle motion is described by $mdU/d\tau = qE(U)$, with the same values of $m$ and $q$.

Section 33.3.5’ (page 521) muses more on beauty in physics.

### 33.3.6 Better than beauty: an experimental consequence involving cyclotron motion

We’ve seen that the manifestly-invariant formula Equation 33.3 reduces to the Lorentz force law as we have been using it, with the one key modification that we must use Einstein’s formula for momentum on its left side. But we have drifted far out into Theoryland. Are there any experiments that can ratify the proposed modification?

First note that when a charged particle orbits in a uniform magnetic field (cyclotron motion), then substituting the magnetic force into Newton’s second law predicts that the orbital period will be independent of energy. The corrected form, in contrast, predicts

The constancy of cyclotron frequency breaks down at high velocity.
deviations from this behavior as the particle’s speed approaches $c$. Not only is this effect seen experimentally; it also imposes an important practical limitation on the design of cyclotron accelerators.\footnote{You’ll work out more details in Problem 33.3. After numerous false starts, the relativistic prediction was experimentally verified in 1914 by G. Neumann.}

**Your Turn 33B**

Work out the period of the orbit as a function of energy and comment on the limiting behavior when particle speed is much smaller than $c$.

### 33.4 Transformation of the Faraday Tensor

It is fun to play with tensors, and nice to have beautiful equations. But finding and confirming the right Lorentz force law has additional benefits. Section 32.3.3 found the transformation properties of the spring tensor by requiring rotation invariance of Newton’s law. Similarly, Section 33.3.1 found the transformation properties of the electromagnetic fields by requiring Lorentz invariance of the force law.

We’ll now work out two more classic examples, which have interesting and falsifiable physical consequences.

#### 33.4.1 Electric and magnetic fields mix under Lorentz boosts

Suppose that in one coordinate system $\vec{B}_3 \neq 0$ but all other components of $\vec{E}$ and $\vec{B}$ are zero. Suppose also that the primed coordinate system is moving at speed $βc$ relative to the unprimed one, along $\hat{x}$. Then Equation 33.1 says that the components $[F'_\alpha]$ will be given by the matrix product $M[F]N^t$, or

$$
\begin{bmatrix}
γ & -γβ \\
-γβ & γ \\
1 & 1
\end{bmatrix}
\begin{bmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & B_1 & 0 \\
0 & -B_1 & 0 & 0 \\
0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
γ & -γβ \\
-γβ & γ \\
1 & 1
\end{bmatrix}
= 
\begin{bmatrix}
0 & 0 & -γβ B_3 & 0 \\
0 & 0 & γ B_3 & 0 \\
γ β B_3 & -γ B_3 & 0 & 0 \\
0 & 0 & 0 & 0
\end{bmatrix}.
$$

(33.8)

The final expression is again antisymmetric, as it must be. Comparing to Equation 33.7, we read off the primed fields:

$$
\vec{B}'_3 = γ \vec{B}_3; \quad \vec{E}'_2 = -γ β c \vec{B}_3.
$$

(33.9)

The second of Equations 33.9 illustrates the mixing of electric and magnetic fields upon Lorentz boosts anticipated in the Prologue.\footnote{See Section 0.4.1 (page 12).} Suppose that a wire stretches along the $y$ axis and is set in motion along $x$ in the presence of the magnetic field just described. Mobile charges are confined to a wire, and so are carried with it. Applying the Relativity Strategy, we ask, what is the situation in the wire’s rest frame? Equation 33.9 says there is an electric field in that coordinate system, which in an ohmic material like copper will give rise to current along the wire. Thus when a coil moves relative to a stationary magnet, we...
have the same explanation for the resulting current as when a moving magnet approaches a stationary coil: In both cases, there is an electric field in the rest frame of the coil that drives charges, resolving the paradox in Hanging Question #A (page 13).

To summarize, \( E \) and \( B \) have no separate identities. They are just bits of some bigger, unified object, the Faraday tensor. If \( B \neq 0 \), then the statement that \( E = 0 \) is not Lorentz-invariant; in our example, it was true in our original coordinate system but not in the boosted one.

In his first relativity paper, Einstein somehow managed to find the right transformations in the ugly, mysterious form Equation 33.9, and show that they were exact invariances of the Lorentz force law and the Maxwell equations, all without the benefit of 4-tensor notation. Today, we view them as consequences of the beautifully simple Equation 33.1, re-expressed in the awkward, but traditional, symbols. The reformulation of relativity using tensor methods was initiated by H. Minkowski and developed by many others.

Was it worth the effort? One reply is that most of us would not have been able to see through the algebra to the happy ending had we tried to guess the right transformation law, and prove the invariance, in the old 3D notation. The lucidity we get from 4-tensor notation was also crucial when it was time to invent general relativity, the more elaborate parts of the Standard Model (Dirac and Weyl spinors, Yang–Mills theory) and beyond (supersymmetry...). Even in electrodynamics, we’ll need that clarity in the following chapter to establish the full invariance of the Maxwell equations and later to prove the local conservation of field energy and momentum.

### 33.4.2 A charge in uniform, straight-line motion

Let’s apply what we have learned to find the fields created by a point charge \( q \) moving uniformly relative to the lab with velocity \( c\beta \hat{x} \). Suppose that at time zero the charge is at the origin of lab coordinates. We wish to know the fields at every position and time.

Rather than solve the Maxwell equations with a tricky moving boundary condition, we can apply the Relativity Strategy. First get the fields in the inertial coordinate system that is itself moving at \( c\beta \hat{x} \) with respect to the lab. In this system, the problem is easy: A point charge \( q \) is at rest at the origin. There is no magnetic field, and the electric field is given by Coulomb’s law.

For brevity, let’s restrict to the xy plane and suppress the z direction from our notation.

---

16 There are vistas in Sections 7.5.3, 34.10, 34.11′, and Section 40.4′b (page 616).
17 Idea 33.1 (page 510). Heaviside actually solved this problem without relativity in 1888, but only by a more strenuous effort.
33.4 Transformation of the Faraday Tensor

Your Turn 33C

Apply the appropriate Lorentz transformation to find that at $z = 0$,

$$
\vec{E}_x = \frac{\gamma(x - \beta ct)}{(y^2(x - \beta ct)^2 + y^2)^{3/2}} \frac{q}{4\pi \varepsilon_0}, \quad (33.10)
$$

$$
\vec{E}_y = \frac{\gamma y}{(y^2(x - \beta ct)^2 + y^2)^{3/2}} \frac{q}{4\pi \varepsilon_0}. \quad (33.11)
$$

The results in Your Turn 33C can be obtained directly by solving the Maxwell equations, but instead, you just found them by applying relativity. They are complicated formulas, but note first the ratio

$$
\frac{\vec{E}_x(t, \vec{r})}{\vec{E}_y(t, \vec{r})} = \frac{x - \beta ct}{y} = \frac{x - x_e(t)}{y - y_*} \quad \text{where } \begin{bmatrix} x_e(t) \\ y_* \end{bmatrix} = \begin{bmatrix} \beta ct \\ 0 \end{bmatrix}.
$$

This ratio, along with $\vec{E}_z = 0$, determines the direction that $\vec{E}$ points. It says that at any moment, $\vec{E}$ points along the line of sight from the particle’s position at that time toward the observer (if $q > 0$).

Think about how remarkable that result is. When we look at a distant charge, we are actually seeing it in the past, due to the finite speed of light. And yet, the electric field at the observer is seen to be directed at the particle’s position at the time of observation, even though simultaneity between that point and the observation is relative! The reason this can occur is that the electric field vector from the charge’s retarded position, which is all that the observer can see, gets bent by the Lorentz boost in exactly such a way as to point in the direction from the charge’s current position at the time of observation.

The magnitude of the electric field is also noteworthy:

$$
||\vec{E}|| = r^{-2} \gamma \left(1 + (y^2 - 1) \cos^2 \theta\right)^{-3/2} \frac{q}{4\pi \varepsilon_0}.
$$

This is isotropic when the velocity is small (and thus $\gamma \rightarrow 1$). But for large $\gamma$, it is peaked around $\theta = \pi/2$ (the equatorial plane). In short,

At any time $t$, $\vec{E}(t, \vec{r})$ points radially outward from the particle’s position at that time to the observation point $\vec{r}$. Its magnitude is nonuniform: Field energy gets squashed into the plane transverse to the particle’s velocity. $||\vec{E}||$ also falls off as the inverse square of distance to that position. \( (33.12) \)

Your Turn 33D

Do a similar calculation to find the $\vec{B}$ field, describe it in words, and relate to Figure 33.1.

---

18See Section 39.5.2.

19You’ll display the field graphically in Problem 33.4.
Figure 33.1: Fields in the \( xy \) plane created by a positive point charge in uniform motion at speed \( 0.93c \) along \(+\hat{x}\). Only the half-plane \( \{y > 0\} \) is shown. \textit{Left}: The electric field lines point away from the particle location at the time of observation. \textit{Right}: The magnetic field lines are circles coming out of the page at the points shown (and diving back into the page in the lower half-plane). The transverse density of field lines is proportional to field strength. Both kinds of field lines are bunched toward the plane perpendicular to the velocity. See also Problem 33.4.

33.5 SUMMARY

Table 33.1 shows our constructions so far, highlighting the parallels and contrasts between 3D and 4D concepts.

33.6 PLUS ULTRA

It is hard to overstate the importance of symmetry analysis in physics. \textit{All three} of the physical interactions that today are considered to be both fundamental and accepted (electroweak, strong nuclear, and the general theory of relativity) are relativistic field theories that were invented as offshoots of electromagnetism, starting with proposed extensions of its \textit{invariance} properties. (The same is true of all the speculative theories that may one day supplant the Standard Model.) In each case, appropriate tensor analysis had to be created or generalized to assist in writing a field theory whose symmetry was manifest. Section 33.6' (page 522) gives some more hints about the Standard Model.

FURTHER READING

\textit{Intermediate}:


Historical, on cyclotron motion: Cushing, 1981.
Table 33.1: Constructions so far (and one to be developed in Chapter 34), highlighting the parallels and contrasts between 3D and 4D concepts and notation.

<table>
<thead>
<tr>
<th>3D</th>
<th>4D</th>
<th>Ref. (also see Appendix B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\vec{r}<em>a' = S</em>{ai'}\vec{r}_i$, where $S^i = S^{-1}$</td>
<td>$X^\nu{}<em>{\lambda} = \Lambda^\nu{}</em>{\mu}X^\mu$ where $[\Lambda]^\nu{}_{[\mu} = 0$</td>
<td>Eqns. 32.1, 32.4; Eqns. 32.15, 32.17</td>
</tr>
<tr>
<td>$\vec{T}<em>{\lambda\mu} = S</em>{\lambda}S_{\mu}T_{ij}$</td>
<td>$T^{\alpha\beta} = \Lambda^\alpha{}<em>{\mu}\Lambda^\beta{}</em>{\nu}T^{\mu\nu}$</td>
<td>Eqn. 32.8; Eqn. 33.1</td>
</tr>
<tr>
<td>$</td>
<td></td>
<td>\vec{r}|</td>
</tr>
<tr>
<td>$\vec{V}_a = ((</td>
<td></td>
<td>S'</td>
</tr>
<tr>
<td>Arc length $s$ is a parameter along a curve with $</td>
<td></td>
<td>d\vec{r}/ds</td>
</tr>
<tr>
<td>Thus, $d\vec{r}' / ds$ is the unit tangent vector along a curve.</td>
<td>Thus, 4-velocity $U = d\Gamma /dr$ obeys $</td>
<td></td>
</tr>
<tr>
<td>$\vec{u} \cdot \vec{\omega}_i = \vec{u}_i \vec{\omega}_i$</td>
<td>$V \cdot W = V^\mu g_{\nu\mu}W^\nu = V^\nu g_{\nu\rho}W^\rho$</td>
<td>Sect. 14.4.1; Sect. 32.6.4</td>
</tr>
<tr>
<td>Symmetry and antisymmetry are rotation invariant properties that a tensor may have.</td>
<td>Symmetry and antisymmetry are Lorentz invariant properties that a tensor may have.</td>
<td>Sect. 32.3.4; Sect. 33.3.2</td>
</tr>
<tr>
<td>Contraction reduces rank by 2, for example, $\text{Tr} \vec{T} = \vec{T}<em>{ii} = \vec{T}</em>{ai}$</td>
<td>Contraction reduces rank by 2, for example, $\vec{T}^{\mu\nu} g_{\nu\rho} = \vec{T}^{\alpha\beta} g_{\alpha\beta}$.</td>
<td>Sects. 13.3.1, 32.3.5; Sect. 32.6.4</td>
</tr>
<tr>
<td>A 3-tensor of rank 2 may be used to express any linear, vector-valued function of a vector via $\vec{f}(\vec{u}) = \vec{\omega}' \cdot \vec{u}$.</td>
<td>A 4-tensor of rank (3,) may be used to express any linear, vector-valued function of a vector via $G(\vec{U})^\lambda = F^{\mu\nu} g_{\mu\lambda} \vec{U}^\lambda$.</td>
<td>Sect. 13.3.1, Eqn. 15.1; Eqn. 33.4</td>
</tr>
</tbody>
</table>

33.3.5’ More on beauty

[Is a physical model likely to be true because it seems beautiful? Surely not—to think so would be to anthropomorphize Nature. Rather, the role of beauty may simply be that a scientist who is moved by a beautiful idea will follow it to the ends of the Earth, without being overwhelmed by the many misgivings that seem to say the idea contradicts some aspect of reality, nor by the myriad distractions of everyday life.

Why did evolution install this imperative in our brains? Certainly humans are programmed to figure things out, and to make connections; the pleasure we get from using these skills may be reinforcement for a behavior that enhanced our survival in difficult times. We habituate, so we need novelty to keep getting that reinforcement. In science, this means that the most powerful jolts come from unexpected connections that nevertheless carry conviction—the quality called “surprising yet inevitable” earlier. We call that beauty, both in art and in science. ]
33.6’ a Extensions of the Tensor Principle

The Tensor Principle was another sweeping generalization that we owe to Einstein, Minkowski, and others in that generation. The equations governing strong and electroweak interactions have additional “internal” symmetries under other groups (called SU(3) and SU(2)×U(1) respectively), and all known fundamental particles are described by quantizing fields that are tensors jointly under the Lorentz group and these additional groups. The tensor structures associated to the extra transformations are called “multiplets”; for example, each flavor of quark consists of a “color triplet” under SU(3); the up- and down-quark color triplets in turn form an “electroweak doublet,” and so on. Leptons such as electron, muon, and tau (and their neutrinos) are all color singlets but some form electroweak multiplets.

Successfully quantizing these field theories required a method that preserves the symmetry. After many false steps, such methods were found, though they still only work if an “anomaly cancellation” condition holds.
33.1 Too much of a good thing?
Finish the discussion in Section 33.3.3 (page 514), as follows: Section 33.3.1 proposed the equation of motion
\[ \frac{d\mathbf{p}_\mu}{dt} = qF_{\mu\nu} \Gamma_{\nu\alpha} \]  [33.3, page 512]
as a manifestly invariant form of the Lorentz force law. But, this is four equations, whereas the Lorentz force law as we initially stated it has only three components. Give a physical interpretation for the “extra” component of the above equation, and explain why we don’t really have to solve four independent equations in three unknown functions \( \mathbf{\Gamma}(t) \) defining the particle trajectory.

33.2 It adds up
A particle of charge \( q \) and mass \( m \), initially at rest, is released in a region of uniform \( \mathbf{E} \) directed along the \( \hat{x} \) axis. Find the subsequent motion. Be sure to check that in the nonrelativistic limit your solution has the expected form.

33.3 Cyclotron motion
This question follows up on Your Turn 33B (page 517). A proton is released into a region of uniform magnetic field (that is, \( \mathbf{B} \) is a constant vector field). Its initial velocity is directed perpendicular to the field. Find the orbital period of the resulting circular motion, in terms of the radius \( r \) of the proton’s orbit, its mass \( m \) and charge \( q \), and the field strength \( ||\mathbf{B}|| \).
Comment on the small- and large-\( r \) limits of your answer (at fixed \( ||\mathbf{B}|| \)).

33.4 Uniformly moving charge
Use a computer to evaluate Equations 33.10–33.11 at time \( t = 0 \) and display that vector field. That is, find \( \mathbf{E} \) everywhere in space, at one moment of time, for a charged particle located at the origin and in uniform motion along the \( x \) axis at speed \( \beta c \). The formulas show that \( \mathbf{E} \) is axially symmetric, so you only need to evaluate and plot it at points in the \( xy \) plane. Choose those points to be a square grid; make sure that no grid point lies exactly at the origin of coordinates.

Various plot styles have various virtues; it is an art to find the most informative presentation. First, note that the formulas show that the direction of \( \mathbf{E} \) is not so interesting (always radially outward from the origin), so you only need to plot the more interesting magnitude. For each of the two cases (a,b) below, make four plots:

- Make a contour plot of \( ||\mathbf{E}(x, y)|| \). If it’s not informative, try instead \( \log ||\mathbf{E}(x, y)|| \). (Because log is a monotonic function, applying it won’t affect the trends of where the field is large and small, but it will compress the large dynamic range.)
- Make a “heat map” (that is, represent the value of the function by color).
- Actually, the overall \( r^{-2} \) falloff is also not very interesting and may make it harder to see the angular dependence. So make a contour plot of \( (x^2 + y^2) ||\mathbf{E}(x, y)|| \). (In this version you won’t need the log trick.)
Also make a heat map for the new function. In each graphic, ensure that your computer uses equal scales for the $x$ and $y$ axes.

a. Make the four graphics for the case $\beta = 0.1$ and comment.
b. Repeat for the case $\beta = 0.9$ and comment.
c. The preceding instructions didn’t tell you what ranges of $x$ and $y$ to use (other than that they should be equal), nor the value of the charge $q$. Why don’t you need to be told these things?
d. Also, restricting to $t = 0$ doesn’t really limit the generality of your results—why not?

33.5 Induced charge

A rigid, conducting sphere of radius $R$ moves with constant velocity $\vec{v}$ through a uniform magnetic field $\vec{B}$. Assume $v \ll c$ and find the areal charge density induced on the sphere to lowest order in $v/c$. 


Manifestly Invariant Form of Maxwell

You boil it in sawdust, you salt it with glue, you condense it with locusts and tape, Still keeping the principal object in view: To preserve its symmetrical shape.

— Lewis Carroll

34.1 FRAMING: THE RULES

The preceding chapter showed that the Lorentz force law is compatible with the hypothesis that physics is Lorentz-invariant, if we assign certain transformation rules to the electric and magnetic fields. Although those rules do appear simple and natural in 4-tensor language, Nature cares little for our æsthetic judgments; experiments give more compelling foundations to a theory. Encouragingly, Section 33.4.1 found that the transformation we assigned to $E$ is what’s needed to push electrons when a coil moving into the field of a magnet is viewed in its rest frame.

Now we turn to a bigger project. We have operationally defined our new field tensor just by studying the Lorentz force law. There is no further freedom. Now we must cross our fingers and hope that the Maxwell equations, which make many more testable predictions, will also be Lorentz invariant with the same field assignments. Establishing that point will be greatly simplified by following some Rules analogous to those in 3D.

This chapter begins by studying fields only, that is, no charges or currents. Then we will construct a charge flux 4-vector, and add it as a source term in our invariant form of the Maxwell equations.

Electromagnetic phenomenon: A suddenly accelerated charge emits a pulse of electromagnetic radiation.

Physical idea: The uniform motions before and after the acceleration give rise to fields that must be matched across an expanding spherical shell.

34.2 FIELD EQUATIONS IN 4D

So far, many of our constructions have closely paralleled the three-dimensional situation. Now a key difference will emerge.
34.2.1 The 4-gradient transforms differently from any 4-vector

Let's use the abbreviation $\hat{\partial}_\mu$ to mean $\partial / \partial X^\mu$, similarly to the notation $\hat{V}_I$ for $\partial / \partial \hat{r}_I$. Then proceeding as in Equation 32.11 (page 494) gives

$$\frac{\partial}{\partial X^\mu} = \frac{\partial X'^{\tau}}{\partial X^\mu} \frac{\partial}{\partial X'^{\tau}} = \Lambda^\alpha_\mu \hat{\partial}_\alpha.$$ 

So far, everything is exactly the same as in 3D (Equation 32.11, page 494). The new wrinkle is that the resulting rule,

$$\hat{\partial}'_\alpha = ((\Lambda^{-1})_\alpha^\mu) \hat{\partial}_\mu,$$  

is different from the one we started with$^1$ ($X'^{\tau} = \Lambda^\alpha_\mu X^\mu$). This issue did not arise in three dimensions, because for rotation matrices $(S^1)^{-1} = S$. But $(\Lambda^{-1})_\alpha^\mu \neq \Lambda$ in general, so the gradient of a scalar field is not itself a 4-vector field.

All is not lost. We do know that

$$\Lambda^i = [ g \Lambda^{-1} [ g ].]$$  

Sometimes we will reduce clutter by writing simply $[ \hat{\partial} ]$ to indicate the four components of this operator arranged as a column vector. For example, the preceding formula combined with Equation 34.1 gives

$$[ g \hat{\partial}' ] = \Lambda [ g \hat{\partial} ].$$  

Thus, if $\varphi$ is a scalar field, then the four functions $g^{\nu\tau} \hat{\partial}_\nu \varphi$ really do transform as a 4-vector field. Here $g^{\nu\tau}$ is defined as the $4 \times 4$ matrix of constants that is inverse to $g_{\mu\nu}$. Even though the two versions of the metric are numerically equal in any E-inertial coordinate system, the notational distinction will prove useful as we build up consistent Rules for dealing with tensors.$^2$

We could now finish stating tensor Rules just like those in 3D, including that the operator $g^{\mu\tau} \hat{\partial}_\tau$ increases the rank of a tensor field by one. In practice, however, a slight elaboration of the Rules lets us write more compact formulas. We abbreviate the operator $g^{\mu\tau} \hat{\partial}_\tau$ as $\hat{\partial}^\mu$, because Equation 34.2 says that it transforms like $X'^\mu$. This operation is often called index raising.$^3$ Then

$$\hat{\partial}^{\tau\tau} = \Lambda^\alpha_\mu \hat{\partial}_\mu$$

looks like the corresponding 3D relation (Equation 32.12, page 494). But instead of explicitly raising every lower index arising from a gradient, we will just keep separate track of each index’s position and subdivide a tensor’s rank into two integers: $\text{Rank} \left( \begin{smallmatrix} p \\ q \end{smallmatrix} \right)$ means $p$ upper and $q$ lower indices. $\text{Rank} \left( \begin{smallmatrix} 1 \\ 0 \end{smallmatrix} \right)$ is therefore our old friend the 4-vector; $\text{Rank} \left( \begin{smallmatrix} 0 \\ 1 \end{smallmatrix} \right)$ is sometimes called a 4-covector.$^4$
Ex. Show that index-raising on any 4-tensor of rank \( \frac{1}{0} \) yields rank \( \frac{1}{0} \), regardless of whether the original arose as a 4-gradient.

**Solution:** Use the transformation of a general 4-covector \( W \):

\[
\varepsilon^{\alpha\beta} W^\nu_{\alpha} = \left[ g W^\nu \right]^{\alpha} = \left[ g \Lambda^{-1} \right]^{\alpha}. 
\]

Next, use Equation 32.20 (page 498) to rewrite this as \( \left[ \Lambda g W \right]^{\nu} = \Lambda^{\alpha}_{\mu} \left[ g W \right]^{\mu} \).

Ex. Will this elaboration go on forever? Do we need yet another class of objects to describe the transformation of \( \partial/\partial X^\mu \)?

**Solution:** Notice that

\[
\frac{\partial}{\partial X^\mu} = \frac{\partial X^\nu}{\partial X^\mu} \frac{\partial}{\partial X^\nu} = \frac{\partial}{\partial X^\mu} \left( g_{\nu\lambda} X^\lambda \right) \frac{\partial}{\partial X^\nu} = g_{\mu\nu} \delta^\lambda_\nu \frac{\partial}{\partial X^\lambda}. 
\]

Multiplying both sides from the left by a metric tensor then shows that \( \partial/\partial X^\nu = \delta^\nu \), which we already have seen is 4-vector—not anything new.

The mnemonic is to say that for derivatives, “lower index in the denominator is an upper index,” just as we previously had “upper index in the denominator is a lower index.” In short, just the two index types already defined (up and down) are enough.

### 34.2.2 Three kinds of invariant product

The doubling of index types causes us surprisingly little trouble in practice. All we need to remember is that

To get invariant formulas, we must only contract indices in pairs, one upper and one lower. (34.3)

**Vector with covector**

Imagine walking along a parameterized curve \( \Gamma(\xi) \), measuring some scalar property \( \phi \) of the world. Then the rate of change of our measurement with respect to \( \xi \) is well defined regardless of what coordinates we use on spacetime. The Chain Rule says that rate is the gradient of \( \phi \) along the curve, or

\[
(d\Gamma^\mu / d\xi)(\partial_\mu \phi). 
\]

This expression contracts a vector with a covector, so we conclude that such a contraction is invariant without any additional metric factor, an instance of Idea 34.3. Indeed, Equation 34.1 gives

\[
\left[ \frac{d\Gamma}{d\xi} \right]^{\nu} \left[ \partial^\nu \right] = \left( \Lambda \frac{d\Gamma}{d\xi} \right)^{\nu} \Lambda^{-1} \left[ \partial^\nu \right] = \left[ \frac{d\Gamma}{d\xi} \right]^{\nu} \Lambda^{-1}. 
\]
Chapter 34 Manifestly Invariant Form of Maxwell

Vector with vector

On the other hand, if we wish to contract two 4-vectors, then we do need a metric to satisfy Idea 34.3: \( \bar{X}^{\mu} g_{\mu\nu} Y^{\nu} \), the rule already found in Section 32.6.4 (page 500). Because \( [g] = \mathbb{1} \), we can invert index raising by another multiplication by \( g \):

\[
\underline{W}_\mu = g_{\mu\nu} W^\nu. \quad \text{index lowering}
\]

Covector with covector

For two covectors, simply convert each to a 4-vector and use the usual invariant product:

\[
(g\mu W_\nu) g_{\mu\nu} (g^{\rho\sigma} V_\rho) = [W^\nu g g V] = W_\rho g^{\nu\sigma} V_\sigma.
\]

34.2.3 The wave operator is the invariant contraction of two 4-gradients

The ideas in the previous section make it straightforward to find a manifestly invariant derivative operator that, when applied to a scalar function, yields another scalar function:

\[
\Box = \partial^\mu \partial_\mu.
\]

It’s called the wave operator, D’Alembert operator, or dalembertian.

Your Turn 34A

Show that \( \Box \) is the same wave operator that we have been writing all along (Section 25.4, page 393), and whose invariances led us to discover the Lorentz transformations in the first place.

We can now take another step. If we apply the wave operator to a tensor of any rank, the result is again a tensor of the same rank. Setting that to zero yields a Lorentz-invariant field equation. That observation immediately suggests the candidate equation

\[
\Box F^{\mu\nu} \gamma = 0
\]

for electrodynamics! Could it really be that simple? Well, no: The Maxwell equations are only first-order in derivatives. But we’ll soon find something almost as simple, and correct.

34.3 GENERAL 4-TENSORS

34.3.1 More examples

Any quantity with \( 4^{p+q} \) components that transform with \( p \) copies of \( \Lambda \) and \( q \) copies of \( (\Lambda^{-1}) \) as we change between E-inertial coordinate systems on spacetime can represent a 4-tensor of rank \( \left( \frac{p}{2} \right) \). We address individual components with Greek indices as usual, but \( p \) of the indices are placed as superscripts and the remaining \( q \) as subscripts.

Extending the list we started in Section 33.2,
34.4 Summary: The Rules in 4D

- The gradient of a scalar function has rank \( (0, 1) \) (Equation 34.1, page 526);
- The Faraday tensor \( F^{\mu \nu} \) has rank \( (2, 0) \) (Equation 33.5, page 513);
- The quantities \( F_{\mu \nu} ^{\alpha \beta} g^{\alpha \lambda} \) constitute a 4-tensor of rank \( (1, 1) \), which we now will denote by \( P^{\mu \lambda} \), and so on.

34.3.2 Symmetry

Let \( A^{\mu_1 \cdots \mu_p \nu_1 \cdots \nu_q} \) be a 4-tensor of rank \( (p, q) \). We saw explicitly in the case \( p = 2, q = 0 \) that the property of being antisymmetric is Lorentz-invariant (Equation 33.8, page 517).

Your Turn 34B

a. Show more generally that if the components of \( A \) are antisymmetric under permutation of some or all of its upper indices in one inertial coordinate system, then \( A \) will have that same property in any other such system (and similarly for lower indices). Similarly, if the components are symmetric under permutations, that property, too, is invariant.

b. Also show that the operation of antisymmetrizing (or symmetrizing) a tensor on some or all of its upper (or lower) indices is invariantly defined.

But beware: There is no invariant sense to (anti)symmetry between an upper and a lower index. We must lower one index, or raise the other, before we can speak invariantly of (anti)symmetry.

34.3.3 The metric is itself a tensor

You now have all the tools to show that the metric is a “tensor from Heaven,” that is, numerically the same when viewed in any inertial coordinate system.\(^5\)

Your Turn 34C

a. The metric as we first introduced it, \( g_{\mu \nu} \), has two lower indices. Show that this matrix indeed gives the components of a 4-tensor of rank \( (0, 2) \), as implied by the notation. \( \text{[Hint: Use Equation 32.20, page 498.]} \)

b. Section 34.2.1 defined the related symbol \( g^{\mu \nu} \) as the inverse matrix to \( g_{\mu \nu} \) (and hence, numerically equal to it). Show that this matrix indeed gives the components of a constant 4-tensor of rank \( (2, 0) \), as implied by the notation. Show that indeed it arises by raising the indices of \( g_{\mu \nu} \).

34.4 Summary: The Rules in 4D

This is getting scary. What saves us from total confusion is that a few Rules make it unnecessary to think much about these intricate transformations. These Rules correspond

\(^5\)See Chapter 14.
to the ones in Section 32.5 (page 495), and are almost as easy to use.

We are exploring the hypothesis that electrodynamics is invariant under Lorentz transformations. To generate Lorentz-invariant equations as candidate laws of Nature, we organize all the dynamical variables into 4-tensors of suitable rank. Some were proved earlier; others are easy (but worthwhile) to prove now:

a’. A 4-tensor of rank \( \frac{p}{q} \) can be represented in a particular inertial coordinate system by \( 4^{p+q} \) numbers, indexed by \( p \) upper and \( q \) lower indices, with a Lorentz transformation law that “acts on” each index in a way appropriate to its up/down status (for example, Equations 33.1 and 34.1).

b’. A 4-tensor field is the same idea, but each entry is a function of \( X \).

c’. Permuting a set of upper indices on the components of a tensor yields another tensor of the same rank (Your Turn 34B), as does permuting a set of lower indices.

d’. The sum of of two tensors with the same rank yields a new tensor of that rank. The new tensor has components that are sums of corresponding components of the two summands. Combining with (c’) then shows that the operation of symmetrizing is invariantly defined.

e’. The collection of all products of the components of tensors of rank-\( \frac{p}{q} \) and \( \frac{p'}{q'} \) itself represents a rank-\( \frac{p+p'}{q+q'} \) tensor called the tensor product, denoted \( A \otimes B \). For example, the tensor product of \( \left( \frac{1}{0} \right) \) and \( \left( \frac{0}{1} \right) \) is the tensor of rank \( \left( \frac{1}{1} \right) \) with component \( \left( V^\mu \otimes j^\mu_{\nu} \right) \) in row \( \mu \) and column \( \nu \) equal to \( V^\mu \otimes \omega_{\nu} \).

f’1. Index lowering is an invariant operation that changes the rank from \( \frac{p}{q} \) to \( \frac{p-1}{q+1} \), and similarly for its inverse (index raising).

f’2. Only contract indices in up/down pairs. The result of such a contraction is again a tensor, with reduced rank \( \left( \frac{p-2}{q-1} \right) \). When tempted to contract two upper indices, first lower one of them, then contract the resulting new lower index with the other upper index. The net effect is to bring the rank down to \( \left( \frac{p-2}{q-1} \right) \) as desired.

f’3. Similarly, to contract two lower indices, first raise one of them.

g’. The operator \( \partial_\mu \) increases the rank of a tensor field by \( \left( \frac{0}{1} \right) \) (Section 34.2.1).

h’. A physics equation of the form \( A = 0 \), where \( A \) is a tensor (or tensor field), is guaranteed to be Lorentz invariant.

i’. The 4D volume element \( d^4X \) transforms to \( d^4X' \) under Lorentz transformations because the jacobian matrix has determinant\( \pm 1 \). Thus, we may convert a tensor field to a constant tensor of the same rank by integrating over all spacetime.

With these Rules, 4-tensor manipulations become so automated that most physicists don’t consciously distinguish between, say, \( E^\mu \) and \( F^\mu \); either one is called “the” Faraday tensor and only index placement is used to tell them apart. If you’ve got one, but you want the other, then you convert by index raising or lowering operations. But beware: If you plan to use index-free (matrix) notation, you need to state which of these quantities

\[ ^6 \text{See Idea 33.2 (page 511).} \]
\[ ^7 \text{Section 32.4.2’ (page 507) makes this idea more explicit.} \]
\[ ^8 \text{So see this, take the determinant of both sides of Equation 32.17 (page 498). For more details, see Section 34.9.3. There is no underscore on the } d^4X \text{ because } d^4X \text{ is not a 4-vector.} \]
you mean, because they differ by some crucial minus signs. Matrix notation is extremely concise, but for that very reason we will usually avoid it, now that we have established our “grammar” of invariant constructions.

34.5 VACUUM MAXWELL EQUATIONS

We wish to establish that the Maxwell equations have the property of form invariance under Lorentz transformations. But they appear elaborate, and have a seemingly ad hoc minus sign. Also, $\vec{E}$ and $\vec{B}$ have complicated Lorentz transformations. Rather than approach the problem directly, let’s instead start from scratch.

Chapters 32–33 explained what “from scratch” could mean, via a new way of thinking, driven by invariance properties. Let’s apply that “Einstein thinking” to the Maxwell equations:

- Abstract away from Maxwell’s version the structural features: The desired equations are first-order in space and time derivatives. They involve an antisymmetric, rank-$(2,0)$ tensor field $F$. In addition, four of them involve charges and currents, while the other four do not. There are also two scalar constants $\varepsilon_0$ and $\mu_0$, or equivalently $\mu_0$ and $c = (\varepsilon_0\mu_0)^{-1/2}$.
- What could the equations be? If they take the form $(\text{tensor field}) = 0$, then The Rules say they’ll be automatically invariant (Section 34.4).
- Once we have guessed candidate equations that meet the criteria, we can ask how they look when phrased in terms of the old-school $\vec{E}$ and $\vec{B}$ fields via the correspondence found in Chapter 33. If they coincide with the Maxwell equations as we’ve been writing them, then we’ll have completed the proof that electrodynamics is Lorentz-invariant.

We could implement the first bullet with the candidate equation

$$\partial_\mu F^{\mu\sigma} = 0,$$

but that can’t be right. For one thing, it’s $4 \times 6 = 24$ equations, because $\mu\sigma$ is an antisymmetric pair, but we only wanted eight equations. Worse, we know all about the solutions to those equations: They say that all six components of $F$ are constants. Too many equations have too impoverished a set of solutions.

But maybe we could reduce the equations without spoiling their Lorentz-invariance. One possibility is to contract indices:

$$\partial_\mu F^{\mu\nu} = 0. \quad (\text{in vacuum})$$

The Rules say that this equation is still Lorentz-invariant, but now it’s just four equations, because there’s only one loose index.
Your Turn 34D

a. Rephrase Equation 34.5 in terms of the traditional $\vec{E}$ and $\vec{B}$ by using the dictionary in Equation 33.6 (page 514). Confirm that indeed it’s precisely the electric Gauss law and Ampère’s law in vacuum—there is no need to tweak those equations, which were secretly Lorentz-invariant all along.

b. There are three ways to choose a pair of indices to contract in Equation 34.4. So far, we’ve only considered one. What about the other two ways?

Notwithstanding your result in (b), a second reduction of the candidate equation is possible, just a bit more subtle:

Your Turn 34E

Show that the totally antisymmetric part of Equation 34.4 is\(^9\)

$$\varepsilon_{\mu\nu} F_{\nu\lambda} + \varepsilon_{\nu\lambda} F_{\mu\nu} + \varepsilon_{\lambda\mu} F_{\mu\nu} = 0.$$  \hspace{1cm} (34.6)

The Rules say that the left side of Equation 34.6 is a tensor, so the statement that these quantities equal zero is Lorentz invariant, and hence a candidate for a law of Nature.

Equation 34.6 may appear to be $4^3 = 64$ equations, because it has three loose indices. Really, however, most of these equations are redundant or automatically satisfied, because a totally antisymmetric 4-tensor of rank $\left(\frac{4}{2}\right)$ has only four independent components.\(^{10}\)

Your Turn 34F

Write down all four independent components of Equation 34.6. You’ll need the expressions obtained by index lowering the identifications we found in Equation 33.6 (page 514):

$$F_{\mu\nu} = \begin{bmatrix} 0 & -\hat{\beta}_{1/c} & -\hat{\beta}_{2/c} & -\hat{\beta}_{3/c} \\ \hat{\beta}_{1/c} & 0 & \hat{\beta}_{2} & -\hat{\beta}_{2} \\ \hat{\beta}_{2/c} & -\hat{\beta}_{3} & 0 & \hat{\beta}_{3} \\ \hat{\beta}_{3/c} & \hat{\beta}_{2} & -\hat{\beta}_{3} & 0 \end{bmatrix}_{\mu\nu}.$$  \hspace{1cm} (34.7)

Once again, you’ll find precisely the magnetic Gauss law and the Faraday law—so they, too, were secretly Lorentz-invariant all along.

Section 34.5\(^*\) (page 542) reformulates Equation 34.6 in a way that is sometimes useful.

---

\(^9\)As mentioned in Section 34.3.2, before we can invariantly antisymmetrize a tensor, we must push all of its indices into matching position, either by lowering the upper ones (as done here) or by raising the lower one.

\(^{10}\)See Problem 34.1.
Figure 34.1: **Unified construction of (a) charge density and (b) charge flux**, an extension of the one in Figure 8.2b (page 116). These spacetime diagrams don’t show the z direction; the blue box is actually a solid 3D region in each panel. Dashed lines indicate charged particle trajectories that make no contribution because they don’t pass through the selected windows. Thus, in (a), trajectory #2 may eventually pass through the spatial region shown, but not at time \(t_c\). Similarly, in (b), trajectory #4 does pass through the selected range of \(ct\) and \(x\) (*green bracket*), and it also crosses \(y_\ast\) (*green dot*), but no point along it does both. In contrast, *blue dots* denote nonzero contributions to (a) the charge density: \(q_i/\Delta^3r\) or (b) the y component of charge flux: \((q_1 - q_2)/(\Delta^3r_\ast\Delta t)\).

### 34.6 THE CHARGE FLUX 4-VECTOR

Equation 34.6 is already complete, because the magnetic Gauss and Faraday laws don’t involve charges nor currents. To complete our job, we now need to upgrade Equation 34.5 to include charges and currents.

#### 34.6.1 A graphical formulation unifies charge density and flux

This section repeats the discussion in Chapter 8 in our new 4D language. For artistic reasons, Figure 34.1 only shows two space dimensions \(x, y\), but \(z\) is understood to be present.

Imagine a swarm of charged particles. Each one’s trajectory is a curve in spacetime, parameterized by proper time \(\tau: \Gamma_{\prec\rightarrow}(\tau)\). Each carries a scalar constant \(q_\ell\) (its charge). As always, we choose an inertial coordinate system on spacetime.

To define charge density at some point \(X_\ast\) (an “event”), set up a small spatial volume element \(\Delta^3X_\perp\), that is,

\[
ct = X^0_\ast = \text{const}, \\
X^1_\ast < x < X^1_\ast + \Delta X^1, \\
X^2_\ast < y < X^2_\ast + \Delta X^2, \\
X^3_\ast < z < X^3_\ast + \Delta X^3.
\]

In Figure 34.1a, the rectangle shown represents \(\Delta^3X_\perp\). Now add up all the charges on lines crossing this element from past to future, divide by volume \(\Delta^3X_\perp\), multiply by \(c\), and call the result \(J^0(X_\ast)\). For example, trajectory #1 contributes \(cq_1/\Delta^3X_\perp\), whereas trajectory #2, which misses the selected volume element, contributes nothing.

Note that the quantity \(J^0\) just defined has units \(\text{coul}/(\text{s m}^2)\). In fact, \(J^0\) is the quantity we’ve previously called \(cp_q\).
Next, define charge flux at $X_\ast$ by setting up a new small 3D element (rectangle in Figure 34.1b). We again call it $\Delta^3X_\ast$, even though one of its directions is in time:

$$
X^0_\ast < ct < X^0_\ast + \Delta X^0,
X^1_\ast < x < X^1_\ast + \Delta X^1,
y = X^2_\ast = \text{const},
X^3_\ast < z < X^3_\ast + \Delta X^3.
$$

(34.8)

Add up all the charges on trajectories crossing this element from smaller to larger values of $y$, and subtract all the charges on trajectories crossing it in the opposite sense. Again divide by $\Delta^3X_\ast$, multiply by $c$, and call the result $J^2(X_\ast)$. Thus, in the sketch trajectory #1 contributes $cq_1/\Delta^3X_\ast$, #2 contributes $-cq_2/\Delta^3X_\ast$, and #3–4 contribute nothing.

Define the other two components $J^1$ and $J^3$ similarly. Thus, all four components of $J_\mu$ have the same units. In fact, $J^i$ are the three components of the charge flux $j_i$ introduced in Section 8.3 (page 116). The advantage of the present formulation is that it treats all four components in the same way. In any inertial coordinate system,

$$
J^\mu = c \text{ times net amount of charge crossing the surface } \{X^\mu_\ast = \text{constant}\},
\text{ from smaller to larger } X^\mu_\ast, \text{ per } d^3X_\ast.
$$

(34.9)

### 34.6.2 $J_\mu$ is a 4-vector

The Tensor Principle claims that all physical quantities can be packaged into 4-tensors. Does $J^\mu$ defined in the preceding section fit?

Chapter 8 considered a small hypercube and showed that, because charge is locally conserved, we must have

$$
\frac{\partial}{\partial t} \rho_q + \nabla \cdot J = 0.
$$

[8.4, page 117]

This continuity equation can be written more elegantly as

$$
\sum_{\mu} \frac{\partial J^\mu}{\partial X^\mu} = 0,
$$

(34.10)

or more concisely still as

$$
\partial_{\mu} J^\mu = 0. \text{ continuity}
$$

(34.11)

Our derivation of Equation 8.4 was valid in any coordinate system, so in particular the form of Equation 34.11 is the same in any inertial system. We also know that $\partial_{\mu}$ form a covector and the index contraction is a Lorentz-invariant operation. Thus, the four quantities

$$
J(X) = \begin{bmatrix}
\frac{c\rho_q(t,\vec{r})}{j(t,\vec{r})}
\end{bmatrix}
$$

(34.12)

must themselves transform as a rank-$\left(\begin{smallmatrix}1 \\ 0 \end{smallmatrix}\right)$ field: the charge flux 4-vector field.$^{12}$

---

$^{11}$See also Section 34.9.3 for a more explicit proof.

November 14, 2023; Contents  Index  Notation  Glossary
34.7 COMPLETE, MANIFESTLY INVARIANT MAXWELL EQUATIONS

We are now ready to add charges and currents to Equation 34.5. Once again, there’s really no freedom! The left side of Equation 34.5 is a 4-vector, so we must set it equal to a 4-vector. We have seen that charges and currents constitute a 4-vector. All we need is a scalar constant of proportionality to make the units work out:

\[
\vec{\nabla}_\mu F^{\mu\nu} = \mu_0 J^\mu \quad \text{and} \quad \vec{\nabla}_\mu F^{\nu\lambda} + \vec{\nabla}_\nu F^{\lambda\mu} + \vec{\nabla}_\lambda F^{\mu\nu} = 0. \tag{34.13}
\]

**Your Turn 34G**

Extend Your Turn 34F to confirm that the first of Equations 34.13 really is equivalent to Equations 0.1 (page 2) and 0.4 (page 2).

The eight beautiful\(^\text{13}\) new equations, Equations 34.5–34.6, have turned out to be exactly the Maxwell equations we have been using since the Prologue! But their complete Lorentz invariance (and that of the Lorentz force law) is now obvious. Along the way, we have also addressed Hanging Question #B (page 13): The form of the equations isn’t arbitrary after all, but rather is dictated by symmetry principles. Moreover, no Levi-Civita tensor appears in Equations 34.13; thus, they are also manifestly invariant under inversions, unlike the traditional formulation in terms of \(\vec{E}\) and \(\vec{B}\).\(^\text{14}\)

Section 34.7′ (page 542) discusses the proper counting of these equations and Hanging Question #D. It also discusses spatial inversion and time-reversal symmetries.

34.8 FOUR-VECTOR POTENTIAL

34.8.1 The Poincaré lemma again implies the existence of a potential

The second of Equation 34.13, together with the Poincaré lemma,\(^\text{15}\) immediately implies that we can always write the Faraday tensor in terms of a four-vector potential:\(^\text{16}\)

\[
F^{\mu\nu} = \vec{\nabla}^\mu A^\nu - \vec{\nabla}^\nu A^\mu. \tag{34.14}
\]

The Rules imply that \(A\) must be a four-vector field in order for \(F\) to be an antisymmetric four-tensor field as desired.

---

\(^{13}\)“Surprising yet inevitable” (Section 33.3.5, page 516).

\(^{14}\)This addresses Hanging Question #E. Nor is any choice of right hand buried in the recipe that converted particle trajectories into \(J^\mu\) (Equation 34.9), nor in the one that operationally defines \(F\) (the Lorentz force law, Equation 33.3).

\(^{15}\)See Idea 15.13 (page 227), which holds in any number of dimensions.

\(^{16}\)Chapter 18 already derived this, but in a way that required a tricky insight.
Your Turn 34H

Work out the corresponding $\vec{E}$ and $\vec{B}$, and show that Equation 34.14 reproduces Equation 18.26 (page 281) when we make the assignments

$$\mathbf{A}^\mu = \left[ \frac{\psi/c}{\mathbf{A}} \right]^\mu.$$

Thus, the potentials we found long ago also adhere to the 4D Tensor Principle. SI units for the 4-vector potential are $[\mathbf{A}] \sim \text{kg m/(coul s)}$.

Gauge invariance is the observation that the Faraday tensor doesn’t change when we replace $\mathbf{A}^\mu$ by

$$\mathbf{A}^\mu = \mathbf{A}^\mu + \mathbf{\Xi}^\mu.$$

(34.15)

Your Turn 34I

a. Prove that last statement starting from Equation 34.14 and connect to Section 18.8.2 (page 281).

b. Show that when we substitute Equation 34.14 into the Maxwell equations, some of them are automatically satisfied.

c. Show that the remaining Maxwell equations become

$$-\Box \mathbf{A}^\nu + \partial_\nu \mathbf{A}^\mu \mathbf{A}_\mu = \mu_0 \mathbf{J}^\nu.$$

(34.16)

Your result establishes that the Maxwell equations can be written as four equations in four unknown functions, even though they started as eight equations in six unknowns (Equation 34.13).

Section 34.8.1 (page 544) discusses the counting in more detail, and introduces an extended notion of gauge field.

34.8.2 Charged particle in uniform motion revisited

For a first look at the benefits of using 4-potentials, we can return to the problem of a charged particle in uniform motion, already solved in Section 33.4.2 (page 518). Choose one inertial coordinate system that moves along with the charge and denote it with a prime. Then components of the 4-vector potential in that system are just those of a point charge at rest:

$$[\mathbf{A}'] = \frac{q}{4\pi\varepsilon_0 c||\mathbf{r}'||} \left[ \begin{array}{c} 1 \\ 0 \end{array} \right].$$
In this expression, \( r = \left[ \frac{\gamma (x - \beta c t)}{z} \right] \). Restricting to the \( xy \) plane gives

\[
[A(t, x, y, 0)] = \begin{bmatrix}
\frac{q}{4\pi \varepsilon_0 c} & \gamma & -\beta y & 1 \\
0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 \\
\end{bmatrix}
\begin{bmatrix}
f(t, x, y) \\
0 \\
0 \\
0 \\
\end{bmatrix}
= \frac{q}{4\pi \varepsilon_0 c} \begin{bmatrix}
\gamma f \\
0 \\
0 \\
0 \\
\end{bmatrix},
\] (34.17)

where \( f(t, x, y) = 1/r' = (\gamma^2 (x - \beta c t)^2 + y^2)^{-1/2} \).

We can now compute the Faraday tensor as usual. For example,

\[
F^{01} = e^{-1} E_x = \partial^0 A^1 - \partial^1 A^0 = \frac{q}{4\pi \varepsilon_0 c} \left( -\frac{\partial}{\partial ct} (\gamma \beta f) - \frac{\partial}{\partial x} (\gamma f) \right) = \frac{q}{4\pi \varepsilon_0 c} \gamma f^3(x - \beta c t).
\]

We have recovered the results already found in Your Turn 33C and Your Turn 33D (page 519). However, sometimes \( A \) is all that’s needed, and Equation 34.17 shows that it was easier to obtain than the electric and magnetic fields.

If a particle changes from one uniform velocity to another one (for example, by coming to a sudden stop), then its 4-vector potential must interpolate between the two respective versions of the preceding result. In Problem 34.3 you’ll investigate the physical meaning of this interpolation.\(^{17}\)

### 34.9 MORE ABOUT \( J \)

The geometric definition of the charge flux 4-vector in Section 34.6.1 is useful for some purposes, for example, to see why it obeys the continuity equation. However, in Chapter 35 it will be helpful to know that another formulation is equivalent to the geometric one.

#### 34.9.1 Delta of a function on spacetime is again a set of delta functions

First we need to review a key fact about the delta function.\(^{18}\) Think of it as a bump, for example, \( \delta(x; \sigma) = (2\pi \sigma)^{-1/2} e^{-x^2/(2\sigma^2)} \) with \( \sigma \) very small. So

\[
\int_{-\epsilon}^{\epsilon} dx \, \delta(x; \sigma) \to 1
\]

if we hold \( \epsilon \) fixed to any positive value and take \( \sigma \to 0 \).

Now define a new function \( f(x; \sigma) = \delta(2x; \sigma) \) and compute the integral, changing variables to \( y = 2x \):

\[
\int_{-\epsilon}^{\epsilon} dy \, f(x; \sigma) = \int_{-\sigma}^{\sigma} dy \, (2\pi \sigma)^{-1/2} e^{-y^2/(2\sigma^2)} \to \frac{1}{2}.
\] (34.18)

Again the limit is taken holding \( \epsilon \) fixed to any positive value and \( \sigma \to 0 \). In the same limit, the integral would have been zero had we chosen any range not centered on \( x = 0 \).

\[^{17}\text{Chapter 50 will revisit the question in greater detail.}\]

\[^{18}\text{This was introduced in Section 0.3.8 (page 11).}\]
Thus, \( f \) has the same properties as those defining \( \frac{1}{2} \delta(x) \). More generally,

\[
\delta(ax) = \frac{1}{a} \delta(x) \quad \text{for positive constant } a.
\]

Next, define \( g(x; \sigma) = \delta(-2x; \sigma) \). Its graph is the same as that of \( f \), so it has the same integral:

\[
\delta(ax) = \frac{1}{|a|} \delta(x) \quad \text{for any constant } a. \tag{34.19}
\]

More generally, if \( h(x) \) is any smooth function that vanishes at an isolated point \( x_* \), then

\[
\delta(h(x)) = \left| \frac{dh}{dx} \right|_{x_*}^{-1} \delta(x - x_*). \tag{34.20}
\]

(Equation 34.19 corresponds to \( h(x) = \pm ax \).) If \( h(x) = 0 \) at several points, then we get a sum with a term for each such point.

We can generalize these results to higher dimensions using

\[
1 = \int d^4x \delta^{(4)}(X) = \int d^4x' \delta^{(4)}(X'), \tag{34.21}
\]

where the first integral is over any small region containing the origin, and the second is over any small region containing \( X_* \). Suppose now that \( G^\alpha \) are a set of four functions of \( X \) that define a new set of coordinates \( X' \), centered at some point \( X_* \). Then instead of the factor \( 1/2 \) in Equation 34.18, we have the inverse of the Jacobian factor. The generalization of Equation 34.20 is therefore\(^{19}\)

\[
\delta^{(4)}(G^\alpha(X)) = \left| \det \frac{\partial G^\alpha}{\partial X^\nu} \right|^{-1} \delta^{(4)}(X^\mu - X'^\mu). \tag{34.22}
\]

For a Lorentz transformation, \( G^\alpha \) is a set of four linear functions. The derivatives appearing in Equation 34.22 are the constant matrix \( \Lambda^\alpha_\nu \). The determinant of that matrix is \( \pm 1 \) because \( \left[ \Lambda g \Lambda \right] = \left[ g \right] \), so Equation 34.22 says \( \delta^{(4)}(X) \) is a 4-scalar field. Equation 34.21 implies that its dimensions are \( L^{-4} \).

### 34.9.2 \( J \) may be formulated in terms of individual trajectories

Here is another set of quantities that may also seem reasonable as a candidate for the 4-current.

Define four functions on spacetime by putting bumps all along each trajectory \( \Gamma_{(\ell)} \):

\[
J^\mu_{alt}(X) = \sum_\ell \int_{-\infty}^{\infty} d\tau q_\ell U^\mu_{(\ell)}(\tau) \delta^{(4)}(X - \Gamma_{(\ell)}(\tau)). \tag{34.23}
\]

The following paragraphs will show that \( J_{alt} \) is equal to the \( J \) defined above. (Already you can check that the units match.)

---

\(^{19}\)Because \( G^\alpha \) are coordinates, they only all vanish at the one point \( X_* \).
Consider any component of Equation 34.23, for example $\mu = 2$, and any starting point $X$. Let’s start by showing that $J^2_{\text{alt}}(X) = J^2(X)$. Let $X_{\perp}$ denote just the other components $(0, 1, 3)$. As in Equation 34.8, let $\Delta^3X_{\perp}$ be a small region about $X$ obtained by varying those three coordinates. We will now show that $J^2_{\text{alt}}$ and $J^2$ give the same result when integrated over this region.

First consider

$$\int_{\Delta^3X_{\perp}} \text{d}(ct) \text{d}xdz J^2_{\text{alt}} = \sum_{\ell} \int_{\Delta^3X_{\perp}} \text{d}(ct) \text{d}xdz \int \text{d}\tau \frac{q_{\ell}U^2(\tau)}{r^2(\tau)} \delta(X^2 - r^2(\tau)) \delta^{(3)}(X_{\perp} - r_{\perp}(\tau)).$$

The things in the brace don’t depend on $t$, $x$, or $z$, so we may bring them to the front:

$$= \sum_{\ell} \int \text{d}\tau \frac{q_{\ell}U^2(\tau)}{r^2(\tau)} \delta(X^2 - r^2(\tau)) \int \Delta^3X_{\perp} \text{d}(ct) \text{d}xdz \delta^{(3)}(X_{\perp} - r_{\perp}(\tau)).$$

The innermost integrations just give 1 if particle #\ell’s transverse coordinates fall anywhere inside $\Delta^3X_{\perp}$ at proper time $\tau$, and otherwise zero.

Now turn to the rest of Equation 34.25. If trajectory #\ell is ever inside the range $\Delta^3X_{\perp}$ and crosses the fixed $y$ that we are considering, then let $\tau_{\ell}$ be the proper time when that crossing occurs. Equation 34.20 gives the remaining integrand (in the brace) as

$$\left(\frac{eq_{\ell}}{d\tau}ight) \frac{d\Gamma^2(\tau)}{d\tau} \delta(\tau - \tau_{\ell}) = \pm eq_{\ell} \delta(\tau - \tau_{\ell}).$$

We get the plus sign if the trajectory crosses from smaller to larger $y$, or the minus sign in the contrary case.

Putting it all together, the only trajectories that make nonzero contributions to Equation 34.24 are those that actually pass through $\Delta^3X_{\perp}$ at the chosen $y$. We may thus restrict the sum to only those trajectories, which we denote by $\sum'_{\ell}$, and so Equation 34.24 becomes

$$\int_{\Delta^3X_{\perp}} \text{d}(ct) \text{d}xdz J^2_{\text{alt}} = \pm \sum'_{\ell} (\pm eq_{\ell}).$$

At last, we can see that Equation 34.26 is the same property that we used to define the current $J^2$ in Equation 34.9. Because the small integration region was arbitrary, that result implies that $J^2_{\text{alt}} = J^2$ everywhere. Repeating the argument for the other three components yields the same conclusion for all of $J_{\text{alt}}$.

### 34.9.3 Another proof that $J$ is a 4-vector

Now we can use our reformulation of the charge flux (Equation 34.23) to show that $J$ is a 4-vector. Indeed, in that equation $dr$ is a 4-scalar, the $q_{\ell}$ are all 4-scalars, we just showed

---

20For a small enough region $\Delta^3X_{\perp}$, there will be at most a single crossing. In Figure 34.1b (page 533), trajectory #4 passes through $\Delta^3X_{\perp}$, but it’s not there when it crosses the chosen $y$ value, so it doesn’t contribute to Equation 34.25. Trajectory #3 never visits the chosen $y$ at all.
that the delta function is a 4-scalar, and \( U \) is a 4-vector (it is the derivative of the 4-vector \( X \) with respect to the scalar \( \tau \).

Section 34.9 (page 544) formulates \( I \) in a more geometrical way.

### 34.10 A DIZZYING VISTA

Boltzmann famously said, “Elegance is for cobblers and tailors.” Should we care that Equations 34.13 are so beautiful?

One pragmatic answer is “yes,” because the manifestly invariant form will make it much easier to finally establish local conservation of energy and momentum (Chapter 35), and indeed the very general relation between symmetry and invariance (Chapter 40). These results were originally obtained without 4-tensor notation, but it’s much harder to do it right without the simplicity we’ve now gained.

Moreover, the train of thought in the last few chapters led Einstein to unravel a seemingly unrelated puzzle. It’s a fantastic detective story: A formal observation about the structure of electromagnetism led Einstein to a hypothesis, with testable quantitative predictions, about the nature of gravitation.

Einstein began by asking himself, what exactly is it that makes some coordinate systems (the inertial ones) particularly good? Why aren’t all systems equally good?

Our discussion of waves on a vibrating string gives a hint.\(^{21}\) Faced with a dynamical equation (for the string’s transverse displacement) with less symmetry than expected (no galilean invariance), we realized that some additional dynamical variable (the velocity of the string) is hiding in the equation, implicitly set to some particular value (zero). Explicitly acknowledging this implicit physical object, and realizing that its value, too, will change under coordinate transformations, restored the full galilean invariance to the string’s wave equation.

What happens if we try the same line of thought with the Maxwell equations? What is the hidden dynamical variable? Einstein argued it’s not the velocity of any luminiferous æther. Rather, Section 32.6.3 characterized the “good” coordinate systems as those in which the invariant interval—a metric function on spacetime—takes its simplest form. Thus, to make progress we should start asking

- What is the origin of the invariant interval function? Is it really a fixed property of spacetime, or could \( g \) itself be a dynamical object? (If so, then we’ll need to propose some new dynamical law for it!)
- Do the Maxwell equations become fully coordinate-invariant if we promote the metric tensor to a dynamical variable, with an appropriate transformation law?

The answer to that last question is “yes.”\(^{22}\) Moreover, Einstein found that again there is essentially only one acceptable equation of motion that a metric tensor could have.\(^{23}\)

---

\(^{21}\)See Chapter 27.

\(^{22}\)See Equation 34.27 (page 545).

\(^{23}\)Here “essentially” means there’s actually a two-parameter family of equations. One parameter is Newton’s
He then asked, what new physical phenomena are predicted if we introduce this new
dynamical variable?

The big clue was a fact from the geometry of curved surfaces: Any metric looks equivalent to any other one, if we only look to first order in excursions about a point.\(^{24}\) Einstein asked, is there any physical property of spacetime that also has this property? His answer was: Yes, the gravitational interaction does, because it can always be eliminated locally by passing to a suitable (accelerating) coordinate system. Once again, “Einstein thinking” suggested that the unique equation of motion dictated by general principles like invariance should then describe all gravitational phenomena, including even those not yet imagined (for example, a new gravitational interaction analogous to magnetism, gravitational radiation, and more), and once again, this vision was borne out.\(^{25}\)

### 34.11 PLUS ULTRA

Every physical quantity carries dimensions, which help us to see its role and to formulate reasonable candidate laws. Now we have seen that every physical quantity also has a tensor character, another meta-property that helps us to see its role and to formulate reasonable candidate laws. We’ve seen this play out in electrodynamics, but the ideas are more broadly applicable—when you study liquid crystals, fluctuating fluid membranes, and so on, these ideas are everywhere.

One could quibble that “Einstein thinking” has merely ratified the Maxwell equations, which were discovered without it. But this sort of thinking was later the indispensable intellectual substrate for Dirac to even propose the right wave equation for relativistic particles with spin 1/2.

\(\text{Section 34.11'}\) (page 545) outlines the relativistic treatment of spin.

---

**FURTHER READING**

**Intermediate:**

- 4-tensors, including rank: Schutz, 2022, chap. 3.

**Technical:**


---

\(^{24}\)That is, we may always find normal coordinates (Equation 7.4, page 103).

\(^{25}\)This thread was the insight needed for Hanging Question #F (page 22).
34.5 The 4D Levi-Civita tensor

We can upgrade the discussion of Section 13.5.3 (page 204) to obtain a tensor “from Heaven” in a 4D space with a metric. As in the 3D case, to get started we must again make a choice, analogous to that of a right-hand convention. Suppose that one orthonormal set of basis vectors has been singled out. This, and any other set obtained from it by restricted Lorentz transformations, will be deemed “right-handed and forward directed in time.” We then define a quadrilinear function \( \varepsilon (T, U, V, W) \) by constructing the 4D parallelepiped with edges given by the arguments, finding its metric 4-volume, and multiplying by \(-1\) if \(\{T, U, V, W\}\) are related to the chosen system by a linear transformation with negative determinant. Equivalently, we can define \( \varepsilon \) by declaring \( \varepsilon_{0123} = +1 \) in any right-handed and forward-directed inertial coordinate system.

The main text avoided introducing \( \varepsilon \) to emphasize the manifest invariance of the Maxwell equations under spatial inversion, but using it does allow an elegant reframing of the homogeneous equations:

\[
\varepsilon^{\mu\nu\lambda\sigma} \partial_{\lambda} F_{\sigma\nu} = 0, \quad \text{or} \quad \varepsilon^{\mu\nu\lambda\sigma} \partial_{\lambda} (\varepsilon_{\rho\sigma\tau} F_{\tau\rho}) = 0.
\]

In this formulation, it is clear that we have just four equations: There is just one loose index. Because the left side is set to zero, the sign ambiguity of \( \varepsilon \) does not spoil inversion symmetry.

34.7a Degeneracy of Maxwell equations

We found eight distinct equations, just like the usual form of the Maxwell equations. Previously we worried that the Maxwell equations are overdetermined, being eight equations in six unknown functions,26 but we found that the system of equations is singular: Two of the eight equations are tautologies, automatically satisfied regardless of what the fields and particles are doing. To see this again, more invariantly,

- Take the 4-divergence of the first set of equations and recall that \( \partial_\mu J^\mu = 0 \) identically. So one combination of these four equations is automatically satisfied.
- Apply \( \varepsilon^{\mu\nu\lambda\sigma} \partial_{\lambda} \) to the second set of equations and recall that partial derivatives commute. Here \( \varepsilon \) is the 4D analog of the Levi-Civita tensor, defined by a choice of orientation on spacetime and \( \varepsilon_{0123} = +1 \). Again, you find that one combination of these four equations is automatically satisfied (and the choice of orientation is immaterial).

A further reduction is possible if we use potentials (Section 34.8.1, page 535).

34.7b Spatial inversion (parity) invariance

One of our goals was to eliminate the Levi-Civita tensor from all of classical physics (Hanging Question #E, page 14). Chapter 15 advocated rephrasing electrodynamics by replacing \( \vec{B} \) by the antisymmetric rank-3 tensor \( \vec{\varepsilon} \), and indeed we see that the spatial block of Equation 33.6 (page 514) does just that. Then our manifestly-invariant forms of the Lorentz force law and Maxwell equations are also manifestly invariant under spatial inversions, because inversion is a particular kind of orthochronous Lorentz transformation and none of our constructions required a choice of right-hand convention.

---

26 Hanging Question #D.
34.7c Time-reversal invariance

Section 32.6.2 (page 497) pointed out that some physical laws contain dissipative processes, such as friction or electrical resistance, that violate time-reversal (or simply “T-”) invariance. Moreover, even some fundamental processes violate CP invariance, although not CPT; hence, they, too violate T.27 Finally, in quantum physics, T invariance is in any case treated separately from other symmetries: Unlike other symmetries, it is implemented by an antilinear operator.

For all these reasons, there is limited utility in preserving manifest T invariance, so we have supposed that a preferred time sense has been specified for spacetime,28 and only discussed orthochronous Lorentz transformations. However, we can still study specialized classical systems without dissipation, for example, the dynamics of a small number of point charges in vacuum, and ask in what sense such systems have an additional discrete symmetry.

The tricky aspect of this discussion is that time reversal is not merely a Lorentz transformation with \( \tau \rightarrow -\tau \). Indeed, when we apply that active transformation to a parameterized trajectory, the result is no longer forward-directed in time, as we have required.29 That is, \( d\tau /d\xi < 0 \). We must instead let

\[
[\Sigma(\tau)] = \Lambda[\Sigma(-\tau)],
\]

which incurs an additional minus sign in the 4-velocity \( U \), and in quantities that involve it.

Quantities like position \( X \), which just suffer an ordinary Lorentz transformation by \( \Lambda \), will be called “even” under time-reversal (or just “T-even”); quantities like \( U \) will be called “T-odd.” Each derivative with respect to proper time flips the even/odd status of a quantity. What can we then say about the electromagnetic field and potential? Inspection of the Lorentz force law shows that it sets a T-even quantity \( (\partial /\partial t) F \) equal to \( qE(U) \). Mass and electric charge are T-even scalars. Using the Lorentz force law to define \( F \) operationally therefore yields a tensor dependent on the chosen forward-time direction. Thus, the Faraday tensor is T-odd.

We now have no further freedom; we must see whether the Maxwell equations are T-invariant with the assignments just found. The homogeneous Maxwell equation (second of Equations 34.13, page 535) is straightforward: The sign change upon time reversal is immaterial.

To study the inhomogeneous Maxwell equation, we first need to investigate the charge 4-flux \( J \). Recall that:

- \( J^\tau = \text{total charge per } d\tau d\xi dz \text{ crossing } X^\tau = \text{const from } X^\tau < 0 \text{ to } X^\tau > 0 \), and so on.
- This definition depends on “from” and “to,” that is, on the sense of parameterization of the trajectories. So \( \dot{J} \) changes sign upon time reversal, unlike \( \dot{F} \). Similarly,
- \( J^0 = \text{total charge per } d^3r \text{ crossing } X^0 = \text{const from } X^0 < 0 \text{ to } X^0 > 0 \). Here there are two minus signs upon change of time orientation (“from”⇒“to” as well as exchange of past and future). So \( J^0 \) does not change sign, unlike \( X^0 \).

All told, \( J \) is T-odd. The inhomogeneous Maxwell equation thus sets a T-odd quantity (contraction of \( \partial F \)) equal to a T-odd quantity \( (\mu_0 J) \), so it is indeed a T-invariant equation.

Finally, \( F \) is built from derivatives of the four-vector potential \( A \), so \( A \) is also T-odd.

27For example, J. Cronin, V. Fitch, and coauthors found CP violation in the weak-interaction decay of kaons. CPT invariance is thought to be universally valid, so CP-noninvariance implies T-noninvariance.
28See Section 32.6.2 (page 507).
29See Sections 31.3.1 (page 477), 32.6.5 (page 501), and Section 32.6.2'.
34.8.1’a Counting equations, again
The main text arrived at four equations, Equation 34.16 (page 536). However, one degree of freedom in \( A \) is unconstrained by (drops out of) the equations, due to their gauge invariance. We may therefore worry that the remaining three degrees of freedom would be overdetermined by the four field equations. What rescues the equations is that one combination is automatically satisfied, as we see by taking the 4-divergence of both sides and using the continuity equation for \( J \).

34.8.1’b \( p \)-form gauge fields
In the language of Section 15.9c (page 236), the 4-vector potential is a 1-form field in four dimensions. Its field strengths are given by its exterior derivative. Therefore it is not completely determined by the fields; adding the exterior derivative of any 0-form leaves the field strengths unchanged because \( d^2 = 0 \).

Some exotic field theories derived from superstrings involve higher-rank antisymmetric tensor fields called “\( p \)-form gauge fields.” They, too, are subject to a field equation \( d F = 0 \), and so can be written as the exterior derivative of a \( (p-1) \)-form potential. Those potentials again have a gauge-invariance property, because adding the exterior derivative of any \( (p-2) \)-form to the potential leaves its exterior derivative unchanged.

34.9’ Geometric status of the charge flux
If you know a little differential geometry, then we can formulate charge flux more generally. Section 34.6.1 assumed that an inertial coordinate system had been chosen, so that the 3-volume \( \Delta^3 \mathbf{X} \) was defined. Now, we will not assume that any inertial coordinate system has been singled out, nor even that such systems exist at all, so our discussion will also apply in general relativity. We only suppose that an orientation on spacetime has been chosen, as well as a preferred time sense (Section 34.7c, page 543).

Start by defining a tensor field of rank (3) called \( \mathbf{J} \) that eats three 4-vectors and returns a number. To define it, construct a small parallelepiped (3-volume) using the given vectors as edges, each scaled by some small quantity \( a \). Some charged particle trajectories cross this 3-volume. For each, multiply the charge carried by that particle by \( \pm 1 \) depending on whether its tangent at the crossing completes the three vectors into a basis with the chosen spacetime orientation (or not). Sum the contributions, divide by \( a^3 \), and send \( a \to 0 \). This quantity is \( \mathbf{J} ( U, V, W ) \). It is indeed trilinear and antisymmetric in its three arguments. Section 15.9c (page 236) called such tensors 3-forms. \( \mathbf{J} \) describes the charges and their trajectories, and depends on the chosen orientation of spacetime, but not on the metric nor on any coordinate choice. Indeed, it continues to make sense in curved spacetime.

We can now define the 4-vector field \( \mathbf{J} \) as the contraction of \( \mathbf{J} \) with the Levi-Civita 4-tensor (Section 34.5’, page 542), an operation sometimes called the Hodge dual. Because \( \mathbf{J} \) also depends on a choice of orientation on spacetime, \( \mathbf{J} \) does not. Because \( \mathbf{J} \) involves the metric, so does \( \mathbf{J} \). In

\( \text{See Section 15.9’c (page 236).} \)

\( \text{Although the metric is used to define proper time, and hence the 4-velocity, any forward-directed tangent to the trajectory may be used when assigning the sign of each contribution. Note that } \mathbf{J} \text{ and } \mathbf{F} \text{ are both even under time reversal (Section 34.7’c, page 543).} \)
Track 2

545

fact, \( F \) is the 4-vector field defined in the main text.

Perhaps the most elegant version of the Maxwell equations is then

\[
d F = 0, \quad d \star F = \mu_0 (\star J),
\]

where \( \star \) in the second equation again denotes the Hodge dual operation and \( d \) is the exterior derivative. In vacuum, these equations are unchanged upon exchange of \( F \) and \( \star F \), an “electric–magnetic duality” transformation.

### 34.11’a Spinors

One of my life’s strongest emotional experiences related to science occurred when for the first time I understood Dirac’s equation.

— Abraham Pais

This section will depart from strictly classical electrodynamics to outline a generalization of the tensor concept. When quantized, the classical spinor fields that we will define here yield the relativistic theory of electrons, neutrinos, quarks, and their interactions.

The main text constructed tensors as multilinear functions of vectors. For example, a rank-three tensor eats three vectors and returns one number (Section 13.5, page 203), and so on. If we feed a rank-\( p \) tensor a collection of basis vectors associated to a particular coordinate system, then in \( D \) dimensions the resulting \( D^p \) numbers are its components in that coordinate system.

Chapter 32 pointed out that the components of a 3D tensor relative to two right-handed cartesian coordinate systems are related by a particular class of linear transformations: those belonging to the rotation group, \( \text{SO}(3) \), or built up by tensor products of \( p \) copies of a matrix in that group. We say that the components “transform” via a linear representation of the group \( \text{SO}(3) \). The simplest linear representation is the scalar; for example, charge transforms under a rotation \( S \) as \( q' = q \). The next simplest is the vector: \( [\vec{v}]' = S [\vec{v}] \). In fact, any group that is defined as a set of matrices has such a natural, or fundamental, linear representation, in which group elements act by ordinary matrix multiplication.

The components of higher rank tensors are then linear representations of the same group formed via tensor products of the fundamental one, in some cases (anti)symmetrized. The key theorem (which we did not prove) says that, up to equivalence, all linear representations of \( \text{SO}(3) \) can be decomposed into blocks obtained in this way.

Later, we graduated to four dimensions. Here we found that any two E-inertial coordinate systems that conform to a chosen orientation of space and time are related by the proper orthochronous Lorentz group \( \text{SO}^*(3,1) \). Hence, the components of a 4-tensor in such a system furnish a linear representation of that group, again starting with a fundamental representation we called “4-vector” (rank \( \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \)): \( [\vec{X}]' = A [\vec{X}] \). We also found another fundamental representation, which we called “4-covector” (rank \( \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \)), but this was equivalent to a 4-vector via index raising (Section 34.2.1, page 526). Then we built up more complex linear representations of the same group by tensor products, in some cases (anti)symmetrized. Again the key theorem (which we did not prove) says that, up to equivalence, all linear representations of \( \text{SO}^*(3,1) \) can be decomposed into blocks obtained in this way.

In the main text, square brackets are often used to flag when the components of a tensor object are being regarded as a row, column, or matrix; but in the rest of this section we will omit that
cumbersome notation for the generalized objects that we construct. Asterisk will represent complex conjugate of each entry. Dagger represents hermitian conjugate: $W^\dagger = (W^*)^\dagger$. If the hermitian conjugate equals the inverse, then $W$ is called a \textbf{unitary matrix}.

\textbf{34.11' Nonrelativistic spin}

Quantum mechanics requires a subtle extension to these ideas, because a wavefunction is not directly observable. In quantum mechanics, multiplying a state vector $|\Phi\rangle$ by an overall complex phase $e^{i\theta}$ does not change any observable, and so these are not physically distinct: They are both said to represent the same \textbf{ray} in state space. Hence, we can only expect that classical symmetry under a group $G$ will be reflected in a “ray representation” on the state space. Fortunately, any ray representation can be written as a true linear representation of a potentially extended form of $G$ called the “covering group.” In nonrelativistic quantum mechanics, when we study rotations then $G$ is $\text{SO}(3)$, whose covering group is $\text{SU}(2)$.\footnote{The special unitary $2\times2$ matrix group $\text{SU}(2)$ is sometimes instead called $\text{Spin}(3)$ in this context. Spatial inversions must be treated separately. The full group of rotations and galilean boosts presents interesting extra features (Bargmann, 1954).} These remarks motivate us to consider fields that transform as representations of SU(2), because when quantized, such fields will create more general particle states than the more familiar tensor-valued fields.

Any group defined as a set of complex matrices actually has \textbf{four} natural (fundamental) linear representations, in which group elements act by matrix multiplication:

\begin{equation}
\eta' = U\eta; \quad \chi' = U^*\chi \quad (34.28)
\end{equation}

\begin{equation}
\lambda' = (U^*)^{-1}\lambda; \quad \zeta' = (U^*)^{-1}\zeta. \quad (34.29)
\end{equation}

However, the second and fourth of these options are duplicates of the third and first, respectively, because $U$ is unitary, so we need not consider them. We now christen complex, 2-component vectors in the first representation as \textbf{3-spinors} of spin rank 1/2. We will now show that furthermore, the third is also equivalent to the first, leaving only one distinct fundamental representation in 3D.

To see the claimed equivalence, first let us introduce more explicit notation analogous to that in ordinary tensor analysis: In this section, indices from the start of the Greek alphabet will run over $\{1, 2\}$. The first representation is then

\begin{equation}
\eta^\alpha = U^\alpha_\beta \eta^\beta, \quad (34.30)
\end{equation}

whereas the third is distinguished from the first by index position:

\begin{equation}
\lambda^{\alpha\beta} = ((U^*)^{-1})^\alpha_\mu \lambda^\mu_\beta. \quad (34.31)
\end{equation}

Next, let $\epsilon$ denote the matrix $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ and write it more explicitly as $\epsilon^{\alpha\beta}$, where

\begin{equation}
\epsilon^{12} = +1. \quad (34.31)
\end{equation}

Next, notice that this new object has a useful property: $U^\epsilon U$ is again an antisymmetric $2 \times 2$ matrix, and hence has only one independent entry. For example, because the determinant of a matrix in SU(2) equals 1 we have

\begin{equation}
(U^\dagger)^1_1 \epsilon^{\alpha\beta} U^2_2 = U^1_1 U^2_2 - U^1_2 U^2_1 = \det U = 1 = \epsilon^{12}. \quad (34.32)
\end{equation}

Thus,

\begin{equation}
U^\epsilon U = \epsilon. \quad (34.33)
\end{equation}
We can now establish the claimed equivalence. Given a two-component complex 3-spinor $\eta_\alpha$ transforming as the first representation in Equation 34.28, let $\lambda^\alpha = e^{a\beta} \eta_\beta$. Then

$$\lambda' = \epsilon \eta' = \epsilon U \eta = (U^\dagger)^{-1} \epsilon \eta = (U^\dagger)^{-1} \lambda,$$

which is indeed the third transformation (left equation in Equation 34.29). So “index raising” via $\epsilon$ converts between these two representations, much as ordinary index raising converts between 4-vectors and 4-covectors. Indeed, Equation 34.33 shows that although $e^{a\beta}$ is defined as a set of four constants, it nevertheless transforms as a 3-spinor on each of its indices. Like the metric in tensor analysis, it is “from Heaven”—that is, rotationally invariant—and furnishes a standard mapping (equivalence) between representations.

Similarly to ordinary tensors, Equation 34.33 also implies that the contraction of two spinors $\eta^{(i)}_\alpha e^{a\beta} \eta^{(j)}_\beta$ is Lorentz-invariant, and motivates a new Rule, that we may “only contract upper with lower spin indices.”

To summarize, in 3D there is only one kind of rank-1/2 spinor. We may also build up more complex linear representations by tensor products, possibly (anti)symmetrized.

The key theorem (which we will not prove) says that, up to equivalence, all linear representations of the covering group can be decomposed into totally symmetric, $p$-fold tensor products of the fundamental spinor representation (“spin rank $p/2$”). Thus, we may construct spinor Rules paralleling those for 3-tensors. The ordinary 3-tensors appear as the integer-numbered entries on this list (they give ordinary representations of $\text{SO}(3)$). Those with half-odd spin rank are new (not encountered in classical physics).

To make the correspondence between SU(2) and rotations explicit, first note that any real 3-vector $\vec{v}$ corresponds to a traceless hermitian matrix, and vice versa, via $\vec{v} \leftrightarrow M = \vec{\sigma} \cdot \vec{v}$, where $\vec{\sigma}$ are the three Pauli matrices:

$$\sigma_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \sigma_2 = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \quad \sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Note that then

$$\det M = -||\vec{v}||^2.$$

For any special unitary matrix, $U M U^\dagger$ is traceless and hermitian with the same determinant as $M$, and linear in $\vec{v}$, so it corresponds to a new vector that’s a rotation of $\vec{v}$. This establishes a correspondence between SU(2) and SO(3) that preserves the product structures of the groups (and the inverse operation). However, the rotation specified by $U$ is the same as the one specified by $-U$, so the correspondence is 2-to-1: SU(2) “double-covers” SO(3).

Under rotations, a wavefunction of spin rank 1/2 has angular momentum eigenvalue $|| \vec{J} ||^2 = (3/4)\hbar^2$, so such a wavefunction can indeed represent a particle with spin 1/2 (such as an electron). The fact that this representation has two components accounts for the Stern–Gerlach observation.

Pauli’s theory went on to construct a rotationally-invariant Schrödinger equation, including spin effects based on the transformation properties we have outlined in this section. But we are after bigger game.
34.11’c Relativistic spin

The preceding discussion can be extended to the group of proper orthochronous\textsuperscript{35} Lorentz transformations, which Section 32.6.2 called $\text{SO}^+(3,1)$. Its covering group is $2 \times 2$ complex matrices, not necessarily unitary, but still with determinant one. That group is called the “special linear group of complex $2\times2$ matrices,” or\textsuperscript{36} $\text{SL}(2,\mathbb{C})$. Because we no longer impose the unitarity condition, this group is larger than $\text{SU}(2)$, as it must be to accommodate Lorentz boosts. To reduce confusion, this time we’ll use the letter $W$ to denote a generic element of the group.

Again there are four fundamental representations. Again the first of Equations 34.28 is equivalent to the first of Equations 34.29; the proof is the same as before, because $\det W = 1$ as in Equation 34.32. Similarly, the second representations in each group are equivalent. However, unlike in 3D our transformation matrices are not necessarily unitary, and so we cannot immediately conclude that the first of Equations 34.28 is just a duplicate of the second of Equations 34.29 and so on. We must therefore keep both of them, and distinguish them carefully. The traditional notational tool for this is to dot indices associated to one of these representations: For example, Equations 34.30 become

\[ \eta'_a = W^\alpha_b \eta_b ; \quad \chi'_a = (W^\alpha)_b \chi_b. \quad (34.35) \]

The distinction between the two transformation rules just given is not superficial like the one between up and down indices on ordinary 4-tensors (or on spinors): The two representations are not equivalent because there is no standard conversion from one type of index to the other.

Each representation gets its own spin tensor from Heaven: Equation 34.31 is augmented by

\[ \varepsilon^{\hat{a}\hat{b}} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \hat{a} \hat{b}. \quad (34.36) \]

The $\varepsilon$ tensors let us raise and lower spin indices as before, without changing their dotted/undotted status. We can also construct a Lorentz-invariant contraction between two dotted spinor indices, or two undotted indices, but not between dotted and undotted. Thus, our earlier 3D spinor Rule becomes “only contract upper with lower indices, and only if they are of the same dot-type.”

This time the key theorem (which we will not prove) says that, up to equivalence, all linear representations of the covering group can be decomposed into irreducible blocks, each of which is obtained as the totally symmetric tensor product of $n$ copies of the undotted representation, combined with the totally symmetric tensor product of $m$ copies of the dotted one (“spin rank $(n/2, m/2)$”). Thus, we may construct spinor Rules paralleling those for 4-tensors.

To set up the correspondence between $\text{SL}(2,\mathbb{C})$ and Lorentz transformations, this time note that any real 4-vector $\underline{X}$ corresponds to a hermitian matrix (not necessarily traceless)\textsuperscript{37} via

\[ \underline{X} \leftrightarrow M = -X^0 \mathbb{1} + \underline{X}_1 \underline{a}, \quad (34.37) \]

Moreover,

\[ \det M = -|\underline{X}|^2. \]

Let $W$ be any complex matrix with determinant equal to 1. Then $WMW^\dagger$ is again hermitian with the same determinant as $M$, and it is linear in $\underline{X}$, so it corresponds to a new 4-vector that’s a Lorentz transformation of $\underline{X}$. The correspondence we have set up between $\text{SL}(2,\mathbb{C})$ and $\text{SO}^+(3,1)$ preserves

\textsuperscript{35}“proper” means determinant $+1$; “orthochronous” means those transformations that do not reverse time; see Section 32.6.2. (Again, inversions must be treated separately.)

\textsuperscript{36}$\text{SL}(2,\mathbb{C})$ is sometimes instead called $\text{Spin}(3,1)$ in this context.

\textsuperscript{37}We already made use of this correspondence when constructing the Stokes parameters (Section 24.2.2, page 385).

November 14, 2023; Contents Index Notation Glossary
the product structures of the groups. But the Lorentz transformation corresponding to $W$ is the same as the one determined by $-W$, so the correspondence is 2-to-1: $\text{SL}(2, \mathbb{C})$ double-covers $\text{SO}^+(3,1)$.

The representations for which $n/2 + m/2$ is an integer correspond to ordinary 4-tensors. The others are new (not encountered in classical physics): They are generically called "4-spinor representations."

With this rather large mathematical framework in hand, we may now start to think about classical fields that, when referred to a particular $E$-inertial coordinate systems, have components that transform in a spinor representation. For example:

- A field with spin rank $(n/2, m/2) = (1/2, 0)$ or $(0, 1/2)$ (a Weyl spinor), when quantized, could represent massless chiral particles.
- The combination of a $(1/2, 0)$ and a $(0, 1/2)$ (jointly called a Dirac spinor) can represent electrons, muons, tau leptons, or quarks.
- The case $(n/2, m/2) = (1/2, 1/2)$ is equivalent to a four-vector field. The sum $n/2 + m/2$ is an integer, so this is an ordinary representation of Lorentz.
- An antisymmetric rank-2 tensor (such as the Faraday tensor $F^\mu_\nu$) can be split into a positive-helicity part, with spin rank $(1, 0)$, plus a negative-helicity part with $(0, 1)$. (These names arise because the Faraday tensor of a plane wave with circular polarization will belong to one or the other of these types, depending on its helicity.)
- Supergravity theories introduce massless fields with spin rank $(3/2, 0)$ and $(0, 3/2)$ (the "gravitino") and so on.

We can now generalize “Einstein thinking” to construct invariant differential equations as candidate spinor field equations, much as Section 34.5 did for the Faraday tensor. Here is one:

$$
\sigma^{\mu}_{\alpha\beta} \partial_{\nu} \chi^{\nu} = 0, \quad \text{Weyl equation} \tag{34.38}
$$

where $\sigma^\mu$ is the identity matrix and $\sigma^\nu$ are the Pauli matrices (compare Equation 34.37). Indeed, quantizing a spinor field that obeys Equation 34.38 yields states describing massless particles of spin rank $(0, 1/2)$.

A little more tinkering is needed to accommodate massive particles, such as electrons, because the left side of Equation 34.38 does not transform in the same way as $\chi$. We must introduce a second spinor field, $\eta$:  

$$
0 = -i \sigma^\mu_{\alpha\beta} \partial_{\nu} \chi^{\nu} + \bar{m} \eta_\alpha; \quad 0 = -i e^{\alpha\beta} \epsilon^{ij} \sigma^\mu_{\alpha\beta} \partial_{\nu} \eta_\alpha + \bar{m} \chi^i. \quad \text{Dirac equations} \tag{34.39}
$$

Substituting one of these equations into the other shows that each field separately satisfies the Klein–Gordon equation $\Box \chi = \bar{m}^2 \chi$. Like the Schrödinger equation, it is second-order in space derivatives. Unlike Schrödinger, the K–G equation is relativistically invariant; it is the appropriate generalization of the wave equation for a field associated to massive particles. The scalar constant $\bar{m}$ has dimensions of inverse length; it is related to particle mass by $m = \bar{m} \hbar/c$.

---

38 A further twist is that the component fields representing spin $1/2$, $3/2$, … must take their values in an anti-commuting number system (Section 40.4b, page 616).

39 To be clear, Einstein did not do this; relativistic spinors were developed by B. van der Waerden in 1929 following earlier work by E. Cartan.

40 Originally Dirac fused $\chi$ and $\eta$ into a single 4-component object, but today they are often separated into irreducible representations of $\text{SL}(2, \mathbb{C})$.
Opinions vary, but many physicists would say that the Weyl and Dirac equations have the surprising-yet-inevitable quality\footnote{Section 33.3.5 (page 516).} that we prize, to an even greater degree than do the Maxwell equations.

**Your Turn 34J**

This section may seem to have wandered far from electromagnetism, so suggest a simple modification to Equations 34.39 that incorporate interaction with a 4-vector potential. Why can’t you make a similar change to Equation 34.38?
34.1 Tensor types

An antisymmetric 3-tensor of rank 2 (such as the magnetic dipole moment tensor, Section 17.3.1, page 253) has only three independent entries. That is, apart from entries that are duplicates or that must equal zero, only 3 remain. Similarly, Section 33.3.1 (page 512) pointed out that an antisymmetric 4-tensor of rank $\binom{4}{2} = 6$ has independent entries.

a. How many independent entries has a symmetric 3-tensor of rank 2 (such as the momentum flux 3-tensor of a fluid, or the moment of inertia of a rigid body) got?

b. How many independent entries has a symmetric 4-tensor of rank $\binom{4}{2} = 6$ got?

c. How many independent entries has a totally antisymmetric 3-tensor of rank 3 got? (Equation 15.10, page 226 introduced one such object.)

d. How many independent entries has a totally antisymmetric 4-tensor of rank $\binom{4}{2} = 6$ got? (Equation 34.6, page 532 introduced one such object.)

e. How many independent entries has a totally symmetric 3-tensor of rank 3 got?

f. How many independent entries has a totally antisymmetric 3-tensor of rank 4 got?

g. How many independent entries has a totally antisymmetric 4-tensor of rank $\binom{4}{2} = 6$ got?

34.2 Uniformly moving charge revisited

A charged point particle moves in a straight line with constant speed. It creates electric and magnetic fields. In this problem you'll find them directly via “Einstein thinking,” instead of using the transformation approach in Sections 33.4.2 (page 518) or 34.8.2.

a. Find a manifestly invariant expression for the Faraday tensor by using the following Hints:

- By translation invariance, it’s enough to find $F_{\mu\nu}$ only for an observer at $X_0 = 0$.
- The particle’s trajectory can be written in parametric form as $X_+ = \tau U$ for constant four-vectors $X_+$ and $U$.
- Replacing $X_+$ by $X_+ + aU$ does not change the trajectory, for any constant $a$.
- Your formula should be an antisymmetric rank $\binom{3}{2} = 3$ tensor constructed out of the four-vectors $U$ and $X_+$ using The Rules.
- Your result must reduce to Coulomb’s law if the particle is at rest in the chosen inertial coordinate system.
- The combinations

$$K^{\mu\nu} = (U^\mu X_+^\nu - (\mu = \nu)) \quad \text{and} \quad \|K\|^2 = K^{\mu\nu}K_{\mu\nu}$$

are useful building blocks for your answer, in part because the latter is equal to something useful when computed in the rest frame of the particle.

b. Specialize to a trajectory moving along the $x$ direction and check that your result is equivalent to the ones in Sections 33.4.2 or 34.8.2.
34.3 Bremsstrahlung I

A positively charged particle is initially in uniform motion along the $x$ axis at speed $0.9c$. At time zero, it abruptly comes to a halt. An observer later measures the electric field at time $t_0 > 0$ throughout the $xy$ plane.

Section 25.3 (page 392) argued that when $r < ct_0$, the observer will see the 4-vector potential of a charge at rest; farther away, however, the observer has not yet “learned” about the deceleration. Hence, close to the particle, the observer sees the usual $1/r^2$ field of a point charge in an inertial coordinate system in which it’s at rest. But far from the particle, the observer sees a field that is crowded into the $yz$ plane and centered on the point where the particle would have been located, had it continued to move (Idea 33.12, page 519).

In the following steps, express all lengths as dimensionless quantities times $ct_0$; then the actual value of $t_0$ will turn out to be immaterial.

a. Make a grid of points at which to evaluate the 4-vector potential. The grid should be fine enough to get reasonably accurate estimates of derivatives by numerical differentiation.

b. Evaluate the 4-vector potential at each of the grid points satisfying $r < ct_0$.

c. Use ideas from Section 34.8.2 (page 536) to evaluate the 4-vector potential at every grid point not satisfying that condition.

d. Repeat (a–d) for later time $(1.001)t_0$, noting that the boundary between the two regions has shifted slightly. Subtract from your previous answer and divide by 0.001 to estimate the time derivative of $\vec{A}$ throughout the $xy$ plane.

e. Do whatever else you need to do to find the electric field at time $t_0$.

f. Get a computer to make a graphical depiction of the magnitude $||\vec{E}(t_0, x, y, 0)||$ over an interesting region of the $xy$ plane. If the range of field values attained is too large to display properly, compress it by taking a logarithm before making the plot.

g. Repeat for initial speed $0.1c$ and comment.

[Hints:]

- $\vec{E}$ points in the $xy$ plane, so a two-dimensional plot is adequate.
- The 4-vector potential field is discontinuous, so you won’t get a very accurate result by numerically differentiating it. However, you do get the right qualitative behavior. This problem is a pathology related to the unrealistic assumption that the charge stops instantly (that is, infinite deceleration).
- Ensure that your plot cover a range of $xy$ values large enough (and also small enough) to show the interesting features. Make sure your computer uses the same scale for the both axes.
- If you wish, you can compare your result to the more complicated formulas in Your Turn 33C, but that’s not the approach you are to use in this problem.
- For the plots in (g,h), you may make a heat map, a contour plot, or a surface plot. Use

---

42 Python users will find useful information in Kinder & Nelson, 2021, §6.4.1, or in the builtin help for numpy.meshgrid.
your judgment about what is clearest. (Why don't you need to know the values of $t_0$ and $q$?)
Initially, Einstein was not impressed [by Minkowski’s formulation] and regarded the transcriptions of his theory into tensor form as “überflüssige Gelehrsamkeit” [superfluous erudition]. However, in 1912 he adopted tensor methods and in 1916 acknowledged his indebtedness to Minkowski for having greatly facilitated the transition from special to general relativity.

—Abraham Pais

35.1 FRAMING: LOCAL CONSERVATION

So far, Chapters 33 and 34 just reformulated old laws, but now it’s time for something more ambitious. We no longer believe that space is filled with gears, pulleys, rubber bands, and so on that carry the EM fields, so we can’t write down any functions for energy and momentum based on intuitions gleaned from mechanics. Instead, we hope to prove a theorem about our system of equations stating that certain quantities are locally conserved. Those quantities include familiar bits corresponding to energy and momentum of point particles, as well as others associated to fields. To get started, we need a good guess for what the field contributions might be. As we have seen in several previous situations, “Einstein thinking” will focus our attention down to such a small space of possible expressions that we can check all (two) of them exhaustively.

This energy is real and can be substantial. In a large magnet, it can be released suddenly (Problem 35.1).

Electromagnetic phenomenon: Superconducting magnets can fail catastrophically. Physical idea: Even a static magnetic field stores energy proportional to volume, and also to magnetic field strength squared.

35.2 WHAT NEEDS TO BE SHOWN AND WHY

- Chapter 6 computed the work that must be done to charge a capacitor. That energy isn’t lost—you can get your investment back. Where is that energy in the meantime? We got a hint: It’s proportional to the volume occupied by electric field. Maybe it’s in the empty space between the capacitor plates.
- Similarly, Chapter 18 computed the work that must be done to set up a current in a coil of wire. If the wire is superconducting, then the energy is not lost—you can get
your investment back. Where is that energy in the meantime? We found that it, too, is proportional to the volume. Maybe it, too, is in the empty space inside the coil.

That is, our hypothesis is that empty space itself can store the energy of static electric and magnetic fields. We need to make that more general and precise.

- Chapter 20 studied nonstatic situations and found energy and momentum fluxes. Here again, we found them to be quadratic in the field amplitudes, although we didn’t yet get the constant of proportionality: We just found how much of the energy and momentum could be extracted by a particular charged test body.
- Finally, Section 25.5.3 found that the energy from our special-case formulas can be irretrievably sent out to infinity by an antenna, regardless of whether there are any receivers to recapture it. Long ago, Hanging Question #H (page 32) asked, “what carries that energy?”

In short, it’s been an ad hoc approach until now. Now that we have unified \( \vec{E} \) and \( \vec{B} \), now that we have unified energy and momentum, it’s time for one big result that covers all these Electromagnetic Phenomena at once. To get it, we’ll generalize the discussion of waves on a string (Chapter 27). We found formulas for energy flux and density, and momentum flux and density. (They, too were quadratic in the amplitude.) Then you proved continuity equations expressing local conservation of energy and momentum.\(^1\) We’ll now attempt the same thing with EM fields.

Using “Einstein thinking,” the strategy will be: Find a family of expressions that all take the form of the sum of the particles’ \( E_\mu \Gamma^\mu \) plus a quadratic function of fields with appropriate tensor properties. Requiring that the expression must also obey a continuity equation then nails down its exact form. Then the field term, whatever it turns out to be, will deserve to be called the “energy and momentum of the fields,” and its continuity equation will be the local conservation law that we wanted to prove. We’ll see that indeed, energy and momentum can slosh locally back and forth between fields and particles, while staying conserved overall.

Certainly the tensor structure will be more complex than in the string/spring metaphor. That’s one reason why we invented our big language in Chapters 32–34.

### 35.3 Continuity Equation for Energy and Momentum in the Absence of Long-Range Forces

First consider a swarm of particles with no external forces and no long-range internal forces. Pointlike collisions are allowed; we suppose that each collision locally conserves energy and momentum, much as we assumed in Chapter 8 that collisions locally conserve electric charge (Figure 8.2). Between collisions, each trajectory \( \Gamma_\varphi(\tau) \) is therefore a straight line, which we parameterize by proper time. Analogously to the charge flux

---

\(^1\)Your Turns 27C (page 425) and 27D.
Chapter 35  Energy and Momentum of Fields

Figure 35.1: **Graphical understanding of Equation 35.1.** The figure is the same as Figure 34.1 (page 533). Again the $z$ axis has been suppressed for visualization. (a) Particle #1 contributes to the density of charge or 4-momentum at the point shown (that is, to $J^0$ or $T^{0\nu}$ at $(ct, \vec{r})$), whereas #2 does not. (b) Particles #1 and #2 contribute to the net flux of charge or 4-momentum (that is, to $J^2$ or $T^{2\nu}$), whereas #3 and #4 do not. #1 crosses from smaller to larger $y$ and hence contributes $q_1/\Delta x\Delta y\Delta (ct)$ or $p_1^{\nu}/\Delta x\Delta y\Delta (ct)$. #2’s contributions have an extra minus sign.

4-vector $J$, define the **energy–momentum flux 4-tensor**$^2$ by a recipe analogous to Equation 34.9 (page 534):

\[
T^{\mu\nu} = \text{net amount of } p^\nu \text{ crossing the surface } X^{\mu} = \text{constant, from smaller to larger } X^{\mu}, \text{ per } d^3 X_1, \text{ times } c. \tag{35.1}
\]

**Your Turn 35A**

Using Figure 35.1, convince yourself that

\[
T^{00} = c \times \text{density of (energy}/c) \\
T^{10} = \text{net flux of (energy}/c) \\
T^{0k} = c \times \text{density of the } k \text{ component of momentum} \\
T^{ik} = \text{net flux along } i \text{ direction of the } k \text{ component of momentum.}
\]

Why do you suppose that some books call $T$ the “stress-energy tensor?”

The last of those results says that the space-space components of $T$ constitute a 3D momentum flux tensor (page 200).

We will call the the energy–momentum flux tensor carried by particles $T_{\text{part}}$, and write an equivalent formula like the one used for $J$ in Equation 34.23 (page 538): Just replace the charge on particle $\ell$ by the 4-momentum on particle $\ell$ at proper time $\tau$:

\[
T^{\mu\nu}(X) = \sum_{\ell} \int_{-\infty}^{\infty} \text{d} \tau \ p_{(\ell)}^{\nu} (\tau) U^{\mu}_{(\ell)} (\tau) \delta^{(4)} (X - \Gamma_{(\ell)} (\tau)). \tag{35.2}
\]

---

$^2$Often abbreviated “energy–momentum tensor.”
35.4 Interactions Seem to Spoil Local Conservation

Figure 35.2: Long-range repulsion of two particles (wavy arrow) would spoil local conservation of particle energy and momentum. Net \( \vec{p}_2 \) appears in the small box due to force exerted on particle \#2 from outside the box, while it is inside.

Your Turn 35B

Using Equation 35.2, convince yourself that \( T \) is a symmetric 4-tensor of rank \((\frac{3}{0})\). Still assuming that the only interactions are local collisions that each conserve energy and momentum, show that \( T_{\text{part}} \) obeys

\[
\frac{\partial}{\partial x_i} T_{\text{part}} = 0. \quad \text{no long-range or external forces} \quad (35.3)
\]

That is, \( T_{\text{part}} \) obeys four continuity equations, expressing the local conservation of each component of the 4-momentum.

35.4 INTERACTIONS SEEM TO SPOIL LOCAL CONSERVATION

35.4.1 Long-range forces

If external forces act on our particles, then we don’t expect their energy or momentum to be conserved: A falling body accelerates (gains momentum). Even mutual forces, if they act at long range, would destroy local conservation: Two distant positive charges, initially at rest, start to accelerate away from each other, so equal and opposite amounts of momentum seem to appear from nowhere at two distant locations (Figure 35.2).\(^3\)

Sections 2.5.1 and 18.11 suggested that there is something else in the box: We must introduce an entity called the “electromagnetic field” in order to rescue locality. Then the repulsion of two particles involves each one getting momentum locally (from the field nearby), and so on. It is time to deliver on this promise, by correctly attributing energy and momentum to fields as well as to particles. It’s not obvious that this can be done consistently. Let’s begin by getting quantitative about the preceding paragraph.

Adapting our proof of the continuity equation (Chapter 8), we again draw a small four-dimensional box (hypercube) and ask how much net momentum enters it by particles crossing its faces (Figure 8.2). As with electric charge,\(^4\) that net change will equal \((-c^{-1} \partial_{\mu} T_{\text{part}}^{\mu}) (\Delta^4 X)\), regardless of any collisions among particles in the box (for example,

\(^3\)Also, each gets not-opposite amounts of kinetic energy, again seemingly from nowhere.

\(^4\)See Equation 8.3 (page 117).
the disintegration shown in the cartoon). Unlike that case, however, this time the net change need not be zero, because energy and momentum can flow across the box walls by some means other than being carried along particle trajectories. Even if a particle does not collide with anything, it is acted on by fields throughout its sojourn in the box.\(^5\)

\[
\Delta_{\text{box}} P^\nu = \text{net } P^\nu \text{ into 4-box} = - \sum_\ell \int_{\tau_{\text{in},\ell}}^{\tau_{\text{out},\ell}} \frac{dP^\nu(\tau)}{d\tau} \delta^{(4)}(X - \Gamma(\tau)\ell).
\] (35.4)

In this formula, we only include those trajectories that actually enter the box; \(\sum_\ell\) denotes the restricted sum. Moreover, we only include the part of each particle's trajectory that is actually spent inside the box. That explains the limits on the \(\tau\) integral. Finally, we only need to include the contributions to \(dP(\tau)/d\tau\) arising from electromagnetic forces on the particles. Although there can also be collisions inside the box involving short-range forces, these locally conserve 4-momentum and so cancel in Equation 35.4.

### 35.4.2 Nonconservation of particle energy and momentum

We now use the Lorentz force law to relate the right side of Equation 35.4 to the fields. As written, however, Equation 35.4 is cumbersome, because of the restricted sum and integral. To make it easier to work with, we now make the unobvious step of multiplying by one, using the identity

\[
1 = \int d^4X \delta^{(4)}(X - X_\ast) \quad \text{for any point } X_\ast \text{ in spacetime.}
\]

For each term \(\ell\) and each value of \(\tau\), make the choice \(X_\ast = \Gamma(\tau)\ell\). Then we move the integration over \(X\) all the way to the left (do it last):

\[
\Delta_{\text{box}} P^\nu = - \int d^4X \sum_\ell \int_{\tau_{\text{in},\ell}}^{\tau_{\text{out},\ell}} \frac{dP^\nu(\tau)}{d\tau} \delta^{(4)}(X - \Gamma(\tau)\ell).
\]

This looks like it’s making our formula more complicated, but now note what happens if we restrict the \(X\) integral to just our little box (hypercube). Then the delta function automatically selects only the trajectories that pass through the box, so we don’t need to restrict the sum. And the delta function also automatically selects only those \(\tau\) values for which a trajectory lies inside the box, so we don’t need to restrict the \(\tau\) integral either.

Using that insight, and the Lorentz force law (Equation 33.3, page 512), gives

\[
\Delta_{\text{box}} P^\nu = - \int d^4X \sum_\ell \int_{-\infty}^{\infty} d\tau \, q_\ell \, F^{\nu\lambda}(\Gamma(\tau)) U(\Gamma(\tau))\delta^{(4)}(X - \Gamma(\tau)\ell).
\]

Use the delta-function to re-express the factor in the brace as \(F^{\nu\lambda}(X)\), and then push it to the left, outside of the \(\tau\) integral. The integral is then \(c^{-1}\) times the electric charge flux four-vector (Equation 34.23, page 538):

\[
\Delta_{\text{box}} P^\nu = -c^{-1} \int_{\text{box}} d^4X \, F^{\nu\lambda}(X) U_\lambda(X).
\] (35.5)

---

\(^5\)To understand the minus sign in Equation 35.4, note that if a particle gains momentum while in the box, then it transports more out when it exits than it had upon entry.
We have now expressed the net change of momentum in the box in terms of electromagnetic fields and the charge flux 4-vector.

For a small enough box, we may approximate the integral as $\Delta^4X$ times the integrand. But Section 35.4.1 argued that this change is also $c^{-1}\Delta^4X$ times minus the 4-divergence of $T$, or

$$\frac{\partial}{\partial t} T_\text{part}^{\mu\nu} = F^{\nu\lambda}J_\lambda.$$  \hfill (35.6)

Because the right side need not equal zero, this formula makes precise what was argued qualitatively before: The energy–momentum flux tensor of particles only does not obey a continuity equation, if long-range forces are present.

### 35.5 ACCOUNTING FOR FIELD CONTRIBUTIONS RESTORES LOCAL CONSERVATION OF ENERGY AND MOMENTUM

Rather than give up, we are hoping to find another contribution to the total energy–momentum flux tensor of the world, attributing 4-momentum to electromagnetic fields, with the properties that:

- $T^{\mu\nu}_\text{field}$ is a symmetric 4-tensor given by a local expression in the fields; and
- $\frac{\partial}{\partial t} \left( T^{\mu\nu}_\text{part} + T^{\mu\nu}_\text{field} \right) = 0.$

That is, we want to find a contribution to the energy–momentum flux tensor depending only on fields and with the property that the total $T^{\mu\nu}$ obeys a continuity equation. Once we prove it, that continuity equation will be a Lorentz-invariant formulation of the local conservation of total energy and momentum.

Equation 35.6 shows what we need:

$$\frac{\partial}{\partial t} T^{\mu\nu}_\text{field} = -F^{\nu\lambda}J_\lambda.$$  \hfill (35.7)

But we can’t prove this until we guess the correct formula for $T^{\mu\nu}_\text{field}$!

To get past this impasse, let’s apply “Einstein thinking.” What sorts of symmetric, rank-two tensors can we build from the Faraday tensor? We already have some anecdotal evidence that stored electrostatic energy is a quadratic function of electric field, with no derivatives ($\propto \vec{E}^2$, Equation 6.2, page 78). And stored magnetic energy is also a quadratic function of magnetic field, with no derivatives ($\propto \vec{B}^2$, Equation 18.14, page 272). Can we write any such expression that is a symmetric, rank-$(\frac{2}{2})$ tensor?

In fact, The Rules allow us to write just two such expressions. Rather than choose one or the other, we must keep our options open and suppose that the tensor we are seeking is some linear combination of them both:

$$T^{\mu\nu}_\text{field} = \alpha F^{\mu\sigma} F_\sigma^\nu + \beta \sigma^{\mu\nu} F_\sigma^\xi F_\epsilon^{\sigma\xi}. \quad \text{provisional formula}$$  \hfill (35.8)

Indeed, the expression above is a tensor of the right rank and symmetry that’s quadratic in fields and has no derivatives. We don’t know the values of $\alpha$ and $\beta$ yet, but already

---

6Sometimes called “Poynting’s theorem,” although independently discovered by Heaviside.
we’ve made a huge simplification: Just those two numbers are all the freedom we have to construct a suitable tensor.

We now take the 4-divergence of our provisional formula:

\[ \partial_{\mu} T^{\mu\nu}_{\text{field}} = \partial_{\mu} (\alpha F^{\mu\sigma} F_{\sigma}^{\ \nu} + \beta g^{\mu\nu} F_{\sigma\lambda} F^{\sigma\lambda}). \]

Use the fact that the fields obey the Maxwell equations, specifically the first of Equations 34.13 (page 535):

\[ = \alpha (-\mu_0^2 F^{\sigma} F_{\sigma}^{\ \nu} + F^{\mu\sigma} \partial_{\mu} F_{\sigma}^{\ \nu}) + \beta g^{\mu\nu} 2(\partial_{\mu} F_{\sigma\lambda}) F^{\sigma\lambda}. \] (35.9)

The first term (first brace) is just what we want! Just choose the value \( \alpha = -\mu_0^{-1} \) and we get Equation 35.7.

We are left with the unwanted other terms (second brace). Can we choose a value of \( \beta \) such that these terms cancel each other identically? That is, can we ensure that

\[ 0 = \frac{\gamma}{\beta} F^{\mu\sigma} \partial_{\mu} F_{\sigma}^{\ \nu} + 2 F^{\sigma\lambda} \partial_{\nu} F_{\sigma\lambda}. \] (35.10)

It’s not as crazy as it sounds, because so far we have only used half of the Maxwell equations to obtain Equation 35.9. The other half indeed say that something involving first derivatives of \( F \) equals zero.\(^7\) Specifically, the quantity enclosed by the brace in Equation 35.10 equals

\[ \partial_{\nu} F_{\sigma\lambda} = -\partial_{\sigma} F_{\lambda\nu} - \partial_{\lambda} F_{\nu\sigma}. \]

In Equation 35.10, this tensor is contracted on \( \sigma\lambda \) with something antisymmetric, so we may replace its first term by \( +\partial_{\sigma\lambda} F_{\nu\sigma} \). Then Equation 35.10 becomes

\[ 0 = \frac{\gamma}{\beta} F^{\mu\sigma} \partial_{\mu} F_{\sigma}^{\ \nu} + 2 F^{\sigma\lambda} (\partial_{\sigma} F_{\nu\lambda} - \partial_{\lambda} F_{\nu\sigma}). \]

Your Turn 35C

Rearrange the last formula to the form

\[ 0 = (\frac{\gamma}{\beta} - 4) F^{\mu\sigma} \partial_{\mu} F_{\sigma}^{\ \nu}. \]

Clearly the identity, and hence Equation 35.10, will hold if we choose \( \beta = \frac{\alpha}{4} \).

Substituting the values we found for \( \alpha, \beta \) into Equation 35.8, we conclude that

\[ T^{\mu\nu}_{\text{field}} = -(\mu_0)^{-1} (F^{\mu\sigma} F_{\sigma}^{\ \nu} + \frac{1}{4} g^{\mu\nu} F_{\sigma\lambda} F^{\sigma\lambda}). \text{ energy–momentum flux tensor of the electromagnetic field} \] (35.11)

This local function of fields meets all the criteria listed at the start of this section.

\(^7\)See the second of Equations 34.13 (page 535).
It is useful to express the four main blocks of Equation 35.11 in the traditional 3D symbols:

\[
T^\mu_\nu^{\text{field}} = \left[ \frac{u}{[\vec{S}] / c} \right]^{\mu\nu},
\]

(35.12)

where \( u = \frac{1}{2} (\varepsilon_0 E^2 + \mu_0^{-1} B^2) \) is energy density and \( \vec{S} = \vec{E} \times \vec{B} / \mu_0 \) is the flux of energy, or \( c^2 \) (density of momentum), also called the Poynting vector.

**Your Turn 35D**

a. Confirm that the 00 component has the form that you expect from Sections 6.3 (page 78) and 18.3.4 (page 272).
b. Then show that its 10 components are consistent with the form anticipated in Section 20.4 (page 302).
c. Why didn’t we need to consider terms that were linear in \( \vec{E} \) in Equation 35.8?

Thus, the formulas that we already guessed for field energy density and Poynting vector need no corrections to account for relativity.\(^8\)

The \( ij \) components may be new to you; they will help us interpret field lines in Chapter 36 and later play a key role for radiation pressure (Section 20.2.3, page 301).

### 35.6 WHAT HAS BEEN ACCOMPLISHED

#### 35.6.1 Poynting’s theorem fits with older ideas

We can now get global conservation laws in the usual way, by integrating \( T_{0\mu} \) over space (see Equation 8.6, page 118).

At a single stroke, our discussion established the local conservation not only of energy, but also of all three components of momentum, via one 4-vector result: Equations 35.6, 35.9, and Your Turn 35C imply that\(^9\)

\[
\partial_\mu \left( T^{\mu\nu}_\text{part} + T^{\mu\nu}_\text{field} \right) = 0. \quad \text{Poynting theorem}
\]

Take a moment to appreciate how surprising this formula is, and how it vindicates Michael Faraday’s intuitions:

- \( T^{\mu\nu}_\text{part} \) is just the mechanical part of the (relativistic) energy and momentum density of charged particles, exactly the same as it would have been without any electromagnetic interactions.

---

\(^8\)Beware that some books introduce a “Maxwell stress tensor” that is minus the space–space part of their “symmetric stress tensor.” In this book, the space–space part of the energy–momentum flux 4-tensor will always be called “momentum flux 3-tensor.”

\(^9\)This was Hanging Question #H.
• $T_{\mu\nu}^{\text{field}}$ locates all electromagnetic contributions in the empty space surrounding the particles and attributes it to the fields themselves.

This viewpoint is such a radical departure from earlier ideas that it’s worth exploring in a simple situation: two identical charged particles at rest. In a first-year class you may have heard that the particles “mutually exert forces on each other” and that the work done against that force to assemble the charges is preserved as potential energy, and can be recovered by later releasing them (letting them fly away from each other). Implicitly the framework is that #1 acts at a distance on #2, creating a potential energy well, a bit similarly to if there were a mechanical spring between them that is initially compressed. But this view begins to struggle when particles are so far apart that each one’s influences take appreciable time to be felt by the other (Hanging Question #H, page 32).

In contrast, now we are saying:

• Each charged particle influences the field in its immediate vicinity (Maxwell equations are local).
• The field next to a charged particle influences the field slightly farther away, and so on throughout space.
• The energy of the complete distribution is the sum of independent contributions from each volume of space.

We can now see how the older viewpoint could be compatible with the new one. We know the electric potential everywhere from Coulomb’s law and superposition, so we have

$$\frac{\varepsilon_0}{2} \int d^3r \|\vec{E}\|^2 = \frac{\varepsilon_0}{2} \int d^3r \|\vec{\nabla}\psi\|^2 = -\frac{\varepsilon_0}{2} \int d^3r \psi \nabla^2 \psi. \quad (35.13)$$

But $\nabla^2 \psi = -\rho_q/\varepsilon_0 = -q(\delta^{(3)}(\vec{r} - a\vec{z}/2) + \delta^{(3)}(\vec{r} + a\vec{z}/2))/\varepsilon_0$. Multiplying this expression by $\psi$ yields four terms on the right side of Equation 35.13. Two of those terms are infinite, but independent of $a$. The other two give

$$\frac{1}{2} \frac{q^2}{4\pi\varepsilon_0} (a^{-1} + a^{-1}),$$

which equals the potential energy from first-year physics.

Our formulas for energy density, energy flux, and momentum flux recover what we found informally in earlier chapters (Your Turn 35D), but now we have more:

• Our earlier explorations merely claimed that electrodynamics was compatible with energy conservation. For example, Section 6.3 argued that a certain amount of work was needed to charge a capacitor, and the corresponding stored energy could be recovered by discharging it, suggesting that that energy must be stored in the field. Similarly, Section 18.3.4 argued that a certain amount of work was needed to establish a current in an inductor, and could be recovered by ramping down the current, again suggesting that energy must be stored. And Section 20.4 argued that light imparts energy, suggesting that this energy must initially be carried by the light. Now we have proved local conservation as a property of the Maxwell equations and the Lorentz force law.
Previously we didn’t show that our expressions had the appropriate Lorentz transformation properties. Now it’s manifest because we followed The Rules.

Previously we only got expressions for energy and momentum flux limited to the case of plane waves. Moreover we didn’t find the correct prefactor. Now we have complete and general formulas.

Finally, the same strategy will also give us an analogous theorem when we later add media in Chapter 53.

Section 35.6.1’ (page 564) discusses other idealized circuit elements in the light of our new ideas.

35.6.2 Magic without magic

It may seem that we have cheated! After all, we just cooked up a quantity precisely so that it would give $2 \mu \gamma_{\mu \nu} = 0$, so what has been proved? But it was highly significant that any such formula could be written at all. The only cookery allowed was the choice of two constants, $\alpha$ and $\beta$, but the theorem we proved was that four functions of space and time are everywhere zero—a potentially infinite number of constraints.

Indeed, it seems magical that our highly constrained guess, Equation 35.8, could be adjusted to satisfy the continuity equation. Chapter 40 will rediscover the energy and momentum conservation laws as consequences of the translational invariance of the Lagrange function giving rise to the Maxwell equations.

FURTHER READING

Intermediate:

Angular momentum flux tensor: Zangwill, 2013, §22.7.4; Weinberg, 1972, p. 46.
35.5’a Angular momentum flux tensor
Any first-year physics book claims to prove that for an isolated system of particles, overall angular momentum is conserved. But we must do better than that. Imagine two isolated opposite charges orbiting each other. Due to their centripetal acceleration, they will emit radiation out to infinity that carries some of their angular momentum (and kinetic energy), slowing the orbit. Similar remarks apply to two charged particles approaching each other on a near-collision course. The main text showed that nevertheless, energy is conserved if we include the field contribution. What about angular momentum? Define the rank-(1,0) tensor field
\[ M^{\nu\lambda} = X^{\nu\lambda\mu} - X^{\lambda\nu\mu}. \]
It’s antisymmetric on \( \nu \) and \( \lambda \), and you can readily show that \( \delta_\mu M^{\nu\lambda} = 0 \). Thus, we get six densities by taking \( \mu = 0 \), leading to six conserved quantities
\[ L^{\nu\lambda} = \int d^4r M^{\nu\lambda}(t, \vec{r}). \]
The spatial bits of this tensor, \( L^{\nu\lambda} \), are the relativistic version of the angular momentum, and we have just shown that they are conserved when we include both particle and field contributions to \( T \). So it’s appropriate to call \( M \) the angular momentum flux tensor.

We may nevertheless worry that the orbital decay phenomenon dictates an “arrow of time,” even in problems that have no dissipative elements. But on the contrary, two orbiting charges can also gain energy and angular momentum if they interact with an incoming, circularly-polarized, wave.

Similar though more complex ideas emerge in gravitation. Two orbiting neutron stars constantly accelerate, which can generate gravitational radiation, leading to a decaying orbit and ultimate merger. Gravitational radiation with the time course characteristic of this process was detected on Earth by the LIGO-Virgo collaborations in 2016.

35.5’b Other tensor structures
Equation 35.8 gave what was claimed to be a complete list of terms with acceptable tensor structure to enter in the energy-momentum flux tensor. Then we saw that indeed these two terms sufficed to find a conserved \( T^{\mu\nu} \). One may ask, however, about two other possible terms involving the Levi-Civita tensor (Section 34.5’, page 542):
\[ \frac{1}{2}(\xi^{\mu}_{\nu\lambda\delta}F^{\lambda\mu\nu}F^{\nu\lambda\mu} + \xi^{\mu}_{\nu\lambda\delta}F^{\lambda\mu\nu}F^{\nu\lambda\mu}) \quad \text{and} \quad \xi^{\mu\nu}_{\nu\lambda\delta}F^{\lambda\mu\nu}F^{\nu\lambda\mu}. \]
Neither of these is zero; for example, the second one is \( g \) times \(-6\vec{E} \cdot \vec{B} / c \). Nevertheless, we know that they will not be needed because they are odd under spatial inversion, and also under time reversal, unlike, say, \( T^{00} \).

35.6.1’ Connect to other idealized circuit elements
The discussion in the main text recalled how our new ideas fit with the behavior of inductors and capacitors. Some further thought is needed to connect to other idealized circuit elements from first-year physics.

November 14, 2023; Contents Index Notation Glossary
Resistors

If \( \int d^3r \left( T_{\text{part}}^0 + T_{\text{field}}^0 \right) \) is always conserved, then where does it go when a resistor or other ohmic element “dissipates” it? To answer this, remember that the first term includes the kinetic energy of all matter, not just the charge carriers in a conductor. Dissipation involves the collision of charge carriers with everything in their way, transferring kinetic energy out to the surroundings as heat.

Batteries

Understanding an electrochemical cell (“battery”) really does require that we acknowledge an additional contribution to the energy density, from the chemical energy of its ingredients. Predicting such energies is ultimately a quantum mechanical calculation, and hence outside the domain of classical electromagnetism, but still we fit it into our framework in Section 10.3.4f (page 158). For example, when we connect a capacitor across the battery, that circuit is momentarily out of electrochemical equilibrium; current flows (limited by the battery’s internal resistance) until the potential drop across the capacitor matches the potential gain across the battery before it was connected. By that point, energy stored in the capacitor equals chemical energy lost by the battery (minus any resistive dissipation along the way to equilibrium).
35.1  **Boom 2008**

The Large Hadron Collider project at CERN suffered something of a setback in October 2008, when, during a test of one of the quadrupole magnets, the magnet failed catastrophically. The resulting “event” lifted a 20-ton magnet off its mountings, filled a tunnel with helium gas, and forced an evacuation (Figure 35.3).

The problem is that a big superconducting magnet stores a lot of magnetic field energy. If any bit of that magnet stops being superconducting, then suddenly the huge electric current generates a lot of heat, converting more of the magnet coil to the nonsuperconducting state. Very quickly, all the stored magnetic field energy ends up as heat. Let’s look at rough numbers. Suppose that the magnet maintains a field of 7 T in a channel of length 3 m and cross-section of area (56 mm)$^2$.

a. Find the total magnetic energy in joules.

b. The magnet is normally kept superconducting by a reservoir of liquid helium. The heat of vaporization of liquid helium is 83 J/mole (you can neglect the additional energy needed to bring He gas up to room temperature). If all the energy in (a) goes to vaporizing helium (and there’s an unlimited supply in the reservoir), how many moles of He gas do we get?

c. Suppose that all that helium gas exits the system via pressure-release valves, then comes up to room temperature. A mole of any ideal gas occupies about 24 liters at room temperature. What volume of helium gas would then flood the underground tunnel near the magnet?

35.2  **Impulse from changing field**

Two opposite walls of a rigid, nonconducting, rectangular box are uniformly charged with surface charge densities $\sigma$ and $-\sigma$ respectively. The positively charged wall occupies the region $0 < x < X$, $0 < y < Y$ of the plane $z = Z$. The other wall occupies the corresponding region of the plane $z = 0$. Inside the box there is a uniform magnetic field $\vec{B} = B_0\hat{y}$. Assume that $Z$ is much smaller than either $X$ or $Y$.

---

*Figure 35.3: See Problem 35.1. [Image from press.web.cern.ch/press/PressReleases/Releases2008/PR14.08E.html]*
a. Use the Lorentz force law to find the impulse experienced by the box (that is, momentum delivered to it) if the magnetic field is suddenly switched off.

b. Find the initial momentum of the electromagnetic field in the box. Make an Insightful Comment.

35.3 Fine point — energy and momentum of fields
The main text stated that the expression

\[ 2\varepsilon^\sigma_\lambda \partial_\mu F^\lambda_\mu, \]

could be replaced by

\[ -2\varepsilon^\sigma_\lambda \partial_\mu F^\lambda_\sigma\mu, \]

where \( F \) is the Faraday tensor. Why is this substitution justified?

35.4 Momentum flux
Now that we have complete formulas for energy flux and momentum flux, find the ratio of their \( z \) components for a plane wave traveling along the \( z \) direction, and interpret it in the light of Section 20.3 (page 302).

35.5 Angular momentum flux
[[Not ready]]
Chapter 36

Interlude: Faraday’s Field Lines

The lines of force, as he called the forces independently considered, stood before the eye of [Faraday’s] intellect as states of space, as tensions, vortices, currents, whatever they might be—this he himself was unable to determine—but there they were, acting on each other, pushing and pulling bodies about, spreading themselves around and carrying the disturbance from point to point.

— Heinrich Hertz, 1889

36.1 Framing: Our Toolkit

Michael Faraday drew a lot of diagrams like the ones in Figure 36.1, and similar ones involving magnets. He found that he could get a consistent picture of both electric and magnetic forces by imagining invisible “lines of force” sprouting out of positive charges and converging onto negative charges (and similarly for the poles of magnets). The magnitude of the forces increased as the lines were crowded laterally. The lines of force were under tension, like stretched rubber bands, yet repelled nearby lines. This “transverse pressure” made the lines want to avoid each other, so they spread out as they left a point charge; the connection to the crowding of field lines also gave rise to the $1/r^2$ force law.

Figure 36.1: Field lines. (a) Electric field lines (streamlines of $\vec{E}$) set up by two opposite point charges. The magnetic field lines set up by two opposite pole tips look the same. The figure is antisymmetric upon reflection through the central plane (dashed line). (b) Two identical point charges or magnetic pole tips. This time the figure is symmetric upon reflection.
It sounds crazy! Even decades later, the Continental *philosophes* were particularly severe on Faraday and his successors. And yet Faraday, with practically no formal education and certainly no vector calculus, used his intuitive picture as a guide to make a discovery that had eluded everyone else: the law of magnetic induction. Maybe his viewpoint belongs in our *toolkit* alongside the others.1 This short chapter will digress from our main study of relativity to outline how Faraday’s imagery expresses features of the energy–momentum flux tensor.

*Electromagnetic phenomenon:* Like charges, and like magnetic pole tips, repel; opposite charges and pole tips attract.

*Physical idea:* These effects reflect the momentum flux 3-tensor of the electromagnetic field next to each object.

### 36.2 FIELD LINE CROWDING IS RELATED TO FIELD STRENGTH

#### 36.2.1 Divergence-free fields

So far, we have expressed electromagnetic phenomena using *vector fields*, not “lines of force.” But Section 0.3.1 (page 7) made a connection: The *streamlines* (integral curves) of a vector field fill space, and they do resemble the curves Faraday drew for the two situations in Figure 36.1. (Today they are often called *field lines.*) As in the figure, they spray out of a positive point charge, then spread apart, indeed as if by mutual repulsion. However, if a negative charge is present, some or all of the field lines crowd back together, terminating on the charge.

Before we study field lines, let’s warm up with a more intuitive system: an incompressible fluid flowing steadily through a pipe. The local velocity near each point inside the pipe amounts to a vector field, whose streamlines are the paths taken by fluid molecules (averaged over thermal motion). Suppose that the flow encounters a constriction in the pipe. Then individual streamlines of the velocity field must converge. We know from watering our gardens that the fluid must also speed up as it passes through the constriction: Slow flow through the hose becomes fast flow through the narrow nozzle. If we observe at a fixed position, we may see a time-independent (steady) fluid velocity there; nevertheless, a speck of dust being swept along will be moving faster at the constriction than elsewhere.

More precisely, $\rho_m$, the density of the incompressible fluid, is constant. If $\vec{V}(\vec{r})$ is the velocity field, then the flux of mass is $\rho_m \vec{V}$, and the continuity equation for mass says

$$\vec{V} \cdot (\rho_m \vec{V}) + \frac{\partial}{\partial t} \rho_m = 0.$$  

Because $\rho_m$ is constant, we conclude that $\vec{V} \cdot \vec{V} = 0$: Incompressible flow has divergence-free velocity. We now recast this familiar result as a property of the streamlines of $\vec{V}$.

Write $\vec{V}$ in terms of its magnitude and direction: $\vec{V} = f(\vec{r}) \hat{n}(\vec{r})$. The divergence-free

---

1The graphical intuition will come back when we study radiation in Chapter 42.
property implies \( \dot{n} \cdot \vec{V} f = -f \vec{V} \cdot \dot{n} \), or

\[
\left( \frac{\dot{n} \cdot \vec{V} f}{f} \right) = -\vec{V} \cdot \dot{n}.
\]  

(36.1)

The left side of this equation is the relative rate of change of the magnitude of velocity as we move along a streamline. The intuition cited above leads us to expect that this should reflect changes in the transverse areal density of a set of neighboring streamlines, that is, the local number per area that pierce a plane perpendicular to the lines. Let’s see if the right side of Equation 36.1 has any such interpretation.

Instead of a constriction, Figure 36.2 shows, a pipe that expands slightly, just in one direction \((y)\). For simple (nonturbulent) flow, \(n\) lies in the \(xy\) plane. At the centerline, \(\partial \hat{n}_x / \partial x = 0\) but \(\partial \hat{n}_y / \partial y > 0\) in the expansion region. Thus, \(\vec{V} \cdot \dot{n} > 0\) for this diverging flow.

We now ask what is happening to the transverse areal density of streamlines. If \(N\) lines enter at the left, spread over cross-sectional area \(wa\), then they must exit into the larger area \((w + 2\Delta x \tan \theta)a\), where \(\tan \theta \approx \hat{y} \approx n_y\) evaluated at the top of the pipe. Because \(\hat{n}_y = 0\) at the center of the box, by a Taylor expansion \(\hat{n}_y \approx (\tan \theta) \approx w^2 (\dot{\hat{y}} \cdot \dot{n}) / \partial y^2\) evaluated at the center.

The area of the rectangle containing the streamlines increases from \(wa\) to \(u(1 + \Delta x \hat{x} \cdot \hat{n})a\). The transverse areal density of streamlines therefore decreases from \(N/\left(\partial x/a\right)\) to \(N(1 + \Delta x \hat{x} \cdot \hat{n})^{-1} / (\partial x/a)\). Its relative rate of change is then \(-\Delta x (\hat{V} \cdot \hat{n}) / \Delta x\), which indeed agrees with the right-hand side of Equation 36.1.

The relative rate of change is the logarithmic derivative. If two functions have the same logarithmic derivative everywhere, then one of them is a constant times the other. We have therefore established that the magnitude of velocity in an incompressible fluid is a constant times the transverse areal density of streamlines. The constant is arbitrary because we could have started with any number of streamlines, but the relative change has physical meaning.

The same result holds for any divergence-free vector field, such as a magnetic field, or electric fields in vacuum. Thus, our result establishes one part of Michael Faraday’s picture.
36.2.2 More general fields

Next, suppose that the pipe in Figure 36.2 has a small leak somewhere, which lets fluid escape to some region we are not considering. Then even though the fluid is incompressible, some mass will escape through the leak, so that the right side of the continuity relation will appear to be nonzero at the leak point, and therefore the velocity vector field has nonzero divergence there. We call the leak a point sink. Conversely, if there is fluid outside the pipe at high pressure, then a small leak can let fluid in, creating a point source where $\vec{V} \cdot \vec{V} \neq 0$. Flowlines can escape through a point sink, or appear from a point source, so to the world inside the pipe, they seem to terminate on such points.

The same remarks apply to the electric field, even though we have no analog of the "fluid": Because the Gauss law says the $\vec{V} \cdot \vec{E} \neq 0$ at charges, electric field lines terminate there. As we approach a point charge through surrounding vacuum, however, the remarks in the preceding subsection remain valid, and so the converging field lines imply diverging field strength.

36.3 ELECTRIC AND MAGNETIC FORCES VIA DERIVATIVES OF FIELD ENERGY

We have seen how the streamlines of $\vec{E}$ contain all the information needed to reconstruct the direction and magnitude of the electric field up to an overall constant, and similarly for $\vec{B}$. In Figure 36.1a, a computer has drawn some $\vec{E}$ field lines for two opposite point charges: We see maximum transverse areal density right at the charges, high density between them, and zero density out at infinity, as expected. Moreover, bringing the charges closer reduces the volume over which the lines are closely packed, and increases the volume in which the lines are sparse. That change in turn reduces the integral of $||\vec{E}||^2$, that is, the stored electrostatic field energy, and so opposite charges attract, as if the lines were real rubber bands under tension.

In contrast, pushing two like charges together (Figure 36.1b) increases the crowding at the central plane and hence the stored energy, so the charges repel. The next section will recover these familiar behaviors by interpreting the electromagnetic part of the momentum flux tensor in terms of Faraday’s imagery.

36.4 FORCES VIA THE MOMENTUM FLUX 3-TENSOR

Your study of mechanics has probably made it clear that often there is both an “energy” approach to a problem and also a different-seeming “force” approach. In any given problem one of those may be easier, so it’s good to understand both. To that end, let’s now look directly at the electric and magnetic forces predicted by our formula for the momentum flux (or “stress”) 3-tensor, $\tilde{T}_{ij}$. Equation 35.11 (page 560) gave it as

$$\tilde{T}_{ij} = -\frac{1}{\mu_0} (c^{-2} \vec{E}_i \vec{E}_j + \vec{W}_{ij} + \frac{1}{2} \delta_{ij} (-c^{-2} ||\vec{E}||^2 - \frac{1}{2} \vec{W}_{kk})).$$
where
\[
\dot{W}_{ij} = F_{i}^{\ k}F_{kj} = \varepsilon_{ikm}\varepsilon_{kjn}\vec{B}_{m}\vec{B}_{n}.
\]

**Your Turn 36A**

Finish showing that
\[
\vec{T} = \varepsilon_{0}(-\vec{E} \otimes \vec{E} + \frac{1}{2}\vec{\nabla}||\vec{E}||^{2}) + \frac{1}{\mu_{0}}(-\vec{B} \otimes \vec{B} + \frac{1}{2}\vec{\nabla}||\vec{B}||^{2}).
\]

(36.2)

#### 36.4.1 Electrostatic forces can be pictured as tension along, or pressure between, field lines

Along the midplane in Figure 36.1a, the electric field points along \(-\hat{x}\). Everything to the left of the midplane transfers momentum to everything to the right with flux of \(\vec{p}_{x}\) equal to
\[
\vec{T}_{xx} = \varepsilon_{0}(-\vec{E}_{x}^{2} + \frac{1}{2}||\vec{E}||^{2}) = -\frac{1}{2}\varepsilon_{0}(\vec{E}_{x})^{2}.
\]

(36.3)

The force that one object \(A\) exerts on another \(B\) is the rate of momentum transfer from \(A\) to \(B\). For Figure 36.1a, the momentum flux density Equation 36.3 is strictly negative throughout the midplane, so when integrated over that plane (or any other separating surface), it predicts a force on the right charge that is directed to the left, that is, attraction, consistent with Faraday’s “tension.”

Along the midplane in Figure 36.1b, the electric field is always perpendicular to \(\hat{x}\). Thus,
\[
\vec{T}_{xx} = \varepsilon_{0}(-\vec{E}_{x}^{2} + \frac{1}{2}||\vec{E}||^{2}),
\]

which is strictly positive, consistent with Faraday’s “transverse pressure.”

#### 36.4.2 Magnetostatic forces have similar pictorial expressions

Now imagine two long, thin, parallel bar magnets. As we bring one pole of the first magnet close to a pole of the second one, then close to those pole tips we find a similar set of field lines to those in Figure 36.1. Because the magnetic terms of Equation 36.2 have the same forms as the electric terms, we get the same results (again consistent with Faraday’s picture), plus small corrections from the other two magnet poles, whose fields are weak at the midplane.

#### 36.5 MAGNETIC INDUCTION

Faraday took his field lines seriously, as objects with some sort of reality. That helped him to suggest that whenever a wire “cut across” magnetic field lines, something physical would happen—its charge carriers would feel a force. Such “cutting across” could happen when a wire was dragged through a static \(\vec{B}\) field (as in a dynamo), or when a motionless wire was subjected to a growing or shrinking \(\vec{B}\) (field lines crowding in to increase their...
transverse areal density, as in a transformer). Eventually, Maxwell and others interpreted those statements as the magnetic part of the Lorentz force law and the field equation today called Faraday’s law, respectively.

FURTHER READING

Intermediate:
Limitations on the pictorial language of field lines: Wolf et al., 1996.

PROBLEMS

36.1 Magnetic stress
Consider the attraction between two rectangular bar magnets placed end-to-end with the first one’s N pole, a flat surface, separated from the other’s S pole, a parallel flat surface, by a narrow gap. You can ignore fringe fields, and indeed all fields outside the narrow gap. Just assume that $\vec{B}$ is uniform in the gap and points in the $\hat{x}$ direction.

a. Show that, for a static, purely magnetic field, $T^{\mu\nu}$ takes the form $\mu C^{\mu\nu}$, where $\mu$ is the energy density of the field and $C^{\mu\nu}$ is a constant tensor that you are to find.

b. Use the continuity equation for momentum to show that the force on each magnet (total rate of transport of momentum) equals $\mu \Sigma$, where $\Sigma$ is the area of the pole faces. (If this is not always true, use the idea behind the equation to describe when it will be true.) Then use Equation 36.2 (page 572) for $T^{\mu\nu}_{\text{field}}$ to evaluate this force.

c. The total energy in the field is dominated by the contribution from the high-field space between the magnet poles, so it’s $\mu a \Sigma$, where $a$ is the distance between poles. Give a second derivation, based on energy conservation, for the force of attraction between the magnets.
CHAPTER 37

Plane Waves in 4D Language

37.1 FRAMING: INDEPENDENT CHANNELS

Our goals are to organize, systematize, consolidate, integrate—and extend our knowledge. Chapter 20 gave some evidence for energy and momentum fluxes in plane waves. We’re now in a position to firm up those results and explore some applications.

Electromagnetic phenomenon: The two polarizations of light do not give rise to visible interference fringes when they are combined; they seem to act as independent channels.

Physical idea: The energy flux of such a superposition gets independent contributions from each polarization.

37.2 LORENZ GAUGE CHOICE

37.2.1 It’s useful

Section 34.8.1 introduced the 4-vector potential via

\[ E^{\mu \nu} = \partial_{\mu} A^{\nu} - \partial^{\nu} A^{\mu}, \]  \[ \text{[34.14, page 535]} \]

which cast the Maxwell equations as

\[ -\partial_{\mu} \partial^{\mu} A^{\nu} + \partial_{\mu} \partial^{\nu} A^{\mu} = \mu_0 j^{\nu}, \]  \[ \text{[34.16, page 536]} \]

with gauge invariance under

\[ A^{\mu} \rightarrow \tilde{A}^{\mu} = A^{\mu} + \tilde{\partial}^{\mu} \Xi. \]  \[ \text{[34.15, page 536]} \]

We could use this freedom to insist on Coulomb gauge as in Section 18.8.2 (page 281). But for radiation problems, it’s nicer to impose a Lorentz-invariant condition,

\[ \tilde{\partial}_{\mu} A^{\mu} = 0. \]  \[ \text{Lorenz gauge} \]  \[ (37.1) \]

---

1Named in honor of L. Lorenz. It’s a Lorentz-invariant condition, but not named for H. Lorentz.
Your Turn 37A

a. Express Equation 37.1 in terms of the 3D variables $\vec{A}$ and $\psi$.
b. Show that in Lorenz gauge, Equation 34.16 becomes four decoupled copies of the D'Alembert equation: $\square \vec{A} = -\mu_0 \vec{j}$, or

$$c^{-2} \frac{\partial^2}{\partial t^2} \psi - \nabla^2 \psi = \frac{\rho_q}{\varepsilon_0} \quad \text{and} \quad c^{-2} \frac{\partial^2}{\partial t^2} \vec{A} - \nabla^2 \vec{A} = \mu_0 \vec{j}.$$  \hspace{1cm} \text{(37.2)}

Unlike our discussion in restricted Coulomb gauge, Equations 37.2 are valid regardless of whether the charge density is zero. They are decoupled, but remember that the Lorenz gauge condition is a constraint linking the four variables $\psi$ and $\vec{A}$.

37.2.2 It’s permitted

Can we really impose Lorenz gauge? Suppose that we had a vector potential not obeying Equation 37.1; that is, $\partial_\mu A^\mu = f$ is some arbitrary function. Now apply a gauge transformation $A^\mu \to A'^\mu + \partial^\mu \Xi$. Then $\partial_\mu A'^\mu = f + \square \Xi$. But we have already found the solution to $\square \Xi = -f$ via its Green function in Chapter 25. So an appropriate $\Xi$ exists to bring any 4-vector potential into Lorenz gauge. The whole argument is a 4D upgrade of one we made in magnetostatics (Section 15.4).

37.3 Plane Waves and the Polarization 4-Vector

Section 32.6.6 (page 502) pointed out that plane-wave solutions to the scalar wave equation can be written in the compact form

$$\Phi(\vec{x}) = \frac{1}{2} C \exp(i \vec{k} \cdot \vec{x}) + \text{c.c.}, \quad \text{(37.3)}$$

characterized by a complex amplitude $C$ and a real 4-vector $\vec{k} = \left[ \begin{array}{c} \omega/c \\ \vec{k} \end{array} \right]$ (the 4-wavevector). These functions solve the scalar wave equation if $||\vec{k}||^2 = 0$ (“$\vec{k}$ is a lightlike 4-vector”). Recall that this is just the condition that the wave moves at speed $c$.

Similarly to the scalar wave equation, the Maxwell equations in Lorenz gauge (Equation 37.2) have plane wave solutions characterized by a lightlike wavevector $\vec{k}$. Unlike the scalar field case, each wave also has a polarization 4-vector $\vec{\xi}$:

$$A^\mu(\vec{x}) = \frac{1}{2} \xi^\mu \exp(i \vec{k} \cdot \vec{x}) + \text{c.c.} \quad \text{(37.4)}$$

This 4-vector field will be in Lorenz gauge if $k_\mu \xi^\mu = 0$. For example, this condition implies $\xi^0 = \xi^3$ if the wave propagates along $+\vec{k}$.

We can apply a gauge transformation with $\Xi$ that is itself of plane wave form (Equation 37.3):
Your Turn 37B

a. Show that when we do this, then $\vec{A}$ will still have the form Equation 37.4, but with $\xi$ replaced by its old value plus $i\vec{k}$.
b. Show that $\vec{A}$ is still in Lorenz gauge, because $k_{\mu}k^{\mu} = 0$.

We can use the freedom you just found to impose the additional condition that $\xi^0 = 0$ (and hence the scalar potential is zero). With that choice, and again assuming the wave propagates along $\hat{z}$,

$$[\xi] = \begin{bmatrix} 0 \\ p \\ q \\ 0 \end{bmatrix}.$$  \hfill (37.5)

Your Turn 37C

a. Work out the Faraday tensor for the plane wave with polarization Equation 37.5. Show that the electric field is parallel to $\xi$, and hence it is perpendicular to $\vec{k}$.
b. Show that the magnetic field is perpendicular both to $\vec{k}$ and to $\xi$.
c. Also confirm that your formula for $\mathbf{E}$ has the expected units.

In short, we have recovered the key results that

- There are only two linearly-independent polarizations of light traveling in a given direction, and
- Both are transverse to the direction of propagation.\(^4\)

This section has shown that gauge freedom allows a nonredundant representation of plane waves, in which the two physical polarizations appear via the values of two parameters ($P$ and $Q$).

37.4 ENERGY AND MOMENTUM CARRIED BY A PLANE WAVE
CONFIRM EARLIER EXPECTATIONS

Chapter 35 found the electromagnetic part of the energy–momentum tensor:

$$T_{\mu\nu}^{\text{field}} = -\mu_0^{-1} \left( F_{\mu\lambda} F^\lambda^{\nu} + \frac{1}{4} g_{\mu\nu} (F^{\alpha\beta} F_{\alpha\beta}) \right).$$ \hfill [35.11, page 560]

Ex. Show that for the Lorenz-gauge plane wave, the time-averaged energy–momentum flux tensor is

$$\langle T_{\mu\nu}^{\text{field}} \rangle = \frac{1}{2\mu_0} k_{\mu} k^{\nu} \xi^* \cdot \xi.$$ \hfill (37.6)

\(^4\)You’ll see this from a different viewpoint in Problem 37.1.
then specialize the result to the case in Equation 37.5.

**Solution:** First work out that

\[
\langle F^\mu_\lambda F^\nu_\sigma \rangle = \frac{1}{4} \left( \left( i(k \mu \xi^\lambda - k^\lambda \xi^\mu) e^{i k \cdot X} - i(k \nu \xi^\sigma - k^\sigma \xi^\nu) e^{-i k \cdot X} \right) \times \left( i(k^\sigma \xi^\nu - k_\nu \xi^\sigma) e^{i k \cdot X} - i(k^\nu \xi^\lambda - k_\lambda \xi^\nu) e^{-i k \cdot X} \right) \right).
\]

Next, use \( |k|^2 = 0 \) and \( k \cdot \xi = 0 \) from Section 37.3. Then note that \( \langle e^{\pm i k \cdot X} \rangle = 0 \), so we only need to keep the cross-terms:

\[
eq \frac{1}{4} \left( k^\mu (\xi^\nu - k^\nu \xi^\mu) + k^\nu (\xi^\mu - k^\mu \xi^\nu) \right) + \text{c.c.} \quad (37.7)
\]

Computing the trace then gives \( \langle F^\mu_\lambda F^\nu_\sigma \rangle = 0 \), so the second term of \( T_{\text{field}} \) is zero and we find Equation 37.6. Finally, substitute Equation 37.5 to obtain the time average of the energy-momentum flux tensor as

\[
\frac{1}{2\mu_0} k^\mu k^\nu \left( |P|^2 + |Q|^2 \right). \quad (37.8)
\]

The trick leading to Equation 37.7 underscores the convenience of using complex exponentials, so make sure you understand it. It is equivalent to noting that \( \sin \) and \( \cos \) each average to zero, whereas the constant 1 averages to 1.

The compact formula Equation 37.6 contains the energy and momentum densities, and the energy and momentum fluxes, of plane electromagnetic radiation. Our previous derivations of those quantities were less compelling, and anyway did not give us the overall constant of proportionality.

You should confirm that Equation 37.6 has units appropriate for energy density. Note that the two polarizations contribute *independently* to the energy and momentum (no cross-terms in Equation 37.8). This implies that they cannot interfere with each other; each polarization can only display interference phenomena with itself.

Figure 37.1 shows one consequence of the last result: Rotating the polarization of light in one arm of an interferometer, with no other change, abolishes its interference pattern.

**Your Turn 37D**

How would Equation 37.6 change if we had instead used a circular polarization basis?

Equation 37.6 is applicable for a single plane wave. It implies that the energy density...
Figure 37.1: [Photographs.] A Michelson interferometer setup. (a) The incoming light is linearly polarized. (b) Interference fringes on the viewing screen produced by the setup in (a). (c) A quarter-wave plate has been added to the upper arm of the interferometer. Light passing through this arm gets its polarization rotated by up to 90°, depending on the orientation of the plate (see Problem 52.1, page 734). (d) Rotating the polarization in one beam by the full 90°, with no other change, seems to abolish the interference. More precisely, the illumination intensity is now uniform, and intensity is what our eyes respond to. See Problem 37.3. (Not shown: Adding a circular polarizer at the output reveals that the pattern in (d) is still periodically modulated, but in polarization, not overall intensity.) See also Media 13. [Realization and photos courtesy William Berner.]

and the momentum flux 3-tensor are related in a simple way:

\[
\langle \vec{T}_{\text{field}} \rangle = \langle \rho_{\text{E,field}} \rangle \hat{k} \circ \hat{k},
\]

regardless of the polarization of the wave. Now consider an isotropic mixture of many different plane waves. More precisely,\(^{6}\)

- The directions \( \hat{k} \) are uniformly distributed.

\(^{6}\)See Chapter 24.
In every direction, the polarizations are uniformly distributed in the plane perpendicular to \( \hat{k} \), and the distribution of frequencies and amplitudes is independent of direction and polarization.

Let \( \langle \cdots \rangle \) denote averaging both over time and over the ensemble \( \{ \mathbf{k}_0, \mathbf{\hat{s}}_0 \} \) of waves. Averaging the preceding equation gives

\[
\langle \hat{T}_{i,i\text{ld}} \rangle = \langle \hat{k} \hat{\rho}_{\text{E,ld}} \hat{k} \rangle.
\]

The right side equals the angular average of \( \hat{k} \hat{k} \hat{\rho}_{\text{E,ld}} \).

A symmetric rank-two 3-tensor like \( \hat{k} \hat{k} \hat{\rho}_{\text{E,ld}}\hat{k} \) need not average to zero. For example, the identity tensor \( \mathbb{1} \) is unchanged by rotations (it’s a “tensor from Heaven”). Indeed, Problem 14.2 (page 220) implies that \( \langle \hat{k} \hat{k} \rangle \) is a constant times \( \mathbb{1} \). Also, rotation does not affect the trace of a 3-tensor, so the rotational average of \( \hat{k} \hat{k} \hat{\rho}_{\text{E,ld}} \) must be \( \frac{1}{3} \mathbb{1} \).

The off-diagonal components of the momentum flux 3-tensor must average to zero. However, the diagonal elements give the pressure, so averaging Equation 37.9 implies the simple conclusion that

\[
p = \frac{1}{3} \rho\varepsilon, \quad \text{equation of state for isotropic EM radiation} \tag{37.10}
\]

As mentioned in Section 20.2.3 (page 301), radiation pressure dominates over the gas pressure of ordinary matter in the early Universe, so Equation 37.10 is crucial for cosmology.

---

10 See Section 13.3.1 (page 196).
37.2’ a Gravitational waves
When we expand Einstein’s gravitational field equation for small fluctuations about flat space, the result is a second-order PDE. It has a complicated tensor structure, until we impose a gauge condition analogous to Lorenz gauge. Then it becomes ten decoupled copies of the usual scalar wave equation, with the same old plane wave solutions! In particular, the polarization tensor is a symmetric, rank-2 tensor. Remarkably, combining a suitable Lorentz gauge and removing residual gauge artifacts, similarly to the analysis of this chapter, again reduces the true number of independent polarizations to just two.

37.2’ b Spin versus polarization
You may ask, “If the quantum analog of light is a spin-one particle, then how can there be only two polarizations? After all, other systems with the same angular momentum (for example, the p-orbitals of a hydrogen atom neglecting spin, or an s-wave, triplet bound state of two spin-$\frac{1}{2}$ particles) have three states!”

It is true that you can always take a hydrogen atom (or a positronium “atom”) in its triplet state, and view it in its rest frame. Then its quantum states furnish a representation of the part of the Lorentz group that leaves its 4-momentum unchanged, that is, the group SO(3), and the usual analysis indeed guarantees three states. But a photon has no rest frame. The subgroup that leaves its 4-momentum unchanged is not SO(3). It does include 2D rotations about $\hat{k}$, but its other two dimensions are not bounded as in SO(3), and so its irreducible linear representations won’t follow the usual rules. There is thus no guarantee that any third polarization must be present, and it’s not.

In contrast, a fundamental particle with spin 1 and nonzero mass, for example a W or Z boson, will indeed have a third polarization state.

Similarly, Section 37.2’a said that the spin-2 graviton has only two polarizations, not the five we might have expected based on nonrelativistic quantum mechanics. This is again possible only because the graviton is massless.

Even weirder things can happen in a theory that contains massless particles but that is not invariant under spatial inversions. When neutrinos were thought to be massless, mathematically consistent theories were written in which only left-handed neutrinos (and only right-handed antineutrinos) existed! There is no Lorentz boost that changes the helicity of a particle moving at speed c, because no boost can overtake such a trajectory. So invariance under the connected part of the Lorentz group does not require that the existence of one helicity entails the existence of the other.
37.1 Plane waves in Lorenz gauge

Consider the plane wave with vector potential \( \vec{A}(t, \vec{r}) = \frac{1}{2} \vec{\zeta} e^{i(kz - \omega t)} + \text{c.c.} \)

a. Find the function \( A^0(X) \) that completes \( \vec{A} \) to a 4-potential in Lorenz gauge. Show that the electric and magnetic fields are proportional to 

\[ \vec{\zeta} - 2(\hat{z} \cdot \vec{\zeta}) \hat{z} \quad \text{and} \quad \vec{\zeta} \times \hat{z}, \]

respectively. The first of these expressions is the projection of \( \vec{\zeta} \) onto the plane perpendicular to \( \hat{k} \).

b. What is the significance of these results for the paradox that the starting formula for \( \vec{A} \) appeared to predict three independent polarizations of light? Compare to the approach in Your Turn 37C.

c. How might we have resolved that paradox without even bothering to compute \( \vec{E} \) and \( \vec{B} \), instead invoking gauge invariance?

37.2 Waves in 4D notation and \( T^{\mu\nu} \)

a. Write down an expression for the 4-vector potential corresponding to a plane wave propagating along \( +z \), in Lorenz gauge with angular frequency \( \omega \).

b. Your answer involves a polarization 4-vector \( \zeta^\mu \). Write down an expression for the most general such \( \zeta^\mu \). Your answer will involve three independent, arbitrary, complex constants.

c. You have found three linearly independent solutions to the wave equation. But we know light has only two independent polarizations! Resolve this discrepancy by calculating the Faraday tensor \( F^{\mu\nu} \) for this wave and making an Insightful Comment.

d. Use your answer to (c) to work out the time-averaged energy–momentum flux tensor for your wave. Your answer will be expressed in terms of \( \omega \), the constants you introduced in (b), and possibly some constants of Nature. Express in words the meaning of each nonzero component of your formula for \( T^{\mu\nu}_{\text{field}} \) in this situation. Make another Insightful Comment about the roles of the two polarizations in your answer.

(Hints: Use Equation 35.11 (page 560). Remember, you’re working in Lorenz gauge; that simplifies the math. Stick to 4-dimensional notation; there’s no need to re-express things in terms of \( \vec{E} \) and \( \vec{B} \).)

37.3 Interference versus polarization

A monochromatic plane wave of light, with angular frequency \( \omega \), gets split into two beams of equal amplitude. The beams travel different distances in vacuum, then recombine traveling in the same direction and land on a screen. We observe the flux of energy at various places on the screen (Figure 37.1, page 578 and Media 13).

Let \( \Delta \) be the difference in path lengths traveled by the two beams.\[^{11}\]

The original

\[^{11}\text{In a real interferometer, light also passes through some glass elements. Their effect on phase difference is lumped into an effective path-length and included in } \Delta.\]
beam is linearly polarized in some direction $\vec{\zeta}$. Along the way, some optical element may rotate the polarization by an angle $\theta$ (without changing anything else), or possibly leave it unchanged.

a. Write an expression for the energy flux in the recombined beam.

b. Write a simpler, more explicit expression for the time-average of your result in (a), including its dependence on $\theta$ and $\Delta$. (You may neglect any overall constant.)

c. Explain the disappearance of interference fringes seen in the figure using the math you just worked out.

d. Suppose that $\theta = \pi/2$, as in Figure 37.1d. Unlike in that figure, however, the recombined beam is subjected to a circular-polarizing filter before it hits the screen. What will we see on the screen now?

37.4 CMBR polarization

The cosmic microwave background radiation fills all of space. Chapter 30 explained why even if the CMBR were perfectly isotropic (the same in every direction) when viewed in one inertial coordinate system, nevertheless in another such system it would appear anisotropic, slightly hotter in one direction than in the antipodal direction.\footnote{Figure 30.3 (page 461).}

We now ask a different, more detailed question. Suppose that in one inertial coordinate system (the “CMBR system”) the radiation is both isotropic and also unpolarized. Will it then appear partially polarized in another inertial coordinate system? To answer this physical question in the context of classical electrodynamics, take the following steps.

Suppose that we are moving at velocity $\beta c \hat{z}$ relative to the CMBR system. Clearly, if we look out in directions $\pm \hat{y}$ we won’t detect any apparent polarization, by azimuthal symmetry of the problem. So let’s consider looking out in one of the perpendicular directions, say $\hat{x}$. Now we wonder if there will be some apparent preference for the polarization along $\hat{z}$ relative to $\hat{x}$, or vice versa.

a. Write down the 4-vector potential corresponding to a plane wave of angular frequency $\omega$, moving along $+\hat{y}$. Express the answer by using a wave 4-vector $\vec{k}$ and a polarization 4-vector $\vec{\zeta}$. Use the usual complex exponential representation, and assume that $\vec{\zeta}$ is real (linear polarization). It will be convenient to work in Lorenz gauge, that is, to require $\vec{\phi}_\mu A^\mu = 0$. What conditions must $\vec{k}$ and $\vec{\zeta}$ obey in order to have a solution to the vacuum Maxwell equations?

b. Now apply a Lorentz boost to $\vec{k}$ and $\vec{\zeta}$, to a coordinate system moving relative to the original one at speed $\beta c$ in the $+\hat{z}$ direction. Confirm that, when viewed in the new coordinate system, the wave still obeys the conditions you found in (a). Find the frequency as observed in this new system. (What is the name for your result?) Find the direction of the wavevector in this new system. (What is the name for your result?)

c. Find the electric field of your wave solution in the original coordinate system. Show that it’s unchanged if you replace $\vec{\zeta}$ by $\vec{\zeta} + \xi \vec{k}$ for any constant $\xi$. Using this freedom, we can simplify the problem by also requiring that $\vec{\zeta}^0 = 0$. Write the most general polarization
4-vector $\zeta^\mu$ obeying all these requirements. Express it in terms of an amplitude $b$ and the angle $\psi$ that the electric field makes with the $x$ axis.

d. Take your boosted polarization vector from (b). Confirm that its electric field, viewed in the new coordinate system, is still transverse. Use the trick in (c) to find an equivalent polarization vector $\zeta$ with the convenient property $\zeta^{x'=0} = 0$. Find the new amplitude $\tilde{b}$ and the angle $\tilde{\psi}$ that the electric field makes with the $x'$-axis. That is, find $\tilde{b}$ and $\tilde{\psi}$ as functions of the original wave’s parameters ($\omega$, $b$, and $\psi$), and $\beta$.

e. Suppose that Earth is bathed in cosmic microwave background radiation that is isotropic and unpolarized in one inertial coordinate system. Section 24.3.2 (page 388) argued that we can regard the radiation coming from any direction in the sky as a superposition of randomly linearly polarized plane waves, whose polarization angles $\psi$ are uniformly distributed. Find the corresponding distribution of polarization angles $\tilde{\psi}$ and comment.
CHAPTER 38

A Simple Spherical Wave

38.1 FRAMING: DIPOLE DOUGHNUT

Plane waves are simple, but we are never literally going to encounter a wave with infinite, planar wavefronts. On the other hand, we do frequently encounter small sources of light (even a single fluorescent molecule) that we view from far away. Our intuition with mechanical waves leads us to expect some sort of expanding ripple—a spherical wave solution to the Maxwell equations. Unlike with water waves, however, we’ll see that the polar graph of energy flux can look like a doughnut.

Electromagnetic phenomenon: Each far-field wavefront of a dipole spherical wave is spherical, but the resulting energy flux is not isotropic.

Physical idea: The amplitude of the electric and magnetic far fields in any direction involve the projection of a constant vector into the plane transverse to that direction.

38.2 AN EXACT SOLUTION WITH SPHERICAL WAVEFRONTS

38.2.1 Analogy to acoustics

We know about spherical waves in acoustics, where the wave function is a scalar. Let’s therefore write the simplest possible generalization to a vector potential as a trial solution, and see whether it can be adjusted to work. Our trial solution is just a constant vector times the scalar spherical wave solution:

\[ \mathbf{A}(t, \mathbf{r}) = \frac{1}{2} \xi \frac{1}{kr} e^{-i\omega t \pm ikr} + \text{c.c.} \] (38.1)

Here \( k \) is a scalar, \( r \) is distance from the origin, \( \xi \) is a constant vector, and as usual \( \omega = ck \). The prefactor \( 1/k \) is a conventional choice designed to make \( (kr)^{-1} \) be dimensionless. The upper sign corresponds to outgoing spherical wavefronts; the lower sign to incoming.

To find the scalar potential, make a similar trial solution for it:

\[ \mathbf{A}^0(t, \mathbf{r}) = \frac{1}{2} \alpha(r) e^{-i\omega t \pm ikr} + \text{c.c.} \] (38.2)

Here \( \alpha(r) \) is an unknown function that we must find. In this chapter it will be convenient to work in Lorenz gauge.
38.2 An Exact Solution With Spherical Wavefronts

Your Turn 38A

a. Impose the Lorenz gauge condition to find what $\alpha$ must be. It need not be a constant, nor even a constant divided by $r$, but you can nevertheless find it.
b. Check that each of the three functions in Equation 38.1 obeys the wave equation.
c. Also check that your answer to (a) has this property. Hence, conclude that Equation 38.1, along with your version of Equation 38.2, solve the Maxwell equations.

Result (b) is not a surprise—sound waves from a point source also have this same form for the air pressure as a function of position and time. What may be surprising, however, is how the wave energy is distributed. The wavefronts of Equation 38.1 (loci where $\vec{A} = 0$) are nice concentric spheres. The amplitude $||\vec{E}||/(kr)$ is also independent of direction. One might guess, then, that the wave sends energy isotropically in every direction. The next section will check, and overturn, that guess: For example, the $\vec{B}$ fields involve the curl of $\vec{A}$, which is not isotropic.

38.2.2 Far fields carry energy in an anisotropic pattern

We could now compute exact expressions for the electric and magnetic fields of the spherical wave. But first, consider what we see when we move very far away from the origin along some direction $\hat{n}$. Out there (near the position $L\hat{n}$), the wavefronts aren’t curved very much, and the solution resembles a plane wave\(^1\) with wavevector $\vec{k}_{pw} = k\hat{n}$ and polarization vector $\vec{\xi}_{pw} = \vec{\xi}/(kL)$. We can therefore apply formulas from Chapter 37.

Your Turn 38B

a. Work out $\vec{\xi}_{pw}^0$.
b. Then find the electric and magnetic far fields in terms of $L$, $\hat{n}$, and $\vec{\xi}$ (perhaps after reviewing Your Turn 37C, page 576).
c. Consider the case where $\vec{\xi}$ is real. How do the amplitudes of the far fields depend on the angle between $\hat{n}$ and $\vec{\xi}$? [Hint: Consider the case with $\vec{\xi}$ pointing along $\hat{z}$.]

Perhaps surprisingly, the fields (and therefore the energy flux) are not at all isotropic: The amplitudes of the far $\vec{E}$ and $\vec{B}$ fields in various directions do depend on angle. They are maximal in the directions perpendicular to $\vec{\xi}$, and zero when we view the wave from far away along the directions $\pm \vec{\xi}$.

There are many other spherical wave solutions, but the simplest one, which we just found, will turn out to be the leading contribution when we work out the radiation from an oscillating electric dipole in Chapter 43.\(^2\) Accordingly, its pattern of energy flux is sometimes called the dipole doughnut pattern (Figure 38.1).

The far fields have another crucial property: Both $\vec{E}$ and $\vec{B}$ fall off with distance as $1/L$. So the energy density, and hence also the energy flux, fall off with distance as $1/L^2$.

\(^1\)There are corrections that are higher order in powers of $1/L$.
\(^2\)We’ll encounter the other spherical waves when we study radiation systematically in Chapters 43–44.
Chapter 38  A Simple Spherical Wave

Figure 38.1: [Mathematical function.] Polar plot of \( \sin^2 \delta \); that is, the surface for which radial distance at any \( \delta, \phi \) equals \( \sin^2 \delta \). This depiction resembles the surface of a toroidal pastry, at least when you are hungry. (Perhaps it looks more like a spherical balloon after we have poked the north and south poles inward to touch.) Your results in Your Turn 38B imply that in the far-field region, the Poynting vector of the spherical wave solution with \( \xi \) along \( \hat{z} \) is directed radially, with magnitude proportional to this function (see also Problem 38.1). The figure makes it visually clear that no energy is directed along \( \pm \hat{z} \), and maximum flux is in the equatorial belt.

We can therefore say, a bit more carefully than in Section 25.5.3 (page 397), that the total energy output passing through a sphere of radius \( L \) approaches a constant as \( L \to \infty \). Any system that creates an exact outgoing spherical wave of this type therefore constantly sends energy all the way out to infinity.

38.2.3 Near fields resemble a time dependent electric dipole

The opposite limit is interesting, too. Instead of expanding for large \( r \) at fixed \( \omega \), we can imagine sitting at a fixed distance from the origin and consider the limit \( \omega \to 0 \), that is, keeping only the leading behavior in powers of \( \omega \). In this “near field” regime, \( \vec{E} \) dominates \( \vec{B} \), and moreover \( \vec{E} \) has a very familiar form: It resembles the dipole field of electrostatics.\(^3\) Farther out, the exact solution smoothly interpolates between that form and the plane-wavy far fields.

38.3 A CIRCULARLY POLARIZED SPHERICAL WAVE?

It’s also instructive to work out the case of complex \( \vec{\xi} \), for example \( \hat{x} + i \hat{y} \). We can explore this case by using the same strategy as Section 38.2 and asking, “What kind of plane wave does the solution look like when we stand far from the origin along some direction \( \hat{n} \)?”

Your Turn 38C

a. Is there any direction in which the far field approaches a circularly polarized plane wave?
b. Is there any direction in which it’s linearly polarized?
c. Is there any direction in which its amplitude is zero?
d. Can you explain your answers intuitively?

\(^3\)You’ll work this out in Problem 38.3.
38.4 INTERFERENCE

[Just as with sound, we can imagine a set of point sources of spherical waves, all vibrating in sync. For example, an incoming plane wave could hit an ordered array of atoms, and set them all in synchronized motion; each will then re-radiate some spherical wave. The total fields that land on a distant projection screen can then form a **diffraction pattern.**

Unlike sound, however, the fact that light has two transverse polarizations complicates matters. There is no way that the crests of a wave traveling along \( \hat{z} \) and polarized along \( \hat{x} \) can cancel the troughs of another wave traveling in the same direction but polarized along \( \hat{y} \), nor will crests combine with crests in the familiar way. Instead, when light from multiple sources lands on a screen, the illumination on each point of the screen involves the vector sums of the \( \vec{E} \) and \( \vec{B} \) waveforms.]

38.5 SUMMARY

The plane wave solutions are exact and simple in either Coulomb gauge (Section 18.9) or Lorenz gauge (Chapter 37). They carry energy and momentum. For any \( \vec{k} \), there is a two-dimensional vector space of plane waves with various polarizations.

In contrast, the exact spherical wave solutions are simpler in Lorenz gauge than in Coulomb gauge. They carry energy and momentum from a point source out to infinity. For any \( k \), we have found a three-dimensional vector space of spherical waves parameterized by \( \xi \) (and later we will find many more). Their wavefronts are spheres (hence the name), but they beam out energy in a dipole pattern. For real \( \xi \), that pattern is maximal in the directions perpendicular to \( \xi \).

---

4Figure 37.1 (page 578) already showed this phenomenon. The striped interference pattern arose because the waves were not quite plane; it disappeared when one beam’s polarization was rotated perpendicular to the other’s.

5See Sections 18.9 (page 283) and 37.3 (page 575).
38.1 Energy flux
a. Continuing Your Turn 38B, derive a formula for the time-averaged energy flux in the far-field region and comment. What directions get the largest energy flux, and why?
b. Repeat for the situation studied in Your Turn 38C

38.2 Doughnut + doughnut = peanut
Create a polar plot analogous to Figure 38.1 (page 586), but for polarization \( \mathbf{r} = (\hat{x} + \hat{y})/\sqrt{2} \), and comment.

38.3 Exact \( \mathbf{E} \) and \( \mathbf{B} \) fields
Your Turns 38B and 38C used far-field approximation. In this problem, instead get exact answers from our exact solution.
a. Work out the curl of Equation 38.1 (page 584) with the choice of outgoing (upper) sign in the exponential. After you get the general answer, specialize to \( \mathbf{r} = \mathbf{r} \) and show that
\[
\mathbf{B} = \frac{\xi}{2k} \hat{\phi} \sin \theta (r^{-2} - i k r^{-1}) e^{-i o t + i k r} + \text{c.c.} \quad (38.3)
\]
where \( o = c k \). [Hint: First prove the geometrical identity \( \hat{r} \times \hat{z} = -\hat{\phi} \sin \theta \).]
b. Work out \( -\mathbf{V} \mathbf{A}^0 - \partial \mathbf{A}/\partial (ct) \), using your result from Your Turn 38A (page 585), and from there find the electric field. After you get the general answer, again specialize as in (a) to show that
\[
\mathbf{E}/c = \frac{i \xi}{2k^2} \left[ \hat{r} \cos \theta (2 r^{-3} - 2 i k r^{-2}) + \hat{\phi} \sin \theta (r^{-3} - i k r^{-2} - k^2 r^{-1}) \right] e^{-i o t + i k r} + \text{c.c.} \quad (38.4)
\]
[Hint: First prove that \( \hat{z} = \hat{r} \cos \theta - \hat{\phi} \sin \theta \).]
c. Now consider the limit \( k r \ll 1 \) (low frequency, or equivalently very short distance), and so retain only terms with the biggest negative power of \( r \) in your expressions for \( \mathbf{E} \) and \( \mathbf{B} \). Connect the result to an electrostatics problem we have discussed in a previous chapter: Where have we seen this expression before?
d. Next, consider the opposite regime, that is, retain only terms with the least-negative power of \( r \). Connect your result to your answers to Your Turn 38B (page 585).

38.4 Angular momentum of fields
Background: EM waves can also carry angular momentum. You found the density of ordinary momentum in Your Turn 35D. So the density of angular momentum \( \mathbf{L}_z \), computed using the origin as reference point, is \( (\mu_0 c^2)^{-1} [\mathbf{r} \times (\mathbf{E} \times \mathbf{B})]_z \). As usual, we will consider only the time average of \( \mathbf{L}_z \).
a. Confirm that the formula given has the appropriate units to be the volume density of angular momentum.
b. Consider the outgoing, exact spherical wave solution (Equation 38.1), with $\mathbf{E} = C(\mathbf{x} + i\mathbf{y})$. Here $C$ is an overall constant with appropriate units. Work out the electric and magnetic fields far from the origin, to leading order in an expansion in powers of $(kr)^{-1}$.

c. Your result in (b) may seem to imply that the density of angular momentum falls with distance as $r(1/r)(1/r) = r^{-1}$. If true, that would imply that the net transport through a sphere of radius $R$ grows without bound as $R \to \infty$. What saves us from that absurd conclusion?

d. Go back to (b) and keep also the next-order terms in the expansion. Then redo (c) retaining those terms and comment.

e. Because everything moves radially outward at speed $c$, the radial component of the flux of $\mathbf{L}_z$ is your answer to (d), multiplied by $c$ to convert units into a flux. Suppose that a sphere of large radius $R$ surrounds the origin and absorbs all the radiation. Before you compute anything: Do you expect physically that the whole sphere will gain any net angular momentum $\mathbf{L}_z$? Why/why not?

f. Now integrate the flux of $\mathbf{L}_z$ over the surface of the big sphere to get the rate at which angular momentum is transferred to the sphere, and check against your answer to (e).

g. Also find the power absorbed by the sphere.

h. Divide the answers to (f) and (g) and comment. [Hint: Review Section 20.3 (page 302).]

---

6Chapter 43 will show that this solution could represent the radiation given off by a rotating electric dipole at the origin, in the electric dipole approximation.
CHAPTER 39

Radiation Green Function Revisited

It is a good thing to have two ways of looking at a subject, and to admit that there are two ways of looking at it.

— James Clerk Maxwell

39.1 FRAMING: LOOKBACK

Chapter 25 found a solution to the D’Alembert equation (inhomogeneous wave equation), but by the unsatisfying method of “lucky guess.” Let’s use “Einstein thinking” to recover that result more directly. Then this chapter and others that follow will give some generalizations to the derivation of radiation in Chapter 25:

- We must remove the limitation to sources with zero net charge everywhere.
- We must also go beyond the case of harmonic time variation. Previously we assumed that only a single frequency was present (the current was assumed proportional to a sine wave in time). It is true that any periodic function can be expanded in Fourier series, and we can analyze each component frequency separately. Moreover, when we have a mole of electrons distributed through a wire and moving in phase, then it makes sense to ignore their particulate character. But when a single electron shakes back and forth, even with a single frequency, its charge density and current at a fixed location are delta functions in time; the Fourier series contains every multiple of the fundamental frequency, and so is not a useful tool. Also, when a single electron flies through space and then hits a wall, that one-time deceleration is not even periodic.
- Chapter 38 found a spherical wave solution, but we still need to show how that wave may be created (see Chapter 43).
- The antenna considered in Section 25.5 did generate a spherical wave, but not the same one that we found in Chapter 38! We need a more general understanding of spherical waves (see Chapter 44).

Electromagnetic phenomenon: In vacuum, a charged particle in uniform motion carries fields along with it but does not radiate.

Physical idea: Looking back from any observation event, there is always exactly one causally connected point on the charge trajectory, and its contribution to the electromagnetic field falls as $r^{-2}$. 
39.2 The Relativity of Time Ordering Constrains Causality

Is the upper-left corner of this page higher or lower than the upper-right corner? Obviously there’s no absolute answer to that question. The higher corner can be made lower by rotating the page. On the other hand, if you stub your toe in the night, and a dog barks on the next block, there doesn’t seem to be any doubt about which happened first.

Before 1905, physicists would have agreed, because galilean transformations have a nice property (Figure 39.1a): Any two G-inertial coordinate systems will agree that event \( R \) is simultaneous with \( P \), that \( S \) precedes \( P \), and that \( Q \) follows \( P \). Geometrically, this is a matter of whether you’re above or below the \( x \) axis, and all G-inertial coordinate systems have the same \( x \) axis.

But turning to Lorentz transformations, which we now believe are invariances of Nature, we found a surprise: An observer who uses an E-inertial coordinate system moving to the right (figure panel (b)) will disagree with the original observer, saying that \( R \) precedes \( P \) (it lies below the \( x' \) axis). ²

Similarly, a left-moving observer (panel (c)) would say that \( R \) (and even \( S \) in the case shown) happen later than \( P \). Interestingly, however, all E-inertial coordinate systems agree that \( T \) is later than \( P \), and \( U \) is earlier. That’s because these points lie beyond the lines at ±45 deg to the axis, and we can never bend the \( x' \) axis past those lines.

In algebraic terms, the temporal ordering of \( P \) and \( Q \) is unambiguous if and only if
\[
|t_Q - t_P| > |\mathbf{r}_Q - \mathbf{r}_P|/c.
\]
We can restate this by using the invariant interval:³ The temporal ordering of two events \( P \) and \( Q \) is unambiguous if \((c\Delta t)^2\) is nonnegative, that is, if \(|\Delta x_{pq}|^2 \leq 0\). Section 32.6.4 introduced the terms timelike separation if \(|\Delta x_{pq}|^2 < 0\), lightlike if it’s exactly zero, or spacelike if it’s positive. Temporal ordering is ambiguous (dependent on which E-inertial coordinate system we choose) if the separation is space-

---

1Recall that a G-inertial coordinate system is one in which the equations of motion take their usual form, in an imagined world governed by newtonian mechanics (Section 26.6.1, page 409).
2In some even faster-moving coordinate systems, \( Q \) precedes \( P \)!
3See Equation 32.21 (page 500).
like.

The relativity of simultaneity just discovered may seem to be a disaster for physics. How can we claim that anything "caused" anything else, if we don’t know which happened first? But it’s not a complete disaster: When two events have timelike or lightlike separation (PT or PU), then we do know for sure which was first. So we can get out of difficulty if we insist that

*If two events are spacelike separated, then neither one may be said to have caused, or even influenced, the other.*

This makes sense when we notice that, in order for two such events to influence each other, one of them would have to send a signal to the other that moves faster than the speed of light in vacuum. Really what we’re asserting, then, is that no signal (causal agent) can move faster than light. This prohibition is consistent with the relativistic velocity addition formula, which always yields a new velocity \( \leq c \). Now we see that the speed limit is also necessary to avoid a physically nonsensical confusion about causality.

### 39.3 GREEN FUNCTION FOR THE D’ALEMBERT EQUATION

#### 39.3.1 Lorentz invariance tightly constrains the Green function

Section 37.2 obtained a version of the Maxwell equations valid in Lorenz gauge (Equation 37.2, page 575):

\[
\Box A^\mu = -\mu_0 J^\mu. \tag{39.1}
\]

This is four decoupled copies of a single equation, so to simplify the notation let’s first solve the scalar D’Alembert equation:

\[
\Box \phi = -J \tag{39.2}
\]

and later add the 4-vector index and factor of \( \mu_0 \). Chapter 25 found a solution to Equation 39.2, but we had to make an unobvious guess (Equation 25.4, page 392). Let’s start over, using “Einstein thinking.”

Equation 39.2 is linear and translation-invariant, so we expect that the solution can be written in terms of a Green function:

\[
\phi(X) = \int d^4 X_\ast \, D_t(X - X_\ast) J(X_\ast). \tag{39.3}
\]

We now use invariance to constrain the possible form of the unknown function \( D_t \), show that there is only one reasonable choice, then confirm that with that choice, the formula Equation 39.3 solves Equation 39.2 for any source function \( J \).

The constraints are that:

- \( D_t \) must be a 4-scalar function of \( \Delta X = X - X_\ast \).

---

*What about quantum entanglement? That’s not part of this book, but every discussion seems to end up, after a lot of analysis, concluding that there’s still no way to transmit useful information faster than \( c \).*

November 14, 2023; Contents  Index  Notation  Glossary
It must have dimensions \((\text{length})^{-2}\), by Equation 39.2. But it cannot involve any constants of Nature, because the equation doesn’t contain any.

- It should vanish when \(\Delta X^0 < 0\), because the behavior of charges in the future cannot affect the values of fields in the past.\(^5\)

The first constraint suggests that \(D_t\) must be a function of the invariant interval.

**Your Turn 39A**

One scalar quantity with the required units is \(\|\Delta X\|^2\). What’s wrong with that choice?

Luckily, there is another option: We can satisfy all the constraints with a function of this form:

\[
D_t(\Delta X) = \frac{1}{2\pi} \delta(\|\Delta X\|^2) \Theta(\Delta X^0).
\]

Taking the factors in turn,

- Soon we’ll see why the prefactor must be \(1/(2\pi)\).
- The delta function is motivated by the idea that\(^6\) electromagnetic influences travel at speed \(c\). Two points can be joined by a path traversed at speed \(c\) only if they are lightlike-separated, that is, \(\|\Delta X\| = 0\).
- The last factor is a “Heaviside step function,” and it enforces causality. Together with the delta function, it says that fields at \(X\) can only be influenced by sources lying in the past light cone\(^7\) of \(X\).

The delta function has dimensions\(^8\) inverse to \((\text{length})^2\). The step function is dimensionless. So our proposal has the required units.

The argument of the delta function is the invariant interval, so this whole factor is 4-scalar. The step function looks noninvariant at first, because Lorentz transformations can affect the temporal ordering of two events: \(\Delta X^0\) may not have the same sign as \(\Delta X^0\). However, this problem can only arise for spacelike-separated events, that is, a pair of events with invariant interval \((\Delta r)^2\) less than zero. The delta function tells us that such events cannot contribute anything to the proposed Green function. Only lightlike-separated events contribute, and Section 39.2 argued that the temporal ordering of any such pair of events is unambiguous.

In short, \(D_t\) is a 4-scalar function. The other ingredient in Equation 39.3 is \(d^4X\), which we saw in Section 34.9.1 is also Lorentz-invariant. Thus, Equation 39.3 is overall a Lorentz-invariant recipe to obtain \(\phi\) from \(\beta\), as desired.

\(^5\)Strictly speaking, the fields must be causal; the potentials could be nonzero outside the light cone, as indeed they are in Coulomb gauge (which spoils manifest Lorentz invariance). We are only setting out heuristic expectations that will help us to formulate a promising guess.

\(^6\)See Section 25.3 (page 392).

\(^7\)Section 32.6.4 (page 500).

\(^8\)See Section 0.3.8 (page 11).
Our trial solution has all the qualitative properties we expect it should have. But we still need to confirm that it really solves the D’Alembert equation! Once that’s done, *everything about radiation will follow from Equation 39.4.*

### 39.3.2 Reformulate and confirm the trial solution

Our proposed Green function is simple, and seems promising. After admiring it, we now rephrase it in a way that obscures its Lorentz invariance but will facilitate checking that it does solve the D’Alembert equation (Equation 39.2).

Begin by substituting the trial solution Equation 39.4 into the right side of Equation 39.3. The resulting expression has four integrals and one delta function. We will now use the delta function to eliminate one of the integrals, specifically the one over \( t = t_\pm \).

Recall from Section 34.9.1 how delta functions transform:

\[
\delta(f(t_\pm)) = \sum \frac{df}{dt} \bigg|_{t_\pm}^{-1} \delta(t_\pm - t_\pm),
\]

where \( t_\pm \) are all the values of \( t \) at which \( f(t_\pm) = 0 \). For our application,

\[
f(t_\pm) = -c^2(t - t_\pm)^2 + R^2 \quad \text{where } R = ||\vec{r} - \vec{r}_s||.
\]

The quantities \( t \) and \( R \) are constants for purposes of evaluating the integral over \( t_\pm \).

There are two solutions to \( f = 0 \): \( t_\pm = (ct - R)/c \) and \( t_\pm = (ct + R)/c \). Of these, however, the second is acausal, and so cannot contribute (the step function in Equation 39.4 eliminates it). Turning to the first,

\[
\frac{df}{dt} \bigg|_{t_\pm} = 2c^2(t - t_\pm) = 2c^2(t + R/c) = 2cR.
\]

Thus, the transformation rule gives

\[
\delta(||\Delta \vec{r}||^2) \Theta(\Delta \vec{r}^2) = \frac{1}{2cR} \delta(t_\pm - t + R/c).
\]

That result lets us easily do the \( t_\pm \) integral in Equation 39.3. The three remaining integrals become

\[
\frac{1}{4\pi} \int d^3 r' \frac{1}{R} \delta(t - R/c, \vec{r}_s).
\]

Chapter 25 already confirmed that \( \phi(t, \vec{r}) \) defined by this formula solves the scalar D’Alembert equation. This time, however, we rediscovered it without having to make a lucky guess, by using “Einstein thinking” (imposing manifest Lorentz invariance and causality) and dimensional analysis.

### 39.4 REMARKS

#### 39.4.1 Upgrade to 4-vector fields

Equation 25.4 is pretty simple: For each location \( \vec{r}_s \) inside the source, it tells us to look back in time to the moment when charges and currents at that location could have influenced
Figure 39.2: [Spacetime diagram.] **Causal look-back.** The cylinder represents a circular loop of current fixed in the $xy$ plane. The past light cone of an observer at time $t = 0$ and position $\vec{r} = 0$ intersects the region of nonzero charge flux in the curve shown. Only points on that intersection contribute to the potentials observed at $t, \vec{r}$.

the fields at $(t, \vec{r})$, then introduce a factor of $1/(4\pi R)$. To upgrade this result to electrodynamics, just use the scalar solution four times with $\vec{J} = \mu_0 \vec{J}^\mu$: Equations 39.3–39.4 become

$$A^\mu(X) = \frac{\mu_0}{2\pi} \int \delta(\|X - X_+\|^2) J^\mu(X_+)$$  \hspace{1cm} \text{Green function solution, Lorenz gauge} \hspace{1cm} (39.5)

Equation 25.4 gives the equivalent form (also in Lorenz gauge)

$$A^\mu(X) = \mu_0 \int \frac{1}{4\pi \|\vec{r} - \vec{r}_s\|} J^\mu(X^0 - \|\vec{r} - \vec{r}_s\|, \vec{r}_s).$$  \hspace{1cm} \text{(39.6)}

Unlike our result in restricted Coulomb gauge (Chapter 25), Equation 39.6 is valid even when the charge density is not everywhere zero. In general, it assigns a nonzero value to the scalar potential.

As a check, for a point charge sitting at rest at the origin $\vec{J} = 0$ and $\rho_0$ is time-independent. So Equation 39.6 reproduces the static Coulomb potential of a point charge. More generally, we have shown that Equation 39.6 with Equation 39.4 gives the fields created (caused) by a general distribution of charges and currents, so it’s called the **causal Green function solution**.\(^\text{10}\) For example, Figure 39.2 shows a current loop as a cylinder in spacetime, and what locus on that loop’s history contributes to fields observed at a particular point.

We now have two formulations of our solution. Each has its merits:

- Equation 39.5 is manifestly Lorentz invariant, and it’s the most useful version for problems with a single charge moving relativistically (synchrotron and Čerenkov radiation, Chapters 45 and 50).
- Equation 39.6 involves one fewer integral. It’s the most useful for nonrelativistic radiation, for example, in the multipole limit (Chapters 43 and 44).

---

\(^9\)L. Lorenz was the first to write the retarded potentials in Lorenz gauge, though not in the manifestly invariant form shown here.

\(^{10}\)Other names for it include **retarded Green function** or **retarded propagator**. The names refer to the fact that the formula “looks back in time.”
39.4.2 Check self-consistency

We’re not quite done. Equation 39.1 is not equivalent to Maxwell unless $\Delta$ is in Lorenz gauge. Does our solution really have that property?

To find out, we must compute

$$\frac{\partial}{\partial X^\mu} A^\mu = \int d^4X^\ast J^\mu(X^\ast) \frac{\partial}{\partial X^\mu} D_t(X - X^\ast).$$

First, note that by the Chain Rule

$$\frac{\partial}{\partial X^\mu} D_t(X - X^\ast) = -\frac{\partial}{\partial X^\mu} D_t(X - X^\ast).$$

After that substitution, we can integrate by parts to find

$$\frac{\partial}{\partial X^\mu} A^\mu = \int d^4X^\ast D_t(X - X^\ast) \frac{\partial}{\partial X^\mu} I^\mu(X^\ast).$$

The right side of this expression is zero, by the continuity equation that any 4-current distribution must obey.

You may be dissatisfied: “The Green function is the response to a blip, but an isolated blip cannot obey the continuity equation!” The logic is that:

- The Green function is indeed a solution to the D’Alembert equation, Equation 39.1, for an isolated blip source.
- If we assemble a lot of blips together into a $J$ field that obeys the continuity equation, then we just showed that the solution will also be in Lorenz gauge;
- and therefore, the combined solution will also solve the Maxwell equations, even though it is constructed out of blips that individually do not.

39.5 POINT PARTICLE EXECUTING SPECIFIED MOTION

39.5.1 The Liénard–Weichert potentials follow from the Green function solution

Let’s return to the wish-list at the start of this chapter. Now that we have found the potentials generated by an arbitrary distribution of charge and current, we can specialize to a point charge, for example, the problem mentioned in Section 39.1 of a single electron undergoing specified motion. If we are given the particle trajectory parameterized by proper time, $\Gamma(\tau)$, then Equation 34.23 gave the 4-current as

$$J(X) = \int d(c \tau) qU(\tau) \delta(\xi)(X - \xi(\tau)).$$  \hspace{1cm} [34.23, page 538]

Substitute into Equation 39.3 with Equation 39.4:

$$\frac{2\pi}{\mu_0 q e} A(X) = \int d^4X^\ast \Theta(X^0 - X^0) \delta(||X - X^\ast||^2) \int d\tau U(\tau) \delta(\xi)(X^\ast - \xi(\tau)).$$  \hspace{1cm} (39.7)

The reasoning here is similar to what we used in magnetostatics, Section 15.5.4 (page 229), and again later in our first look at radiation, Your Turn 25A (page 392).
The retarded time is the time when the past light cone of the observation event intersects the particle trajectory.

The definition implies that \( t_\tau \) is the solution to
\[
ct - ||\vec{r} - \vec{\Gamma}(t_\tau)|| = c\Gamma^0(t_\tau) \quad \text{and} \quad \Gamma^0(t_\tau) < t.
\]

Thus, \( t_\tau \) depends on the observer’s position and time, because \( t_\tau \) does. This remark will be important in a moment,\(^\text{12}\) when we begin varying \( \vec{X} \). Similarly, if we define \( \vec{\beta} = d\vec{r} / d(ct) \) then \( \vec{\beta}_\tau = \vec{\beta}(t_\tau) \) also depends on \( \vec{X} \).

Next, use the remaining delta function in Equation 39.8 to remove the remaining integral, via Equation 34.20 (page 538)
\[
(39.8) = \int d\tau U(\tau) \left| \frac{d}{d\tau} \left( ||X - \Gamma(\tau)||^2 \right) \right|^{-1} \delta(\tau - t_\tau).
\]

The derivative inside the absolute value is
\[
-2(\vec{X} - \vec{\Gamma}(t_\tau))_\mu U^\mu_{\tau}.
\]

Let \( \vec{R}_\tau = \vec{r} - \vec{\Gamma}(t_\tau) \). Then the derivative is
\[
= 2(c(t - t_\tau)U^0_{\tau} - \vec{R}_\tau \cdot \vec{\beta}_\tau \frac{dt}{d\tau} |_{t_\tau})
= 2c \frac{dt}{d\tau} |_{t_\tau} (c(t - t_\tau) - \vec{R}_\tau \cdot \vec{\beta}_\tau).
\]

\(^{12}\)The retarded time also depends implicitly on the particle’s trajectory, but we will hold this fixed.
Equation 39.9 gives the quantity $c(t - t_r)$ as $R_r$. Thus, we have

$$-2(X - \Gamma(t_r))_\mu U^\mu_r = 2c \frac{dt}{dr}|R_r(1 - \vec{R}_r \cdot \vec{\beta}_r).$$

(39.12)

This quantity is always positive, because $||\vec{\beta}|| < 1$, so we may drop the absolute value in Equation 39.10.

Combining Equations 39.7–39.11 yields a compact result called the Liénard–Weichert potential formula:

$$A(X) = \frac{\mu_0 q c}{4\pi} \frac{1}{-(X - \Gamma(t_r)) \cdot U^\mu_r U_r.}$$

where $||X - \Gamma(t_r)||^2 = 0$.

(39.13)

Both parts of Equation 39.13 are manifestly Lorentz-invariant. However, sometimes it is useful to render the first part in another form by using Equation 39.12:

$$A(X) = \frac{\mu_0 q c}{4\pi} \left[ \frac{1}{\vec{\beta}_r} \right] |R_r - \vec{\beta}_r |^{-1}. \quad \text{(39.14)}$$

To use this formula, first solve Equation 39.9 for $\tau_r(X)$ and substitute into Equation 39.14. If we like, we can also replace $\mu_0 c$ by $1/(\varepsilon_0 c)$.

### 39.5.2 Uniform motion yet again

To gain confidence in the Liénard–Weichert potential formula, let’s revisit the problem of a constant-velocity trajectory, whose fields we have already found by other means. Suppose that a point charge $q$ moves along the $z$ axis at speed $\beta c$ and that it passes $\vec{r} = 0$ at $t = 0$. Thus, its trajectory can be written as $\vec{\Gamma}(t) = \beta ct \hat{z}$ and Equation 39.9 says that $t_r$ is the solution of

$$c(t - t_r) = R(t_r).$$

(39.15)
Figure 39.5: Solutions to Equation 39.15. In these figures, the time direction is suppressed but the y direction is shown. An observation is made at P. At the time of observation, a charge is at event Q, at earlier times the charge sits lower on the z axis. (a) Circles centered on βct, of radii c(t − t), for four choices of time t prior to the observation time t. The locations of the charge at those times are shown as stars on the z axis, and circles centered on those points are shown with matching colors. The circles cover the entire plane, so one of them will certainly hit the observation point P. Moreover, the circles never intersect, so only one of them hits P. (If the x direction had been shown, the circles would instead be nested spheres.) (b) Two sticks with lengths in the ratio β (heavy lines) are joined by a hinge. The short stick is fixed while the long one swings around until its endpoint (dashed circle) meets the ray QP. See text for the argument that again establishes a unique solution.

In this simple situation, we will now show explicitly that Equation 39.15 always has exactly one solution.

First proof (1D)

Figure 39.4 is a spacetime diagram that establishes this claim in a special situation, where the observer is sitting on the z axis.

Second proof (> 1D)

Even when that is not the case, we can use rotation invariance to choose coordinates for which x = 0 (although y may not be zero). Figure 39.5a then illustrates that, for any observation point P in the yz plane, exactly one of the circles drawn intersects P. Hence, there is exactly one contribution to Equation 39.14.

Third proof

To evaluate the fields, it will be useful to have yet another graphical proof of the point just made.15 Again suppose that we have been given a choice of field point P, an event whose time is t. Figure 39.5b shows an example, along with the charged particle’s location Q at observation time t. This information determines the angle θ between the line QP and the z axis.

What we need to find is another point, called Q, in Figure 39.5b, which is the charge’s position at some earlier time t. Thus, the distance QQ equals βc(t − t). We want to know

15Recall this chapter’s epigraph.
Chapter 39 Radiation Green Function Revisited

Figure 39.6: **Geometry needed to evaluate Equation 39.14.** We are given observation time $t$ and position $P$ (the point $P$). We also know where the charged particle is located at $t$ (point $Q$). We wish to find a prior point $Q$, on the trajectory that satisfies $R_r = c(t - t_r)$, which will also allow us to evaluate that quantity and the rest of Equation 39.14.

whether we may choose $t_r$ such that also the distance $Q_rP$ equals $c(t - t_r)$ (Equation 39.15), and if so, how many such choices exist.

Imagine two sticks joined by a hinge. The ratio of the sticks’ lengths is $\beta$. Place the free end of the shorter stick at $Q$, and align it along the $z$ axis. Hold the short stick in place and pivot the long stick about the hinge point. The long stick’s end then sweeps out a circle (dashed in the figure), which clearly intersects the ray from $Q$ through $P$ at exactly one point. Now imagine rescaling both sticks’ lengths by a common factor, holding the short one along the $z$ axis with its endpoint always at $Q$. There will always be exactly one rescaling that makes the free endpoint pass through $P$.

**Your Turn 39B**

Figures 39.4–39.5 were drawn assuming that the observer is ahead of the charged particle at the time of observation, that is, $z_o > \beta c t$. Redraw them to make sure they still work in the contrary case.

We again conclude that there is exactly one contribution to Equation 39.14 for any field point $X = \left[ \begin{array}{c} x' \\ y' \\ z' \end{array} \right]$. 

**Evaluation of the potentials and fields**

Now we must evaluate the expression $R_r - \beta \vec{R}_{r,3}$ appearing in Equation 39.14. Figure 39.6 shows a perpendicular dropped from $Q$ to the segment $Q_rP$ in red. Notice that there are two right triangles with a common angle $\psi$, so they are similar: $\triangle Q_rNP \sim \triangle Q_rMQ$, or

$$\frac{R_r}{\vec{R}_{r,3}} = \frac{\beta c(t - t_r)}{Q_rM}.$$ 

Rearranging gives

$$\frac{R_r}{\beta c(t - t_r)} = \frac{\vec{R}_{r,3}}{Q_rM}.$$ 

Also, Equation 39.15 gives $R_r = c(t - t_r)$, so we have

$$\beta \vec{R}_{r,3} = Q_rM.$$
Hence, the quantity we need is
\[
R_r - \beta \vec{R}_{r,3} = R_r - \overrightarrow{Q,\mathbf{M}} = \overrightarrow{M}\mathbf{P}
\]
\[
= \sqrt{\overrightarrow{Q}^2 - \overrightarrow{M}^2} = \sqrt{\vec{r}_\perp^2 + (z_p - \beta ct)^2} - (\beta R_r \sin \psi)^2
\]
\[
= \sqrt{(1 - \beta^2)\vec{r}_\perp^2 + (z_p - \beta ct)^2}.
\]

The square root is always real, because \( \beta < 1 \). Finally, substitute this result into Equation 39.14 and the similar formula for vector potential:
\[
\psi(t, \vec{r}_p) = \frac{q}{4\pi \varepsilon_0} \left( \left(1 - \beta^2\right)\vec{r}_\perp^2 + (z_p - \beta ct)^2 \right)^{-1/2}
\]
\[
\vec{A}(t, \vec{r}_p) = \frac{q\mu_0}{4\pi} \beta c \vec{z} \left( \left(1 - \beta^2\right)\vec{r}_\perp^2 + (z_p - \beta ct)^2 \right)^{-1/2}.
\]

These reproduce the results we got by Lorentz-transforming the fields of a point charge at rest in Section 33.4.2.

We can now find the electric and magnetic fields by changing variables to
\[
\vec{u} = \vec{r}_p - \beta ct \vec{z}
\]
and letting \( g(\vec{u}) = \left( \gamma^{-2}\vec{u}_\perp^2 + \vec{u}_3^2 \right)^{-1/2} \). Thus,
\[
\vec{E}(t, \vec{r}_p) = \frac{q}{4\pi \varepsilon_0} \left( \gamma^{-2}\vec{u}_\perp^2 + \vec{u}_3^2 \right)^{-1/2} \vec{g}
\]
\[
\vec{A}(t, \vec{r}_p) = \frac{q\mu_0}{4\pi} \beta c \vec{z} \left( \gamma^{-2}\vec{u}_\perp^2 + \vec{u}_3^2 \right)^{3/2}.
\]

So using cylindrical coordinates \( u_\perp, \varphi, \vec{u}_3 \),
\[
\vec{B} = \vec{\nabla} \times \vec{A} = \frac{q\mu_0 \beta c}{4\pi} \left( u_\perp \frac{\partial \vec{g}}{\partial \varphi} - \frac{\partial \vec{g}}{\partial u_\perp} \right)
\]
\[
= q \frac{\mu_0 \beta c}{4\pi} (-1)(-\frac{1}{2})g^3 \gamma^{-2}2u_\perp \hat{\varphi} = q \frac{\mu_0 \beta c}{4\pi} \frac{\gamma u_\perp}{u_\perp^2 + \gamma^2\vec{u}_3^2} \hat{\varphi}.
\]

The magnetic field is always pointing in the azimuthal direction.

Next, get the electric field \( \vec{E} = -\vec{\nabla} \psi - d\vec{A} / dt \) by using the Chain Rule:
\[
\vec{E} = \frac{q}{4\pi \varepsilon_0} \left( -\vec{\nabla} g - (\beta/c) \vec{z} (\beta/c) \frac{\partial \vec{g}}{\partial u_3} \right)
\]
\[
= \frac{q}{4\pi \varepsilon_0} \left( -\vec{u}_\perp (-\frac{1}{2})g^3 \gamma^{-2}2u_\perp - \vec{z} (\frac{1}{2})g^3 2u_\perp + z \beta^2 (\frac{1}{2})g^3 2u_3 \right)
\]
\[
= \frac{q}{4\pi \varepsilon_0} g^3 (\vec{u}_\perp \gamma^{-2} u_\perp + \vec{z} \beta^2 u_3).
\]

Note that \( \vec{u}_\perp u_\perp + 2u_3 \) is just \( \vec{u} \). Substitute the definition Equation 39.16. Thus,
\[
\vec{E}(t, \vec{r}_p) = \frac{q\gamma}{4\pi \varepsilon_0} \frac{\vec{r}_p - \beta ct \vec{z}}{(r_\perp^2 + \gamma^2(z_p - \beta ct)^2)^{3/2}}.
\]
Equations 39.17–39.18 are the same results we obtained by applying a Lorentz transformation to the electrostatic field surrounding a static point charge.\(^\text{16}\)

Our solution says that there is a spatial region with large electric and magnetic field strengths, which moves at speed \(\beta c\). The energy flux \(\mathbf{E} \times \mathbf{B}\) is nonzero, but that just describes the translational motion of the lump of energy associated to those fields. It’s not surprising: The charge is surrounded as it moves by a region with fields. A small volume close to the trajectory sees energy flow into it as the particle approaches, then drain back out as the particle recedes, but no energy escapes completely to infinity.

### 39.5.3 Coda

The preceding section was a lot of work just to rediscover results we obtained earlier (Equations 39.17 and 39.18)! One justification is that once we are confident in the Liénard–Weichert formula, we can use it on more difficult problems, such as radiation by an accelerating charge.\(^\text{17}\) Also, even the uniform velocity derivation will be useful in another context (Čerenkov radiation, Chapter 50), where the solution by Lorentz transformation will not be available.

---

\(^{16}\text{See Your Turns 33C and 33D, which however had the particle moving along the } x\text{ axis and evaluated only in the plane } z = 0.}\)

\(^{17}\text{See Problem 42.2.}\)
39.4 Gravitational radiation
Continuing Section 37.2a (page 580), weak gravitational radiation with a source again involves a D'Alembert equation. The Green function we have found therefore also applies to (weak) gravitational radiation.
39.1 Wiggler

Figure 39.7.
CHAPTER 40

Vista: Variational Formulation

The devine nature doth it selfe possesse
In immortallitie, and everlasting peace,
Remoovd farre of from mortall mens affairs,
Neither our sorrows, nor our dangers shares,
Rich in it selfe, of us no want it hath,
Nor moovd with meritts, nor disturbd with wrath.

— Lucy Hutchinson’s 1653 translation of Lucretius (60 BCE)

40.1 FRAMING: NOETHER THEOREM

Our derivation of $\mathcal{I}_{\text{aus}}$ in Chapter 35 may have seemed magical—we desired a result (locally conserved energy and momentum), stated some constraints (Lorentz invariant tensor of the appropriate rank, quadratic function of fields), and found a simple candidate expression that worked with only a little adjustment. But conservation laws should not be magical accidents. In this chapter, we’ll see that they are general consequences of symmetries.

Stepping back a bit, we may notice some habits of highly successful physical theories:

- They are Lorentz invariant.
- They are specified by differential equations, either in time (for particle mechanics) or in spacetime (for fields). That is, they are local; for example, they don’t involve products of field values at two distant points.
- They generally admit a variational formulation; for example, Newton’s Second Law arises as the condition for an action functional to be extremal, and a similar result holds for relativistic mechanics as we review below.

We’ll now see how these themes play out in electrodynamics.¹ Then we’ll see how a variational formulation establishes conservation laws corresponding to continuous invariances of a field theory, a result known as E. Noether’s theorem.²

Electromagnetic phenomenon: .
Physical idea: [[Not ready]].

---

¹K. Schwarzschild obtained the variational formulation in 1903—hence without the benefit of the relativistic invariance that will greatly assist us.
²We will present an extension of Noether’s original result, but she had the key insight. (An unrelated theorem is due to M. Noether.)
40.2 VARIATIONAL FORMULATION OF NEWTONIAN MECHANICS

Given any particle trajectory, we compute its action by evaluating the action functional, which is the time integral of kinetic minus potential energy. For one-dimensional motion,

\[
S[x(t)] = \int_{t_i}^{t_f} dt \; \mathcal{L}(x(t), \frac{dx}{dt}). \quad (40.1)
\]

Here the notation \(S[x(t)]\) means that \(S\) depends on an entire trajectory \(x(t)\). For a particle moving in 1D, the lagrangian density \(\mathcal{L}\) is an ordinary function of two variables, with \(x(t)\) substituted for the first argument and \(\frac{dx}{dt}\) for the second:

\[
\mathcal{L}(x, \frac{dx}{dt}) = KE - PE = \frac{m}{2} \left( \frac{dx}{dt} \right)^2 - U(x).
\]

Let’s characterize those trajectories for which \(S\) is extremal over the space of all trajectories with fixed values at two time points: Substitute \(t = t_i + \Delta t\), where \(x(t_i) = \Delta x(t_i) = 0\). Expanding \(S[x(t) + \Delta x(t)]\) to first order in \(\Delta x(t)\) and using the Chain Rule gives

\[
\Delta S = \int_{t_i}^{t_f} dt \; \Delta \left[ \mathcal{L}(x(t), \frac{dx}{dt}) \right] = \int_{t_i}^{t_f} dt \left( m \frac{dx}{dt} \frac{d^2x}{dt^2} - \Delta x \frac{dU}{dx} \right).
\]

Now integrate the first term by parts. The boundary terms equal zero because we consider only variations that hold the endpoint values fixed:

\[
\Delta S = \int_{t_i}^{t_f} dt \left[ -m \frac{d^2x}{dt^2} - \frac{dU}{dx} \right] \Delta x.
\]

The only way this first-order variation could equal zero for any variation \(\Delta x(t)\) is if the terms in square brackets cancel at each time:

\[
m \frac{d^2x}{dt^2} = - \frac{dU}{dx}. \quad (40.2)
\]

That last formula is Newton’s law.

Generalizing to many interacting particles, we find that we can always re-express newtonian mechanics as a statement about the variation of an action functional of the form Equation 40.1. For example, two masses joined by a spring have

\[
\mathcal{L}(\vec{r}_1, \vec{r}_2, \frac{d\vec{r}_1}{dt}, \frac{d\vec{r}_2}{dt}) = \frac{1}{2} \left( m \left\| \frac{d\vec{r}_1}{dt} \right\|^2 + m \left\| \frac{d\vec{r}_2}{dt} \right\|^2 - k \| \vec{r}_1 - \vec{r}_2 \|^2 \right). \quad (40.3)
\]

Whatever our lagrangian density function, the same reasoning as was given earlier yields the Euler–Lagrange equations

\[
\frac{d}{dt} \left( \frac{\partial \mathcal{L}}{\partial (dx_\alpha/dt)} \right) - \frac{\partial \mathcal{L}}{\partial x_\alpha} = 0. \quad (40.4)
\]

For the example Equation 40.3, the index \(\alpha\) runs over the six components of \(\vec{r}_1\) and \(\vec{r}_2\).

The first term on the left of Equation 40.4 denotes the result when we:

---

3Many authors shorten “lagrangian density” to “lagrangian.”
Differentiate $\mathcal{L}$ with respect to one of its velocity variables, then
Substitute values of $\{x_i\}$ and $\{dx_i/dt\}$, obtaining a function of time, and
Take a derivative with respect to time.

(The second term denotes the derivative with respect to one of the position variables, again followed by substituting values of $\{x_i(t)\}$ and $\{dx_i/dt\}$.)

If, moreover, the action functional has some invariance, for example under translations or rotations, then that fact is also reflected in the resulting equations of motion. For example, Equation 40.3 defines a scalar quantity, invariant under overall rotations; indeed, Your Turn 26A (page 409) involved a set of rotationally-invariant equations. We can also see at a glance that integrating Equation 40.3 over time gives an action that is invariant under spatial or time translations; again, these properties are reflected in the equations of motion.

In short, the lagrangian density is a single function that compactly contains all the dynamics of a mechanical system via its Euler–Lagrange equations, including the invariances of that dynamics.

40.3 VARIATIONAL FORMULATION OF FIELD EQUATIONS

40.3.1 Local lagrangian densities and their variational equations

We now upgrade the variational formulation to accommodate fields. Consider action functionals of fields of the generic form

$$S[\text{traj}] = \int d^4X \mathcal{L}(\phi, \partial \phi),$$

(40.5)

where for a scalar field $\phi$, the lagrangian density $\mathcal{L}$ is an ordinary function of five variables (the field and its four space and time derivatives at any point). Action functionals of this form are called local. More generally, for a multicomponent field (for example, the 4-vector potential in electrodynamics), $\mathcal{L}$ is a local function of five variables for each component.

Given a “trajectory” (field history) $\phi(X)$, Equation 40.5 instructs us to substitute it and its derivatives into $\mathcal{L}$ and integrate the resulting function over spacetime.

We will also require that $\mathcal{L}$ be a 4-scalar function of the fields. Because $d^4X$ is also a 4-scalar,\(^4\) therefore $S$ will be a Lorentz scalar. Also, the field theories we will consider are invariant under translations (in space or time).

Adapting the preceding discussion, instead of Equation 40.4 we get the generic variational field equation

$$\frac{\partial}{\partial \mu} \left( \frac{\partial \mathcal{L}}{\partial (\partial_{\mu} \phi)} \right) - \frac{\partial \mathcal{L}}{\partial \phi} = 0. \quad \textbf{Euler–Lagrange equation (fields)} \quad (40.6)$$

This time, the first term on the left denotes the result when we:

\(^4\)The Rules: Section 34.4. Or take the determinant of both sides of Equation 32.17 (page 498).
Find derivatives of $\mathcal{L}$ with respect to one of its four derivative variables, then substitute values of $\phi$ and $\partial_\mu \phi$, obtaining a function of $X$, and only then take a derivative with respect to $X^\mu$.

(The second term denotes the derivative with respect to undifferentiated $\phi$, again followed by substituting values of $\phi$ and $\partial_\mu \phi$.)

A specific choice for the lagrangian density of a scalar field could be a constant times

$$\mathcal{L}(\phi, \partial_\mu \phi) = \frac{1}{2}( -\partial^\mu \phi \partial_\mu \phi - \lambda^{-2} \phi^2).$$

(40.7)

Evaluating Equation 40.6 then gives $-\partial^\mu \partial_\mu \phi - \lambda^{-2} \phi = 0$, sometimes called the Yukawa equation for its role in an early theory of nuclear forces.

**Your Turn 40A**

Show that the static solutions of the Yukawa equation fall exponentially with distance as $\exp(-r/\lambda)/r$, and hence could mediate short-range interactions (such as the nuclear force).

40.3.2 A simple lagrangian density leads to the Maxwell equations in vacuum

Can we find an action functional meeting all of the symmetry requirements, and whose variational equation recovers the Maxwell equations? Let’s begin with some “Einstein thinking.”

We may begin by formulating fields in terms of the 4-vector potential $A$, so that half of the Maxwell equations become identities, not dynamical equations. The remaining equations are linear in $A$, and second-order in its derivatives (Equation 34.16, page 536). So $\mathcal{L}$ must be a quadratic function of $A$, with at most two derivatives. It should be a 4-scalar, to ensure Lorentz-invariant field equations of motion, as well as being gauge- and translation invariant. There are very few such functions:

- The expressions $(\partial_\mu A^\mu)^2$ and $\partial_\mu \Box A^\mu$ are not gauge invariant. The expression $A_\mu A^\mu$ is not gauge invariant, and moreover contains no derivatives.
- $F^{\mu\nu} F_{\mu\nu}$ is the only remaining option, so we now explore it.

Following our recipe, we find the first order variation of our candidate action functional $S[A(X)] = (\text{const}) \times \int d^4 X F^{\mu\nu} F_{\mu\nu}$ and ask under what condition it will equal zero:

$$\Delta S = 0 = 2 \int d^4 X F^{\mu\nu} (\Delta F_{\mu\nu}) = 2 \int d^4 X F^{\mu\nu} (\partial_\mu \Delta A^\nu - \partial^\nu \Delta A_\mu)$$

$$= 4 \int d^4 X F^{\mu\nu} (\partial_\mu \Delta A^\nu)$$

$$= -4 \int d^4 X (\partial_\mu F_{\mu\nu}) \Delta A^\nu. \quad (40.8)$$

**Your Turn 341 (page 536).**

---

5When quantized, the field $\phi$ was once associated to particle states that could represent pions.

6Your Turn 341 (page 536).
For this quantity to vanish regardless of $\Delta A$, we must have that $\partial \mu F_{\mu \nu} = 0$. That indeed completes the Maxwell equations in vacuum.\footnote{Set $\mathcal{L} = 0$ in the first of Equations 34.13 (page 535).}

Section 40.3.2' (page 616) mentions another candidate term for the lagrangian density.

### 40.3.3 Fields plus charged particles

Suppose that charged particles are present and executing prescribed motions; that is, we don’t inquire yet into the equations of motion for particles. We can construct the charge flux 4-vector $\mathcal{J}$ as in Section 34.9.2 (page 538). Then $\mathcal{J}$ obeys the continuity equation for charge,

$$\partial_{\mu} \mathcal{J}^{\mu} = 0. \quad [34.11, \text{page 534}]$$

We may add $A_{\mu} \mathcal{J}^{\mu}$ to our lagrangian density because:

- This term is Lorentz scalar and its integral over $d^4X$ is translation invariant.
- $\mathcal{J}$ is gauge invariant, so under gauge transformation by $\Xi$ we have (Equation 34.15, page 536)

$$\int d^4X J_{\mu} A^{\mu} \mapsto \int d^4X (J_{\mu} A^{\mu} + J_{\mu} \hat{\partial}_{\mu} \Xi).$$

Integrating by parts shows that the second term equals zero.
- This term is linear in $A$, so it will contribute a term to the variational equations of order zero in $A$, as desired for the source term in the Maxwell equations.

Combining the pure-field term from Section 40.3.2 with the particle term just found and choosing constants that give appropriate units gives finally

$$\mathcal{L}(A, \hat{\partial} A) = \mathcal{L}_{f} + \mathcal{L}_{fp} = \frac{1}{e} (-\frac{1}{4\mu_0} F_{\mu \nu} F^{\mu \nu} + A_{\mu} \mathcal{J}^{\mu}). \quad (40.11)$$

The two terms are labeled $f$ for the field part and $fp$ for the field–particle interaction. The overall factor of $1/c$ gives our action functional the traditional units ($J$ s). The factor $-1/(4\mu_0)$ will be justified when you work out:

**Your Turn 40B**

Show that the corresponding Euler–Lagrange equations are indeed Maxwell with charges and currents (Equation 34.13, page 535 or Equation 34.16, page 536).

Until now, we have assumed that particle motions were given. We can extend the theory to include equations of motion for the particles as well as the fields by adding a kinetic energy term for each one. To find the appropriate expression, we once again resort to “Einstein thinking.” The action of a free particle must be a single number computed from the particle’s trajectory $\Gamma (\xi)$. It must be Lorentz scalar and local. The only obvious choice is the total elapsed proper time, $\int dr$. That quantity has the wrong units, but we
can correct that defect with the available constants (particle mass and speed of light), yielding a proposal for the particle part of the action:\(^8\)

\[
S_p[\text{trajectory}] = -mc \int d\xi \sqrt{-\left|d\Gamma/d\xi\right|^2}.
\]  

(40.12)

There is a tricky point here. Proper-time parametrization is often convenient, but proper time implicitly depends on the trajectory’s speed. We want all dependence on the trajectory to be explicit so that we know what we’re doing when we vary it. Luckily, Equation 40.12 is invariant under change of parameter, so we may specify that \(\xi\) is any fixed parameter choice, for example, covering the fixed range from 0 to 1, and consider variations \(\Delta\Gamma\) that equal zero at \(\xi = 0, 1\). After computing the variations we need, at the end we can, if we wish, specialize to proper-time parameterization.

If many point charges are present, we give each one its own kinetic energy term. Ex. Add Equation 40.12 to the integral of Equation 40.11 and use Equation 34.23 to express the charge 4-current in terms of the particle trajectory. Show that the complete action functional thus obtained leads to a variational equation that is precisely the Lorentz force law for the particle (Equation 33.3, page 512).

Solution: \(\mathcal{L}_f\) does not depend on the trajectory, so we wish to find the first-order variations of \(\delta A_\mu(X)\) and of \(\delta p\). Start with the first of these:

\[
\delta S_{ fp} = \frac{1}{c} \int d^4X \frac{\partial A_\mu(X)}{\partial X} q \int (c d\xi) \delta^{(d)}(X - \Gamma(\xi) - \Delta\Gamma(\xi)) \left( \frac{d\Gamma}{d\xi} + \frac{d\Delta\Gamma}{d\xi} \right)\mu
\]

Taylor expand to find the displaced 4-vector potential, then integrate by parts:

\[
\Delta S_{ fp} = q \int d\xi \left( \frac{\partial A_\mu}{\partial X} \right)_{\Gamma(\xi)} \Delta\Gamma^\nu \frac{d\Gamma^\mu}{d\xi} - \Delta\Gamma^\mu \frac{\partial A_\mu}{\partial \Gamma^\nu} \left( \frac{d\Gamma^\nu}{d\xi} \right)_{\Gamma(\xi)}
\]

We can pull out a common factor if we first rename the indices in the second term:

\[
\mu \rightarrow \nu \text{ and } \nu \rightarrow \mu
\]

\[
\Delta S_{ fp} = q \int d\xi \Delta\Gamma^\nu \left( \frac{\partial A_\mu}{\partial X} \frac{d\Gamma^\mu}{d\xi} - \frac{\partial A_\nu}{\partial X} \frac{d\Gamma^\nu}{d\xi} \right) = q \int d\xi \Delta\Gamma^\nu F_{\nu\mu}(\Gamma(\tau)) \frac{d\Gamma^\mu}{d\xi}.
\]

Next, turn to the kinetic term and expand in \(\Delta\Gamma\):

\[
S_p + \Delta S_p = -mc \int d\xi \left( -\left\| \frac{d\Gamma}{d\xi} \right\|^2 \frac{d^4\Gamma}{d\xi} \| \right)^{1/2}
\]

\[
= -mc \int d\xi \left( -\left\| \frac{d\Gamma}{d\xi} \right\|^2 - 2 \frac{d\Gamma^\mu}{d\xi} \frac{d\Delta\Gamma^\mu}{d\xi} + \ldots \right)^{1/2}
\]

\(^8\)See Equation 32.21 (page 500). The minus sign is needed because the length-squared of a timelike vector is negative. Interestingly, this action is formally the same as the transit time that led to Fermat’s principle in geometrical optics (Section 21.5.4, page 333), but promoted to Minkowski space.
= -mc \int \xi \left( -\frac{d\Gamma}{d\xi} \right)^2 \left( 1 - \frac{2}{-\|d\Gamma/d\xi\|^2} \frac{d\Gamma^\mu}{d\xi} \frac{d\Delta\Gamma^\mu}{d\xi} + \cdots \right)^{1/2}

\Delta S_\rho = mc \int d\xi \left( -\frac{d\Gamma}{d\xi} \right)^2 \left( 1 - \frac{2}{-\|d\Gamma/d\xi\|^2} \frac{d\Gamma^\mu}{d\xi} \frac{d\Delta\Gamma^\mu}{d\xi} \right)^{1/2}

It is safe now to specialize to proper-time parameterization, \( \xi = \tau \), and use Equation 32.25 (page 502). Integrating by parts gives

= -mc \int d\xi \Delta\Gamma_\nu \frac{d^2\Gamma^\nu}{d\tau^2} (c^2)^{-1/2}.

At last we can combine our two terms for the first-order variation and ask that they equal zero for arbitrary \( \Delta\Gamma \). This happens only if the trajectory everywhere satisfies

0 = qF_{\nu\mu} \frac{d\Gamma^\mu}{d\tau} - m \frac{d^2\Gamma^\nu}{d\tau^2}.

We have arrived the Lorentz force law.

In short, we have found that all of electrodynamics admits a formulation as a variational principle. Instead of starting with the Maxwell equations and the Lorentz force law, we can specify the theory with the action functional given above.

### 40.4 Continuous Invariances Lead to Conservation Laws

#### 40.4.1 Scalar field example

Section 40.1 promised to show that conservation laws are linked to invariances. Before investigating this claim in electrodynamics, let’s again begin with a simpler system, consisting of a single scalar field \( \phi \). We now explore the consequences of a continuous symmetry, that is, a field transformation that leaves the equations of motion form-invariant and that changes fields by an infinitesimal amount. Accordingly, consider a general local transformation, that is, one for which

\[ \phi(X) \rightarrow \tilde{\phi}(X) = \phi(X) + \epsilon D[\phi, \tilde{\phi}] + \cdots. \] (40.13)

Here the ellipsis denotes terms of higher order in a bookkeeping parameter \( \epsilon \); from now on, we will drop such terms without comment. \( D \) is a local expression in fields and their derivatives, which is to be evaluated at each spacetime point \( X \). We suppose that the action \( S \) is unchanged under substituting the expression just given, for any trajectory \( \phi(X) \); then we ask for consequences in the situation where \( \phi \) also obeys the variational equation associated to its action functional.

Here are two examples:

- A translation (shift of \( X \) by a constant 4-vector \( \epsilon \hat{b} \)) corresponds to the local functional \( D[\phi, \tilde{\phi}] = b^\mu \tilde{\partial}_\mu \phi \), as we see by Taylor expanding \( \phi \).
Chapter 40 Vista: Variational Formulation

Next, consider a set of two scalar fields, each with its own lagrangian density of the form Equation 40.7. Then

\[ D \left[ \begin{bmatrix} \dot{\phi}_1(t) \\ \dot{\phi}_2(t) \end{bmatrix} \right] = \left[ \begin{bmatrix} -\dot{\phi}_1(t) \\ \dot{\phi}_2(t) \end{bmatrix} \right] = T \left[ \begin{bmatrix} \phi_1(t) \\ \phi_2(t) \end{bmatrix} \right] \] (40.14)

implements an infinitesimal rotation in the internal space of \( \phi \)'s components (not in physical space). Here \( T \) is the generator of rotations in internal space (Section 3.7.2, page 46).

40.4.2 Consequences of invariance: the Noether theorem

We cannot assume that the lagrangian density is unchanged by an invariance, but we at least know that its change, if any, must be a total derivative (because its integral, the action, was assumed to be unchanged upon substituting Equation 40.13). Thus, for each infinitesimal invariance of the system we must have

\[ \mathcal{L} \to \mathcal{L}(\hat{\phi}, \partial_t \hat{\phi}) = \mathcal{L}(\phi, \partial_t \phi) + \varepsilon \partial_\mu M^\mu(\phi, \partial_t \phi). \] (40.15)

Here \( M^\mu \) is some local functional of fields and their derivatives that we can find from the chosen lagrangian density and the invariance under consideration. Continuing the two examples in the preceding section, Equation 40.7 gives

* For translation by \( h \),

\[ M^\mu = \frac{1}{2} b^\mu \hat{\phi} \partial_t \phi - \frac{1}{2} \lambda^{-2} b^\mu \phi^2 = \frac{1}{2} b^\mu (\|\hat{\phi}\|^2 - \lambda^{-2} \phi^2). \]

**Your Turn 40C**

Find \( M \) for the case of internal rotations (Equation 40.14).

We will now find a 4-vector field associated to our assumed invariance that obeys a continuity equation, and hence defines a conserved “charge,” when \( \phi \) is a solution of the variational equation. To do this, first substitute Equation 40.13 into Equation 40.15:

\[ \mathcal{L}(\phi, \partial_t \phi) + \varepsilon D[\cdots] \left( \frac{\partial \mathcal{L}}{\partial \phi} \right) + \varepsilon (\partial_\mu D[\cdots]) \frac{\partial \mathcal{L}}{\partial (\partial_\mu \phi)} + O(e^2) = \mathcal{L}(\phi, \partial_t \phi) + \varepsilon \partial_\mu M^\mu + O(e^2). \]

Next, rephrase the first term by using the Euler–Lagrange equation. Comparing the sides of this equation then shows that the 4-vector quantity\(^9\)

\[ \mathcal{J}^\mu = \frac{\partial \mathcal{L}}{\partial (\partial_\mu \phi)} D[\phi, \partial_t \phi] - M^\mu \] obeys \( \partial_\mu \mathcal{J}^\mu = 0 \) (40.16)

for any field trajectory that solves the equations of motion.

Equation 40.16 is the identity we were seeking, often called the **Noether theorem**. Returning to our two examples,

---

\(^9\)We introduced the new generic symbol \( \mathcal{J}^\mu \) for the flux under construction, to distinguish it from \( J^\mu \) which is always specifically electric charge flux.
For translation by \( \mathbf{b} \),

\[
\mathcal{J}^\sigma = -\mathcal{D}^\sigma \phi \mathcal{D}_\mu \phi + \left( -\frac{1}{2} b^\sigma \mathcal{D}^\mu \phi + \frac{1}{2} \lambda^{-2} b^\sigma \phi^3 \right). \tag{40.17}
\]

We can summarize all four of the associated continuity equations as

\[
\mathcal{D}_\mu \mathcal{T}^{\mu\nu} = 0 \quad \text{where} \quad \mathcal{T}^{\mu\nu} = -\mathcal{D}^\mu \phi \mathcal{D}_\nu \phi - \frac{1}{2} \lambda^{-2} (\mathcal{D}^\mu \phi \mathcal{D}_\nu \phi - \lambda^{-2} \phi^2). \tag{40.18}
\]

In fact, the symmetric tensor \( \mathcal{T} \) just defined is the energy–momentum flux 4-tensor of the scalar field theory under consideration. Tracing the derivation reveals that

\textit{Time-translation invariance implies energy conservation, whereas the three spatial translation invariances imply conservation of momentum.}

For the internal rotation invariance example, \( \mathcal{J}^\mu = -(\mathcal{D}^\mu \phi) \mathcal{I} \). Its corresponding conserved quantity is then \( \int d^3 r \; \mathcal{J}^0 \) (see Equation 8.6, page 118).

### 40.4.3 Translational invariance of electrodynamics leads to the same \( \mathcal{T} \) as was found previously

Let’s upgrade these ideas to electrodynamics, with the lagrangian density Equation 40.11. The derivation is a bit subtler than in the scalar field case because, in addition to translation invariance, electrodynamics is also gauge invariant. It will be most convenient to consider a combined operation, in which an infinitesimal translation by \( \varepsilon \mathbf{b} \) is combined with a gauge transformation by \( \mathcal{A}^\mu \).

\[
\mathcal{A}^\mu = \mathcal{A}^\mu + \varepsilon \mathbf{b} \lambda (\mathcal{D}_\lambda \mathcal{A}^\mu - \mathcal{D}^\mu \mathcal{A}_\lambda),
\]

where the second term is translation and the third is gauge transformation.

The recipe given earlier starts by working out

\[
D[\mathcal{A}, \mathcal{D} \mathcal{A}]^\mu = \mathcal{D}_\lambda F^\mu_\lambda.
\tag{40.19}
\]

Equation 40.11 gives the change of field lagrangian density as

\[
\Delta \mathcal{L} = \frac{1}{2 \mu_0 c} \left[ \mathcal{D}_\mu (\varepsilon \mathcal{D}_\lambda F^\lambda_\mu) - \mathcal{D}_\mu (\varepsilon \mathcal{A}_\lambda F^\lambda_\mu) \right] F^{\mu\nu}.
\]

Although we may not use the variational equations to simplify this expression, the homogeneous Maxwell equations are fair game because they are identities, consequences of our decision to use the 4-vector potential as our dynamical variables.\(^{11}\) Thus, we may replace \( \mathcal{D}_\mu F^\lambda_\nu \) by \( \mathcal{D}_\nu F^\nu_\mu - \mathcal{D}_\mu F^\lambda_\nu \):
So indeed, the change is a total derivative (Equation 40.15), with

\[
\frac{\mathcal{M}_\lambda}{\mathcal{M}_\lambda} = \frac{-1}{4\mu_0 c} b_\lambda E_{\mu \nu} F^{\mu \nu}.
\]  

(40.20)

Hence, for any \( b \) we get a continuity equation for the quantities analogous to Equation 40.16:

\[
\mathcal{J}^\lambda = \frac{\delta \mathcal{L}}{\delta (\partial_\mu A^\nu)} \partial^\mu \left[ \partial_\nu A^\lambda - \mathcal{M}^\lambda \right]
= \frac{1}{\mu_0 c} \left[ F^{\lambda \nu} b_\sigma F_{\sigma \nu} + \frac{1}{4} b_\lambda E_{\mu \nu} F^{\mu \nu} \right] = b_\sigma \frac{1}{\mu_0 c} \left[ -F^{\lambda \nu} F_{\sigma \nu} + \frac{1}{4} \epsilon^{\sigma \lambda F_{\mu \nu} F^{\mu \nu}} \right].
\]  

(40.21)

This expression is the contraction of \(-b/c\) with the electromagnetic energy–momentum flux tensor obtained in Chapter 35.\(^{12}\) Its continuity equation, (40.16), is the result we already found in Section 35.5 (page 559), but now exposed as a consequence of translation symmetry.

### 40.5 PLUS ULTRA

- Remarkably, the classical limits of all known fundamental physical theories are expressible as variational principles. There may not be any satisfying “explanation” for this grand overarching theme of physics, but perhaps it’s relevant that the quantum version of any such theory can be straightforwardly constructed by a path integral: Simply divide the action by \( \hbar \) (which has units of action), multiply by \( \sqrt{-1} \), and exponentiate to obtain a phase. Integrating that phase over all trajectories yields quantum amplitudes.

- The emphasis we have given to conserved quantities may seem puzzling: In classical physics, one can always take the complete solution and evolve backward in time to time zero, so that every feature gives rise to a “constant of the motion.” What we have found is that, in a local field theory, continuous invariances give rise to conserved quantities that are local, and hence additive over objects that start and end well separated by vacuum (for example, Equation 40.18 or its electrodynamic analog). These are the sorts of conservation laws that are useful for understanding collisions.

- This chapter did not claim that symmetry was the only way to get conservation laws. Special field theories in one space and one time dimension can actually have infinitely many local conserved quantities (they are integrable), despite being interacting.

### FURTHER READING

*Semipopular:*
Intermediate:
Variational principles in general: Feynman et al., 2010a, chap. 19.
Variational formulation of electrodynamics, field-theoretic Noether theorem: Coleman, 2019; Freeman et al., 2019; José & Saletan, 1998; Zangwill, 2013, §24.4; Melia, 2001, chap. 6; Peskin & Schroeder, 1995, chap. 2; Weinberg, 2005b. [[Reader, please help me out: Section 40.4.3 opens with an elegant move (see footnote 10), but I cannot remember who taught it to me, nor have I found it in any of the well-known textbooks I consulted. I am not now, nor ever have been, clever enough to invent this gambit, so if you know who did, or even where it may appear in print, I'd like to hear.]]

Technical:
Historical: Noether, 1918; English translation at arxiv.org/abs/physics/0503066.

 Supersymmetry: Weinberg, 2005a; Schwarz & Schwarz, 2022.
40.3.2’ Axion term
The main text discarded the candidate term because it is a total 4-divergence. Therefore its integral over $d^4X$ is a boundary term, by the divergence theorem, and hence makes no contribution to the local variation of $S$.

There is an interesting loophole, however. On a topologically nontrivial space, this term need not integrate to zero, because the 4-vector potential may not be globally well defined. Then this “axion” term may affect the physics of, for example, topological insulators (Vazifeh & Franz, 2010).

40.4’ Angular momentum
Applying Noether’s theorem to infinitesimal translations led us to a continuity equation for the energy–momentum flux tensor, and thence to conservation laws. Similarly, applying it to infinitesimal Lorentz transformations leads to a continuity equation for the angular momentum flux tensor (Section 35.5’a, page 564). Of the resulting six conserved quantities, the ones associated to spatial rotations are angular momenta. The ones associated to Lorentz boosts involve the velocity of the system’s overall center of energy (the relativistic version of center of mass), which is also conserved.

40.4’b Classical fermion fields and supersymmetry
At least at the symbolic level, the analysis of this chapter can be extended to include classical fields that, when quantized, lead to fermionic particles (in contrast to the scalar and vector fields we considered). The mind-boggling insight is that the appropriate fields must take their values in an anticommuting number system, not the usual real numbers.

More remarkable still, the introduction of such fields leads to the possibility of transformations some of whose parameters (generalizing $b$ and $\bar{b}$ in the main text) are also anticommuting variables. Although it sounds like moonshine, theories with such supersymmetry can be written, and a generalized Noether theorem can be written leading to fermionic conserved quantities for the new class of invariances. We have seen that the parameters of a symmetry transformation may themselves transform, for example under rotations. The anticommuting parameters of supersymmetry transformations transform as spinors under rotations and other Lorentz transformations.

Supersymmetric field theories have many theoretically attractive features, and arise in models for real condensed matter phenomena. Their use to describe fundamental particles remains an intriguing unsettled possibility.
40.1 Consequences of galilean invariance
Illustrate the reasoning in Section 40.4 with a Newtonian system: two masses joined
by a spring (Equation 40.3, page 606). There are ten symmetries corresponding to the
infinitesimal Galilean group transformations.
Beams, Radiation and Scattering

Optical studies from Roger Bacon’s *De multiplicatione specierum*. The diagram shows light being refracted by a spherical glass container full of water.
41.1 FRAMING: DIFFRACTIVE SPREADING

Earlier chapters gave several examples of monochromatic (single-frequency) solutions to the vacuum Maxwell equations. Chapter 37 gave one of the simplest: a linearly polarized plane wave:

\[
\vec{A}(t, \vec{r}) = \frac{1}{2} \vec{\xi} e^{-i\omega t + ikz} + \text{c.c.,} \tag{37.4, page 575}
\]

where \( \omega = kc \) and \( \vec{\xi} \) is a real, constant vector. (In Lorenz gauge,\(^1\) there is also a scalar potential \( \psi \).) Another simple solution is a circularly polarized plane wave, obtained by replacing \( \vec{\xi} \) in the previous formula by one of the complex unit vectors that we called \( \vec{\xi}_{(+)} \) (Equation 18.34, page 285) times a suitable constant. There are also spherical waves, for example the ones found in Section 38.2.1. None of these solutions, however, could be called a “beam” of light. A beam travels mainly in one direction, unlike a spherical wave, and has finite width, unlike either spherical or plane waves.

Chapter 21 gave one approach to beams of light, but in an approximation that neglected diffraction. When a perfect plane wave encounters an opaque barrier with a pinhole, Chapter 22 explored diffractive spreading that results. Naively we might expect the light to spray out on the other side in all directions, not just the one direction it was traveling in to begin with, like a water wave hitting a narrow gap in the wall surrounding a lagoon.

And yet, in everyday experience, when light from a distant source (such as the Sun), or from a near but extremely parallel source (such as a laser) hits a finite-size pinhole, it continues for many wavelengths without spreading much. So in some approximate sense, a real beam seems physically possible. This chapter will begin by analyzing and quantifying that statement. Then we will find some exotic beams that have recently proven very useful in applications.\(^2\)

\textit{Electromagnetic phenomenon:} A structured beam of light can transfer far more angular momentum than a circularly polarized plane wave.

\textit{Physical idea:} Optical vortex beams exert gradient forces and torques.

41.2 GAUSSIAN BEAMS

---

\(^1\)Section 37.3 (page 575), Section 37.2.

\(^2\)See Section 41.4.3.
Let’s look for vacuum solutions to the Lorenz-gauge Maxwell equations (Equation 37.2, page 575) that have definite angular frequency \( \omega = k_c \). That is, we seek functions of the form \( \mathbf{A}(X) = \frac{1}{2} \mathbf{A}(r)e^{-ik_cct} + \text{c.c.} \), where

\[
\begin{align*}
\left( -ik_c \right)^2 + \nabla^2 A &= 0 \quad \text{Helmholtz equation (41.1)}
\end{align*}
\]

and Lorenz gauge \( \partial_{\mu}A^\mu \) says

\[
-ik_c A^0 + \nabla \cdot A = 0. \tag{41.2}
\]

41.2.1 Paraxial approximation creates a mathematical analogy to the Schrödinger equation

The preceding equations are exact, but now we’ll look for a linearly polarized solution traveling mostly along \( \hat{z} \) that has nearly planar wavefronts close to the \( z \) axis, but which is limited in its extent along \( x, y \). To be very specific, suppose that the frequency is appropriate to red visible light, \( \omega = 2\pi c/(600 \text{ nm}) \), and the transverse extent is circular with radius \( \omega = 1 \text{ mm} \), similar to the beam of a laser pointer. We will use physical reasoning to motivate a class of trial solutions, then substitute into the Maxwell equations to nail down an unknown function.

We may hope that such a solution would take a form reminiscent of those studied in Chapter 21: In some region, the solution is a plane wave modulated by a slowly varying function of coordinates:

\[
\mathbf{A}(t, \mathbf{r}) = \frac{1}{2} \mathbf{A}(\mathbf{r})e^{ik_c(\mathbf{r} - \mathbf{ct})} + \text{c.c.} \tag{41.3}
\]

Let’s also attempt to find a solution for which the direction of polarization is constant: \( \mathbf{\hat{z}}(\mathbf{r}) = \mathbf{\hat{z}} u(\mathbf{r}) \), where \( \mathbf{\hat{z}} \perp \hat{z} \). Substituting into the Helmholtz equation yields

\[
\frac{1}{2}(\nabla^2 u + 2ik_c \nabla_z u - k_c^2 u + k_c^2 u)e^{-ik_c(\mathbf{r} - \mathbf{ct})} = 0.
\]

We can now give a precise meaning to the phrase “traveling mostly along \( \hat{z} \)”: If

\[
(\nabla_z)^2 u \ll 2k_c \nabla_z u \quad \text{and also} \quad (\nabla_{z_\perp})^2 u, \quad \text{paraxial conditions (41.4)}
\]

then we may drop the corresponding term in the wave equation, leaving

\[
(\nabla_{z_\perp})^2 u + 2ik_c \nabla_z u = 0. \quad \text{paraxial equation (41.5)}
\]

The paraxial approximation (dropping \( (\nabla_z)^2 \)) is valid when the length scale over which \( u \) changes along \( z \) is much larger than both its transverse extent, and also the wavelength.

The paraxial equation is an old friend: It’s mathematically equivalent to the Schrödinger equation in two dimensions, with \( z \) playing the role of “time.”

---

3There will also be a scalar potential dictated by Lorenz gauge (Problem 41.4). Had we chosen to work in Coulomb gauge, then there would be no scalar potential (Section 18.8.3), but there would be a complicating constraint on \( A \).
It will shorten our formulas to nondimensionalize. Because the relevant transverse length scale is $\omega$, let $\tilde{x} = x/\omega$ and $\tilde{y} = y/\omega$. Because our problem is not isotropic, it will be convenient to rescale $z$ differently,\footnote{The length scale $\frac{1}{2} k_s \omega^2$ is sometimes called the \textbf{Rayleigh range}.} letting $\tilde{z} = 2z/(k_s \omega^2)$. Then the paraxial equation takes a universal form
\begin{equation}
(\tilde{V}_\perp)^2 u + 4i \frac{\partial u}{\partial \tilde{z}} = 0.
\end{equation}

\textbf{Your Turn 41A}
\begin{enumerate}
\item Establish Equation 41.6.
\item Show that its conditions for applicability are that $(4 \omega k_s)^{-1} \frac{\partial^2 u}{\partial \tilde{z}^2} \ll \text{both } 4 \frac{\partial u}{\partial \tilde{z}} \text{ and also } \ll (\tilde{V}_\perp)^2 u$.
\end{enumerate}

You have some relevant experience with the Schrödinger equation from earlier courses. For example, you know that a spatially confined wavepacket will spread over time, according to the Uncertainty Principle, and that a gaussian profile will spread gradually while retaining a gaussian form. In the present context, we conclude that the corresponding beam of light, whose amplitude falls off rapidly in the direction transverse to propagation, will retain that property, and hence remain within the domain of the paraxial approximation. Let’s make that expectation precise.

Call the dimensionless transverse radial distance $\tilde{\rho} = ||\tilde{r}_\perp||/\omega$. A circular, gaussian wavepacket at $\tilde{z} = 0$ takes the form $\exp(-\tilde{\rho}^2)$, but we must allow for it to spread and distort. So consider a trial solution of the form
\begin{equation}
\tilde{u}(\tilde{z}, \tilde{r}_\perp) = \exp(M), \text{ where } M(\tilde{z}, \tilde{r}_\perp) = -\tilde{\rho}^2 + \tilde{z}S(\tilde{\rho}) + \cdots.
\end{equation}

In these formulas, $S$ is a function we must find and the ellipsis represents higher-order terms in $\tilde{z}$. We will be content with just the first-order correction, because in practice $\tilde{z}$ is numerically small.

\textbf{Your Turn 41B}

Estimate $k_s \omega^2$ using the typical numbers appearing earlier, and compare to the size of a typical lab setup. If we restrict to $z/(k_s \omega^2) \ll 1$, will that impose a severe limitation on the applicability of our calculation for tabletop apparatus?

\subsection{41.2.2 The gaussian beam spreads slowly, although its wavefronts curve}

The preceding section argued that

\begin{quote}
We seek a solution to the Maxwell equations of the form Equations 41.3 with 41.7, satisfying the condition stated in Equation 41.6, in the region with $\tilde{z} \ll 1$.
\end{quote}

To find it, we now solve Equation 41.6.
Set up cylindrical coordinates $\rho, \varphi, z$. Recall\(^5\) that the two-dimensional Laplace operator is $(\nabla_{\perp})^2 = (\partial/\partial \rho + \rho^{-1})^2(\partial/\partial \varphi) + \rho^{-2}(\partial^2/\partial \varphi^2)$. In the axially-symmetric case we are studying, the paraxial equation therefore becomes

$$[(\partial/\partial \rho + \rho^{-1})^2(\partial/\partial \varphi) + \rho^{-2}\partial^2/\partial \varphi^2 + 4i\partial/\partial z]e^M = 0$$

(41.8)

$$0 = (\partial/\partial \rho + \rho^{-1})((-2\rho + zS')e^M) + 4iSe^M$$

$$0 = -2 + 2S'' + (-2\rho + zS')^2 - 2(z\rho^{-1}S' + 4iS).$$

(41.9)

We can now approximate by dropping all terms with the small quantity\(^6\) $\tilde{z}$, to find $S \approx i(-1 + \rho^2)$. So the trial solution Equation 41.7 is

$$u \approx \exp(-\tilde{\rho}^2 + i\tilde{z}(-1 + \rho^2) + \cdots).$$

(41.10)

Inspecting the modulus and phase of our solution, we see that

To first order in $\tilde{z}$, the beam width does not change, but its phase acquires a $\rho$ dependence.

Equations 41.3 and 41.10 give the small-$\tilde{z}$ limit of a solution to the paraxial equation called the \textit{gaussian beam}. In fact, many lasers do generate beams well approximated as gaussian.

Having found a solution, we should now confirm that it satisfies the conditions for paraxial approximation to be valid. The gaussian beam profile cuts off at around 1, so the conditions in Your Turn 41A both amount to $(k, \omega)^{-2} \ll 1$, which is certainly well satisfied in our example situation. Thus, Equation 41.10 indeed gives a beam that does not spread appreciably over typical lab dimensions. Had we kept the terms with $\tilde{z}$ in Equation 41.9, and higher powers in $\tilde{z}$ dropped in Equation 41.10, then we would have found eventual diffractive spreading away from the \textit{beam waist} at $z = 0$ (and also nonconstant $\zeta$).

Now put the pieces together (Equations 41.3 and 41.10):

**Your Turn 41C**

Examine the phase of our solution at $t = 0$ and show that already at first order in $\tilde{z}$, the wavefronts begin to curve as $\tilde{z}$ increases from zero. [\textit{Hint:} A wavefront is a level set of the phase. The phase of a wave is the imaginary part of the log of its complex amplitude.]

### 41.3 Optical-Vortex Beams Transport Angular Momentum Even When Linearly Polarized

Sections 20.4–20.5 mentioned that a circularly-polarized plane wave transports angular momentum along its direction of propagation. You found in Problem 20.2 that this angular

---

\(^5\)See Section 5.3.2 (page 68).

\(^6\)Notice that $\tilde{z}S'/\rho = 2i\tilde{z}$ is nonsingular, justifying this step.
momentum is physical (for example, it can be extracted by a charged particle) and amounts to \( \pm \hbar \) per photon in the quantum theory of light. A linearly polarized plane wave, in contrast, does not exert any such rotatory forces.

Perhaps surprisingly, there are linearly polarized paraxial beams that can carry arbitrarily high angular momentum per photon! Indeed, \( 30 \hbar \) is readily achievable, for example via holographic construction.\(^7\) The angular momentum carried by an optical vortex beam has technological uses, in part because it can be large enough to be useful for micromanipulation (an “optical torque wrench”).

You’ll work out details of such optical vortex beams in Problems 41.3–41.4. Specifically, the solution you’ll find is called a Laguerre–Gaussian beam with “winding number” \( \pm 1 \).

### 41.4 Bessel Beams

#### 41.4.1 An idealized solution with no diffractive spreading at all

All of the beams studied in the preceding sections exhibited diffraction: They spread, just as in acoustics. Can we imagine a beam that doesn’t spread? Such a beam would have a vector potential whose amplitude is translationally invariant in one direction, for example,

\[
\vec{A}(t, \vec{r}) = \frac{1}{2} e^{-i k_z t + i \vec{k}_x \vec{r}} u(x, y) \zeta + \text{c.c.} \tag{41.11}
\]

Compared to Equation 41.3, this trial solution forbids any \( z \) dependence in \( u \) but relaxes the previous assumption that \( \vec{k}_z = k_z \). Also, this time we seek an exact solution (without paraxial approximation).

The Helmholtz equation now reduces to

\[
k_s^2 u + \rho^{-1} \frac{\partial}{\partial \rho} \left( \rho \frac{\partial u}{\partial \rho} \right) + \rho^{-2} \frac{\partial^2 u}{\partial \varphi^2} - \beta^2 u = 0. \tag{41.12}
\]

As with our study of gaussian beams, we’ll begin with the axially symmetric case, where \( u = u(\rho) \) is independent of \( \varphi \). Then we can solve

\[
\rho^2(k_s^2 - \beta^2)u + \rho \frac{du}{d\rho} + \rho^2 \frac{d^2 u}{d\rho^2} = 0. \tag{41.13}
\]

The solution of Equation 41.13 takes a standard form when we introduce the dimensionless variable \( \nu = \rho \sqrt{k_s^2 - \beta^2} \); then \( u(x, y) = J_0(\nu) \), where the Bessel function \( J_0 \) is the solution to

\[
u^2 \frac{d^2 J_0}{d\nu^2} + \nu \frac{dJ_0}{d\nu} + \nu^2 J_0 = 0. \tag{Bessel equation, order 0} \tag{41.14}
\]

#### 41.4.2 A physically realizable approximation to the ideal

The Bessel beam solution that we just found has some remarkable features:

\[^7\text{Media 16 illustrates how such high transport can move large (micrometer) objects rapidly, overcoming the high viscous friction in the microworld.}\]
Chapter 41 Beams: Gaussian, Vortex, and Bessel

- It is an exact solution to the Maxwell equations (no paraxial approximation).
- It is translation invariant along its direction of propagation, like a plane wave.
- *Unlike* a plane wave, it concentrates its energy along its central axis.

In this sense, it may be called a “nondiffracting beam.” But that sounds paradoxical—spreading seems inevitable, especially in the light of the mathematical analogy to quantum mechanics noted earlier.

The Bessel beam avoids contradiction because it is not really confined in the transverse direction: Although its amplitude decreases with increasing distance from the central axis, it does so slowly, as $e^{-r^2/2}$. Indeed, the fraction of energy flux along $\hat{z}$ that passes through a transverse disk with any finite radius is zero, because there is always an infinite total flux outside that radius. So a true Bessel beam is just as unattainable in practice as a true plane wave. We can now ask, is there any *approximately* Bessel beam that is experimentally realizable, yet still has less transverse spreading than a gaussian beam?

J. Durnin and coauthors found a simple answer. The setup involves a plane wave that impinges on an opaque barrier with a narrow, annular (ring-shaped) aperture. We expect that any such confinement of light will lead to a complicated diffraction pattern downstream, whose intensity profile evolves as we look at various downstream positions $z$. The beam then encounters a thin focusing lens, placed at a distance $f$ from the aperture. The value $f$ is chosen to equal a parameter of the lens called its focal length, which will be defined in a moment. That’s the complete setup, easily arranged in a lab.

To find the downstream pattern of illumination, we first simplify the mathematical problem by specializing to paraxial approximation (Section 41.2.1). Thus, we are only interested in the intensity close to the central axis, and we ask how the electromagnetic field changes as we move downstream from the aperture at $z = 0$. Because of the mathematical analogy between the paraxial equation and the Schrödinger equation, we can think of our task as finding a “time” evolution operator for a free particle in 2D. Following Figure 41.1, let $\vec{r}_b$, $\vec{r}_l$, and $\vec{r}_s$ denote 2D vectors at the barrier, lens, and screen respectively.

Figure 41.1: [Cartoon; not to scale.] Cross-section of Durnin and coauthors’ setup. The vectors $\vec{r}_b$, $\vec{r}_l$, and $\vec{r}_s$ denote transverse distances from the centerline in the planes of the barrier, lens, and screen, respectively, and $f$ is the lens’s focal length. The barrier on the left has a narrow, ring-shaped opening of radius $a$. 
Your Turn 41D

a. Show that
\[ u(\hat{r}_1, z) = -\frac{ik_x}{2\pi z} \int d^2\hat{r}_1' u_0(\hat{r}_1') e^{(ik_x/(2z))||\hat{r}_1 - \hat{r}_1'||^2} \]  
(41.15)

is a Green function solution to the paraxial equation.

b. Show that as \( z \to 0^+ \), Equation 41.15 reproduces the starting profile \( u_0 \).

[Hint: Review Section 22.3 (page 342).]

To find the diffraction pattern from an aperture, we make the idealization that the vector potential at \( z = 0 \) is \( u_0 = 0 \) everywhere except within the opening, where it is a constant (due to plane-wave illumination).\(^8\) Then we carry out the integral in Equation 41.15. However, Durnin's setup contained another element: The lens at \( z = f \).

We can model a thin lens as simply introducing a delay as light passes through, or more precisely, an extra phase factor\(^9\) that depends on transverse position \( \hat{r}_L \) in the plane \( z = f \). The delay is maximal at the thickest part of the lens, that is, at \( \hat{r}_L = 0 \), and decreases as we move outward. In paraxial approximation we only need small values of \( ||\hat{r}_L|| \), so we can write a Taylor series:

\[ \text{delay} = \text{const} - \frac{1}{2cD}||\hat{r}_L||^2 + \cdots \]  
(41.16)

The constant in the second term must have dimensions \( L^{-2} \), so we wrote it as \( \frac{1}{D} \) divided by a length scale \( L \). The single quantity \( \text{const} \) contains all we need to know about the lens's curvature and refractive index; Section 23.3.1 (page 375) called it the focal length.

We now use Equation 41.15 twice:\(^10\) once to get from \( z = 0 \) to \( z = f \), then again to get beyond the lens to \( z = 2f \). The first step yields the field just prior to the lens:

\[ u(\hat{r}_L, f) = -\frac{ik_x}{2\pi f} e^{(ik_x/(2f))||\hat{r}_L||^2} \int \text{slit} d^2\hat{r}_B u_0(\hat{r}_B) e^{(ik_x/(2f))||\hat{r}_B||^2} e^{-ik_x/f}\hat{r}_L\cdot\hat{r}_B, \]  
(41.17)

According to Equation 41.16, the lens introduces a constant times \( \exp(-ik_x c(2cD)^{-1} ||\hat{r}_L||^2) \), which cancels the first exponential in Equation 41.17.

Next, we introduce another Green function to find the field at \( z = 2f \):

\[ u(\hat{r}_B, 2f) = \left( -\frac{ik_x}{2\pi f} \right)^2 e^{(ik_x/(2f))||\hat{r}_B||^2} \int \text{lens} d^2\hat{r}_L \left[ e^{(ik_x/(2f))||\hat{r}_L||^2} e^{-ik_x/f}\hat{r}_B \cdot \hat{r}_L \right. 
\times \left. \int \text{slit} d^2\hat{r}_B u_0(\hat{r}_B) e^{(ik_x/(2f))||\hat{r}_B||^2} e^{-ik_x/f}\hat{r}_L\cdot\hat{r}_B \right]. \]

---

8See Section 22.4.4 (page 351).
9See Section 23.3.1, especially Equation 23.21 (page 376).
10The logic is similar to Section 22.4.3 (page 347).
Now rearrange:
\[ = (\text{const})e^{i\vec{k}_s \cdot \vec{r}} \int_{\text{slit}} d^2 \vec{r}_b \left[ u_0(\vec{r}_b) e^{i\vec{k}_s \cdot \vec{r}} \right]^2 \]
\[ \times \int_{\text{lens}} d^2 \vec{r}_l e^{-(\vec{r}_s^2 + \vec{r}_b^2)} e^{i\vec{k}_s \cdot \vec{r}} \right]^2 \]. \quad (41.18)

If the integral over \( \vec{r}_s \) were unrestricted, then combining the exponentials and completing the square shows that the inner integral would be of Fresnel form, yielding a constant times \( e^{-(\vec{r}_s^2 + \vec{r}_b^2)} \). A real lens is finite in extent, but we will suppose that its diameter is much larger than the aperture. So the complex amplitude at the screen is
\[ Q \int_{\text{slit}} d^2 \vec{r}_b u_0(\vec{r}_b) e^{-(\vec{r}_s \cdot \vec{r})} \]
where \( Q \) is a normalization factor.

Within the approximations we made (paraxial; thin lens; large lens diameter), we conclude that the beam profile is the Fourier transform of the aperture function. For a narrow circular slit, \( u_0 \) is a constant over a narrow ring \( \| \vec{r}_b \| = a \) and zero elsewhere, so
\[ u(\vec{r}_s, 2f) = (\text{const}) a \int_0^{2\pi} d\varphi' e^{-(\vec{r}_s \cdot \vec{r})} \cos \varphi' \quad \text{where} \quad \rho = \| \vec{r}_s \|. \]

The remaining \( \varphi' \) integral is another representation of the Bessel function, so
\[ u(\vec{r}_s, 2f) = (\text{const}) J_0((k_s a / f) \rho). \quad (41.19) \]

The time-averaged intensity is then proportional to \( |u|^2 \).

Thus, the setup in Figure 41.1 should indeed create an approximately Bessel beam. Figure 41.2 shows a demonstration. Even at distances beyond \( z = 2f \), Section 41.4.1 suggests that we may expect the intensity profile to be nearly unchanging as the beam propagates. Figure 41.3 shows that indeed the central intensity maximum does not widen appreciably over many times the focal length. For comparison, a gaussian beam at the same wavelength and spot size has a Rayleigh range of only about \( 0.1 \) mm. Passing a plane wave through a 10 \( \mu \)m pinhole indeed yields a beam that spreads rapidly over just a few centimeters.

41.4.3 Application to microscopy

Bessel beams, well-investigated nondiffracting beams, have been shown to possess greater resistance to scattering media and to have a greater penetration depth than gaussian beams (63–66). However, these nondiffracting profiles come with bright side lobes, which are a corollary of the self-healing property but can greatly deteriorate the optical sectioning performance of a light sheet at the same time. By combining a Bessel
41.4 Bessel Beams

Figure 41.2: [Experimental data.] **Almost-Bessel beam**, produced via the setup described in the text. A laser beam with wavelength 532 nm impinged on an annular aperture: a ring with diameter \(2a = 3.9\) mm and width 25 \(\mu\)m. A lens with diameter 50 mm and focal length \(f = 60\) mm was positioned at \(z = f\) and viewed at \(z = 2f\). The graph shows the intensity of the light along a line through the center; each peak appears as a ring on the focal plane, with spacing for the innermost rings in rough agreement with our prediction (the square of Equation 41.19). **Right:** intensity sampled along a radial line through the center. [Data courtesy Lucas Hanson.]

Figure 41.3: [Experimental data.] **Nearly diffractionless propagation of a beam.** The apparatus is the same as in Figure 41.2, but this time the light intensity was viewed at up to twelve focal lengths away from the lens. **Left:** Detail showing that the width of the central intensity maximum does not change appreciably. Data have been normalized so that the light collected over the entire image is the same for each curve. **Right:** Corresponding full 2D patterns of illumination. [Data courtesy Lucas Hanson.]

beam with structured illumination or two-photon excitation techniques (67, 68), the side lobes of the Bessel beam can be sufficiently suppressed, and an isotropic resolution can be achieved while maintaining a large FOV (up to 60 um). In the case of an Airy beam, image artifacts introduced by the side lobe can be effectively reduced by deconvolution (69, 70), and a substantially extended FOV (up to 160 um) as well as improved resolution
over Bessel-beam and gaussian-beam illumination have been demonstrated (70).” – Liu et al., 2020


FURTHER READING

Semipopular:
Simon, 2016.
Optical vortices as mechanical actuators: Media 16.

Intermediate:
Peatross & Ware, 2015, chap. 10–11.
Vortex and Bessel beams: Simon, 2016; Milonni & Eberly, 2010, chap. 7; Jones et al., 2015, chap. 4; Andrews & Bradshaw, 2022, chap. 8; Zangwill, 2013, chap. 16; Smith, 1997, chap. 3; Freeman et al., 2019, chap. 12.
Demonstration of Bessel beams: Basano & Ottonello, 2005; McQueen et al., 1999.

Technical:
Ruffner & Grier, 2012; Curtis & Grier, 2003.
Roichman et al., 2008.
Lee et al., 2010.
Simon, 2016; Allen 1992; Allen et al., 1999; Loudon, 2003;
http://en.wikipedia.org/wiki/gaussian_beam; Pampaloni and Enderlein;


https://doi.org/10.1364/OPEX.12.005448.


Ji et al., 2020.
41.1 Display wavefronts
Continuing Your Turn 41C (page 622), use a computer to display some wavefronts for $-0.2 \leq \hat{z} \leq 0.2$. [Hints: By the axial symmetry of your solution to Your Turn 41C, it’s enough to show just the intersections of the wavefronts with the $\hat{x}\hat{z}$ plane, that is, to show curves in that plane. Limit your graph to a relevant range of $\hat{x}$ values (a range where the amplitude is nonnegligible). To make the effect easier to see, set the constant $(k_s \omega)^2$ unrealistically small, say 30.]

41.2 The razor’s edge
Chapter 22 discussed the illumination pattern when an infinite plane wave impinges on a half-plane obstruction. The case of a narrow beam is also tractable in paraxial approximation. Consider a beam with $k_s = 2\pi/(543 \text{ nm})$ whose profile at $z = 0$ is a half gaussian:

$$u(r_L) \propto \begin{cases} \text{e}^{-\left(\frac{x^2+y^2}{(2\omega^2)}\right)} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

Thus, there is no spatial confinement in $y$. Suppose that $\omega = 2 \text{ mm}$ and we observe intensity on a screen at $z = 9 \text{ m}$.

Both the profile function and the Green function from Your Turn 41D (page 625) factorize into products of $x$ and $y$ factors. So the illumination intensity at the screen is the product of a function of $x$ times a function of $y$.

a. Express each of those functions in terms of a one-dimensional integral.

b. Then evaluate the integrals, for example, numerically, over the range $\pm 8 \text{ mm}$ and compare to the behavior in Figure 22.1 (page 341).

c. For more insight, choose some interesting values of $x$ and $y$ and show the integral as an arrow plot (similar style to Figure 22.4). Unlike the situation in that figure, however, your arrows will not all be of equal length.

41.3 Whirl
In this problem, you’ll find another beamlike solution to the Lorenz-gauge Maxwell equations, with more interesting structure than the ones in Section 41.2.

a. Extend the discussion in Section 41.2 to find some more solutions of the form Equa-

Figure 41.4: See Problem 41.2. A gaussian laser beam with diameter about $2 \text{ mm}$ and wavelength $543 \text{ nm}$ impinged perpendicularly on a vertical, conducting straight edge that interrupted the left half of the beam. The resulting illumination was viewed on a screen at distance $9 \text{ m}$ from the obstruction. [Photo courtesy William Berner.]
tion 41.3. Again try \[ \hat{\zeta} = \zeta \hat{x}, \]
but now let your solution depend on azimuthal angle \( \varphi \).

The circular symmetry of the paraxial equation suggests that we look for solutions that are proportional to \( e^{im\varphi} \). But any such solution will have infinite derivatives on the symmetry axis (\( \phi = 0 \) in cylindrical coordinates), unless it also vanishes at least as fast as \( \rho^{m} \). So replace Equation 41.7 by the trial solution

\[
\tilde{u}_{m} = e^{im\varphi} f_{m}(\tilde{z}, \tilde{\rho}), \quad \text{where} \quad f_{m} = \tilde{\rho}e^{-\tilde{\rho}^{2} + 2S_{m}(\tilde{\rho}) + \cdots}, \quad \text{and} \quad m = +1. \tag{41.20}
\]

Here \( \tilde{\rho} = \rho / \omega \) is the dimensionless radial coordinate defined in the main text. Find the unknown functions \( S_{m}(\tilde{\rho}) \) by using the same approach and approximations as in Section 41.2, and comment.

b. Repeat Your Turn 41C and Problem 41.1 for your solution in (a), again with the value \((k_{x}, \omega)^{2} = 30\). This time, you’ll need to make a 3D plot of a surface in \( \bar{x} \bar{y} \bar{z} \) space. Only display one wavefront. Describe in words how it evolves over time. [Optional: Make an animation to support your words.]

c. Construct a circularly-polarized beam solution of this type. [Hint: Try a linear combination of solutions you know.]

d. Optional: Carry on with higher values of \( m \).

41.4 Twirl

In this problem, you’ll continue to study solutions to the Maxwell equations of the generic form Equation 41.20:

\[
\vec{A}(t, \vec{r}) = \frac{1}{2} \zeta \frac{e^{i(k_{x}x + k_{y}y)}}{\sqrt{x^{2} + y^{2} / \omega}}, \quad \text{where} \quad \vec{z} = \frac{z}{2k_{z}\omega^{2}}, \quad \zeta \quad \text{is a constant vector, and} \quad m \quad \text{is an integer.}
\]

Section 41.2 found an approximate solution of this form with \( m = 0 \) and \( f_{0} = u = \exp(-\tilde{\rho}^{2} + i\tilde{z}(1 - \tilde{\rho}^{2})) \). In Problem 41.3, you studied the cases \( m = \pm 1 \). In this problem, you’ll see a possibly surprising feature of the \( \pm 1 \) solutions. You won’t need the detailed form of the functions \( f_{m} \); what matters is that, as you have shown, approximate solutions of the above form do exist.

For concreteness, suppose that the solution is linearly polarized: \( \vec{\zeta} = \zeta \hat{x} \). We would like to see whether the beams of light given by Equation 41.21 with \( m \neq 0 \) can make particles twist around in the \( xy \) plane, despite not being circularly polarized. To do this, we need the flux of momentum (force per area) crossing each point of the transverse plane \( \{z = 0\} \), as a function of \( \tilde{x} \) and \( \tilde{y} \) (or cylindrical coordinates \( \rho \) and \( \varphi \)). We are especially interested in the azimuthal component of this momentum flux, that is, in \( \vec{\phi} \cdot \hat{T}_{\text{field}} \cdot \hat{\tilde{z}} \).

Although this is a component in cylindrical coordinates, it will be convenient to first work out the cartesian components, then convert at the end. It may be useful to recall that

\[
\frac{\partial \rho}{\partial x} = x / \rho = \cos \varphi; \quad \frac{\partial \varphi}{\partial x} = -\rho^{-1} \sin \varphi; \quad \text{and so on.}
\]

a. Starting from Equation 35.11 (page 560), derive a formula for this quantity in terms of components of \( \vec{E} \) and \( \vec{B} \).
Before proceeding, note that all of our solutions have $\frac{\partial f}{\partial z} \propto 1/(k_s w^2)$, but such quantities are much smaller than $\partial e^{ik_r z} / \partial z \propto k_s$. This means that when we calculate fields on the plane $z = 0$, the factors $\exp(2S_m)$ will be negligible and may be dropped. For example, in this problem you may take $f_{\pm 1}$ to have the simple form $\frac{\rho}{w} e^{-\left(\frac{\rho}{w}\right)^2}$, which is real and independent of $z$.

b. Use the Lorenz gauge condition to find the scalar potential\(^{13}\) $\psi$ corresponding to Equation 41.21 in terms of $k_s$, $m$, the function $f_m$, and its $\rho$ derivatives.

c. Find expressions for the four components of $\vec{E}$ and $\vec{B}$ fields found in (a), and set $z = 0$. Are the fields perpendicular to the $z$ axis?

d. Substitute your results from (c) into your formula from (a), then simplify by finding the time average.

e. The maximum torque along $\hat{z}$ that this beam could exert (if fully absorbed) is the integral over the $xy$ plane of the angular momentum flux $\rho \hat{\phi} \cdot \vec{T}_{\text{field}} \cdot \hat{z}$. So evaluate this in terms of $k_s$, $w$, $m$, and the overall amplitude factor $\zeta$ from your result in (d). Comment on the distinction between the cases $m = -1, 0, \text{ and } 1$.

41.5 **Twirl (circular polarization)**

Work Problem 41.4, but with the following modifications: The beam is gaussian $(m = 0)$, but circularly polarized.

---

\(^{13}\)Recall our treatment of the spherical wave solution (Your Turn 38A).
CHAPTER 42

Vista: J. J. Thomson’s Pictorial Explanation of Radiation

42.1 FRAMING: KINKS

The Gauss law implies that a static positive point charge creates an electric field that is directed radially outward and falls as \( r^{-2} \). That behavior is quite different from radiation. For example, the energy density of the static charge falls as \( r^{-4} \), too fast to transport any energy to infinity.\(^1\) But there are many other solutions to the Maxwell equations. In particular, when a charge is in motion, then it’s no longer a spherically symmetric source, so we need not expect a spherically symmetric solution. Indeed, Sections 33.4.2 and 34.8.2 found bunching of the field into the equatorial plane.

This chapter will extend the discussion to accelerating charges\(^2\) by abstracting just one more qualitative fact from Chapter 39: Disturbances in the field propagate at the fixed, finite speed \( c \). Starting from that observation, and Michael Faraday’s field-line concept, J. J. Thomson built a pictorial explanation that gives most of the qualitative features of the electric field arising in radiation. Adding Faraday’s law of induction will then allow us understand the magnetic field as well. Later chapters will work through the analytic details, but it’s good to have this intuition first.

Electromagnetic phenomenon: The pulse of radiation from a suddenly accelerated charge consists of fields that are transversely polarized, have maximal strength in the plane perpendicular to the acceleration, and fall with distance as \( 1/r \).

Physical idea: Field lines may not terminate in empty space; to connect regions before and after an impulse, they must kink.

42.2 ELECTRIC FIELDS FROM A SUDDENLY ACCELERATED CHARGE

Figure 42.1b shows the trajectory in spacetime of a particle that is moving at uniform velocity along \( \dot{x} \) from time \( -\infty \) till time zero, then decelerates to rest at a point \( P \). At some later time \( t_f \), we ask what the fields look like throughout space.

- A nearby observer, at \( O_2 \), sees \( \vec{E} \) pointing radially outward from the charge at \( P \), because by \( t_f \), enough time has elapsed for this observer to learn that the charge has come to rest.
- A distant observer, at \( O_1 \), sees \( \vec{E} \) pointing radially outward from a different point \( R \), located where the charge would have been at time \( t_f \), had it not decelerated (Sec-

---

\(^1\)See Section 25.5.3 (page 397).
\(^2\)You began this program in Problem 34.3 (page 552).
Electric field lines from a charge that decelerates. The field lines were drawn by numerically evaluating the Liénard–Weichert formula (Equation 39.13, page 598) and finding the streamlines of $\vec{E}$; the text describes why the resulting figure had to look like this. Bottom: A charged particle initially moves at constant velocity along the $\hat{x}$ axis at speed $0.8c$, then decelerates to rest at point $P$ during an interval $c\Delta t \approx 0.3\, m$ near time zero. In this cross-section, future-directed light cones from the start and end of the deceleration appear as the dashed lines. Had the particle continued instead of stopping (dotted), it would have arrived at point $R$. Top: Snapshot of the corresponding $\vec{E}$ field lines at one time $ct_i = 2\, m$, in the upper $xy$ half-plane. (The full 3D picture would have more field lines, obtained by rotating the ones shown about the $\hat{x}$ axis.) In this cross-section, the same two light cones appear as the dashed circles. At higher initial velocity, the field lines in the outer region would be more crowded toward the $\hat{y}$ direction (see Problem 42.3).

We now connect up the two regions just described. We know that $\vec{\nabla} \cdot \vec{E} = 0$ in vacuum, so field lines cannot terminate anywhere except on the charge itself. In the intermediate region, the field lines must join up as drawn in the figure, in order that every line continues out to infinity. We can see that:

3. This conclusion has nothing to do with the mental state of the observer. At this point in space and moment in time, no instrument could distinguish the trajectory from one that is in eternal, uniform straight-line motion.

4. See Section 36.2.2 (page 571).
An observer at \( \mathbf{O}_s \), for example, sees a pulse of \( \vec{E} \) directed transversely to their line of sight away from \( \mathbf{P} \). These kinks lie on a spherical shell whose radius expands outward in time at speed \( c \).

The longest kinks are those located at right angles to the direction of acceleration (that is, along the \( \pm y \) axis); there is no kink for an observer on the \( \pm x \) axis (the direction of acceleration).

We can summarize both of these observations by saying that the kink is roughly directed along \( \hat{r} \times (\hat{r} \times \vec{a}) \), where \( \vec{a} \) is the acceleration and \( \hat{r} \) is directed toward the observer.

Now that we understand the direction of the field in the kink region (transverse to the line of sight), we can ask about its strength. Chapter 36 showed that \( ||\vec{E}|| \) is proportional to the transverse areal density of the field lines, which in turn is the total length of all the lines in a volume, divided by that volume. The stretching needed to accommodate the kinks without breaking any line crowds more line length into the kink region than...
there would be without any acceleration. Hence, the field is stronger in that region (the spherical shell in Figure 42.1a) than outside it. Let us see how much stronger.

Consider an observation made at $O$, a particular angle $\theta$ from the $\hat{x}$ axis. The charge decelerates from velocity $v$ to 0, so the kink joins a line pointing away from $P$ to one pointing away from $R$, which sits a distance $vt_f$ to the right of $P$ (Figure 42.1), or $vt_f \sin \theta$ in the direction transverse to the field line. The acceleration occurs over a time interval $\Delta t \approx v/a$, so the thickness of the shell between the dashed circles is $\approx cv/a$.

Imagine drawing $N$ lines emerging from the charge. We wish to find the total length of a bundle of field lines that originally occupied area $d\Sigma$ before passing through a spherical shell of radius $R$. A total of $N d\Sigma/(4\pi R^2)$ lines enter, bend sideways, and travel a distance $vt_f \sin \theta$ through the shell. Thus,

$$\text{total length of lines} \approx \left( \frac{N d\Sigma}{4\pi R^2} \right) (vt_f \sin \theta) \approx \frac{N t_f a \sin \theta}{4\pi R^2 c}.$$

The radius $R$ of the sphere is $R = ct_f$, so we find that $||\vec{E}||$ is proportional to the acceleration, to $\sin \theta$, and to $1/R$. Thus, we have found the hallmarks of radiation from an accelerated charge:

- The electric field is transverse to the line of sight from observer to source.
- The electric field is strongest in the equatorial plane $\theta = \pi/2$.
- The electric field falls with distance as $R^{-1}$, not $R^{-2}$.
- The electric field is proportional to the magnitude of the acceleration.

Figure 42.2 gives another representation of these fields.

### 42.3 MAGNETIC FIELDS ALSO HAVE A RADIATION CONTRIBUTION

Next, let’s find the $\vec{B}$ field lines by using the Faraday law.

Consider the same charged particle trajectory as before. Figure 42.3 shows the electric field lines at a time $t_1$. The preceding section argued that at a fixed point $S$ in the $xy$ plane, an observer will initially see a small, radial electric field ($\propto r^{-2}$), because they are outside the expanding shell and so has not yet learned about the deceleration (Figure 33.1b), then...
around $t_1$ a pulse of $E_x$ ($\propto r^{-1}$), and then back to small field at later time. Point $V$ is inside the shell, and so sees the magnetic field of a charge at rest, which is zero. But as the leading (outer) boundary of the shell sweeps past (point $T$), the observer sees $E_x$ increasing over time.

Figure 42.3 also shows a small rectangular surface element surrounding $T$, coming out of the page in the $yz$ plane. The top and bottom edges of this rectangle straddle the boundary of the shell. Integrating Faraday’s law over this surface element shows that $\vec{B}_z$ must be nonzero on the top or bottom edge. But the top edge sees only a static electric field, so $\vec{B}_z = 0$ there.

The field lines must form a figure of revolution about the $x$ axis, by axial symmetry, and they may not terminate at the origin (nor anywhere else), so consider lines that are circular loops in planes parallel to the $yz$ plane. Thus, just below $T$ in the figure we have $\vec{B}$ pointing out of the page, with field lines forming rings in planes parallel to the $yz$ plane where $E$ is changing.

A similar argument applies at point $U$ on the trailing boundary of the shell. Here $E_x$ is falling over time, but $\vec{B}$ can only be large on the top edge of the rectangle. So again we find $\vec{B}$ pointing out of the page. Also, $||\vec{B}|| \propto R^{-1}$, because we previously found that $\vec{E}$ has that sort of falloff. All told, we have that:

- In the pulse region, $\vec{B}$ is directed azimuthally and falls as $R^{-1}$, so
- In that region, $\vec{E} \times \vec{B}$ is directed radially outward and falls as $R^{-2}$,

which are two more hallmarks of radiation.

Because we found that the energy flux falls with distance as $R^{-2}$, the total power sent through a distant sphere of radius $R$ is $4\pi R^2 \times (\text{flux})$, which approaches a constant at large $R$. This is the sense in which radiation “sends energy all the way out to infinity.”

FURTHER READING

*Intermediate:*

Historical: Thomson, 1904.

Purcell & Morin, 2012; Smith, 1997.
42.1  **Standing start**

Problem 34.3 asked you to find the fields generated by a charge that suddenly decelerates, but the result was incomplete: There was a discontinuity in the potential, arising from the unrealistic assumption of instantaneous deceleration. Figure 42.2 (page 635) shows the result of a better calculation. In this problem, you’ll do something similar for a different but related situation.

a. Make a sketch like Figure 42.1 (page 634), but this time, show the expected field lines from a charge that starts at rest, then accelerates to velocity \(-0.8c\hat{x}\) over \(c\Delta t \approx 0.3\ m\) and is observed at time \(= 2m/c\).

b. A convenient model trajectory is

\[
x(t) = -\frac{1}{2}(0.8)(0.1\ m)(\log(\cosh(\frac{-ct}{(0.1\ m)})) + (ct/(0.1\ m))) + \log 2.
\]  

(42.1)

Graph this function and comment on why it has the desired features.

c. Now get a computer to make a precise version of your figure by evaluating the Liénard–Weichert potentials and finding appropriate numerical derivatives. As in Problem 42.2b,d, for each desired point in the \(xy\) plane at the observation time you will first need to find the retarded time by numerically solving an algebraic equation.

42.2  **Bumper car**

A point charge \(q\) sits motionless at the origin \(\vec{r} = 0\) for a long time, then gets bumped, causing it to move along the \(x\) axis. It then returns to its original position and sits there forever. All told, its trajectory in the lab system is specified by the function

\[
\vec{r}(t) = \begin{bmatrix} x_0e^{t^2/(2\tau^2)} \\ 0 \\ 0 \end{bmatrix}.
\]

Here \(x_0\) and \(\tau\) are constants. You can measure all lengths and \(ct\) values in meters or any other arbitrary unit. For concreteness, choose the twitch duration parameter to be \(c\tau = 0.2\) times that unit. Also choose \(x_0\) such that the maximum velocity achieved during the twitch is \(0.4c\).

A second point charge \(-q\) sits forever at the origin without moving. Thus, at early and late times there is no net charge nor current anywhere; near time zero, there is a transient charge separation.

a. What is the value of \(x_0\)?  [**Hint:** It must have appropriate units, be constructed from the given parameter values, and fulfill a condition stated above.]

b. Find a formula for the Lorenz-gauge 4-vector potential set up by the charges. Your formula should be exact (no multipole nor far-field approximation), but it will be implicit. That is, it involves the retarded time, which is determined by an algebraic (not differential) equation.

c. Before any detailed calculation, explain why for a field point (observation point) \(\vec{r}\) in the \(xy\) plane (that is, \(\{z = 0\}\)), the \(z\)-component of the resulting electric field will be zero.
d. Consider an interesting range of time values from something less than zero to something greater than zero. For each of several time values in that range, set up a grid of points that covers an interesting region of the $xy$ plane. [Hint: Dimensional analysis may help you to make an initial guess for what is an “interesting” range.] The grid should be fine enough to get reasonably accurate estimates of derivatives by numerical differentiation.

e. Then use a computer to evaluate your result from (b) numerically at each grid point, at each of the time values that you chose.

f. Use your result from (e) to evaluate $\vec{E}$ in the $xy$ plane at each time value. (Why is it good enough just to show the $xy$ cross-section?)

g. Make a graphical depiction of the magnitude $||\vec{E}(t, x, y, 0)||$ at each chosen time. For example, you may wish to make a heat map; that is, show this scalar quantity as color on a plane. Or you may prefer a contour plot or surface plot. Use your judgment about what is clearest.

h. Identify as many visual features of the result as you can, and explain why they arose. This important step also helps you check your results for reasonableness. For example, at any time there will be some places on the $xy$ plane that have “not yet learned” about the motion of $q$, and others that have “already forgotten about it.” What are those regions? What should be the field there? Do your graphics show that behavior?

i. Optional: A picture may be worth a thousand words, but a movie is worth many pictures, so get your computer to make an animated graphic showing the time development.

j. Optional: Create some other meaningful representation, using your own judgment, that shows something else interesting about this system, or about an interesting related system and has features that confirm general conclusions.

Notes:

- Python users may find `numpy.meshgrid`, and its builtin help description, to be useful. If you use it, make sure you understand the two options indexing="xy" versus indexing="ij" and choose the one you want. (To see the distinction, try it out with small arrays.)
- If you use `numpy.gradient`, check its documentation and experiment on a small array to make sure you know exactly how it works. Or just do your own subtraction to estimate a gradient.
- Make sure your computer uses the same scale for the $x$ and $y$ axes.
- If the range of values attained is too large to display properly, compress it before making the plot. For example, you could use a monotonic function like $n$-th root, or logarithm, for this. Alternatively, in Python the keyword arguments `vmax` and `vmin` for `plt.imshow` may be useful.

42.3 Relativistic bremsstrahlung

a. Make a sketch like Figure 42.1 (page 634), but this time show the expected $\vec{E}$ field lines from a charge that starts out moving at $0.95c$, then comes to rest over $c\Delta t \approx 0.3$ m and is
observed at time $= 2m/c$. What do you expect to be different in this highly-relativistic situation, compared Figure 42.1?

b. Now get a computer to make a precise figure. As in Problem 42.2, for each desired point in the $xy$ plane at the observation time you will need to find the retarded time by numerically solving an algebraic equation.
CHAPTER 43

Electric Dipole Radiation

43.1 FRAMING: *DOUBLE EXPANSION*

This chapter will show in a special situation that, as foreshadowed in Chapters 25 and 42,

- Charges emit electromagnetic radiation when accelerated,
- In the far-field region, the radiation is polarized transversely to the line of sight, and
- Far from the source, its energy flux falls with distance like $1/r^2$.

The special situation, which is frequently relevant in practice, is a limit in which the source is distributed over a region whose size is much smaller than the outgoing wavelength. Thus, it makes sense to attempt a *double expansion* in both size/wavelength and size/(distance to observer).

Unlike Chapter 25, this time, we make no restriction that charge density is everywhere zero. Remarkably, once again a multipole expansion will help us out.

*Electromagnetic phenomenon:* Diatomic, homonuclear molecules have little infrared activity, but more complex ones can be strongly IR-active gases.

*Physical idea:* $\text{O}_2$ and $\text{N}_2$ have no electric dipole moment, even when set into vibration.

43.2 THREE LENGTH SCALES

Suppose that some charges executing prescribed motions are confined to a region of size $\approx a$ centered on the origin of coordinates. So their positions $\vec{r}_n$ all satisfy $||\vec{r}_n|| < a$. We observe fields at a field point $\vec{r}$. In statics problems (Chapters 3 and 17), we found a great simplification if we are only interested in the far fields, that is, the case $r \gg a$. The situation is a bit more subtle in dynamics, however, because a third length scale enters, allowing a richer set of asymptotic situations.

To understand the new length scale, in this chapter we suppose that charge density and flux vary periodically in time with some frequency $\omega$. Then the quantity $c/\omega$ has dimensions of length, and indeed, it will be the wavelength of emitted radiation (divided by $2\pi$). One common situation is then

$$r \gg c/\omega \gg a. \quad \text{far field (dynamic)}$$ (43.1)

For example, an atom ($a \approx 0.1 \text{ nm}$) may emit light ($c/\omega \approx 100 \text{ nm}$), which may be observed at a distance $r > 1 \text{ cm}$. However, it is also useful to break this strong condition down into components, to see which conclusions depend on which conditions. Accordingly, the
Figure 43.1: **Three limiting cases.** The small source regime is \( r \gg a \) (**hatched region** extending to right). The multipole limit is \( c/\omega \gg a \) (**hatched region** extending upward). The far field regime is the intersection of these, with the additional condition \( r \gg c/\omega \) (**solid triangle** extending up and to the right).

following sections will introduce

**r \gg a**  \[ r \gg a \text{  small source regime} \]  \hfill (43.2)

and

**c/\omega \gg a.**  \[ c/\omega \gg a. \text{  multipole regime} \]  \hfill (43.3)

Note that even when both Equations 43.2 and 43.3 hold, we may nevertheless not be in the far field, because \( r \) may not be \( \gg c/\omega \). For example, a small microwave dipole antenna generates fields that we may choose to measure in a zone that is not many wavelengths away. Figure 43.1 shows the three regimes just defined and their relations.

43.2.1 Small source

Let’s first recall familiar steps from statics: Let \( \vec{R} = \vec{r} - \vec{r}_s \). Please review why

\[ R = r - \hat{r} \cdot \vec{r}_s + \cdots \]  \hfill (43.4)

\[ R^{-1} = r^{-1} \left( 1 + \frac{\vec{r}_s}{r} + \cdots \right). \]  \hfill (43.5)

In each case, we have kept the first two orders of a power series in \( a/r \); the ellipses denote terms of higher order, which are subleading in the small source regime (Equation 43.2).

---

1Section 3.4 (page 40).
Our general, Green-function solution gives the vector potential in Lorenz gauge as

\[ A^\mu(t, \vec{r}) = \frac{\mu_0}{4\pi} \int d^3r_s \frac{1}{R} J^\mu(t - R/c, \vec{r}_s). \]  

[39.6, page 595]

We need to be careful with our approximation. In the 1/R factor, the second and higher terms in Equation 43.5 can be dropped—they make contributions to \( A \) that are suppressed by powers of \( a/r \) relative to the first term. But in the argument of \( J \), we must keep the first subleading term of Equation 43.4 because, although it is smaller than the leading term,

- Its overall magnitude tends to a constant, not zero, as \( r \to \infty \), and
- When we take \( J \) to vary harmonically in the next section, this additive term will turn into a multiplicative factor that cannot be dropped.

Moreover, we’ll see that the apparently leading term does not give rise to any radiation. Thus, dropping the subleading term just mentioned might fool us into thinking radiation is not possible at all!

However, the still-higher terms really may be dropped in small source approximation. Thus,

\[ A^\mu(t, \vec{r}) = \frac{\mu_0}{4\pi r} \int d^3r_s J^\mu(t - r/c + \hat{r} \cdot \vec{r}_s/c, \vec{r}_s). \]  

small source (43.6)

Equation 43.6 is the generalization of Equation 25.9 (page 395) to situations where the net charge density is not everywhere zero.

### 43.2.2 Harmonic time variation

Let’s suppose that the source charges and currents \( J \) vary harmonically in time with some angular frequency \( \omega \). That is, assume

\[ J^\mu(t, \vec{r}_s) = \frac{1}{2} e^{-i\omega t} \vec{J}^\mu(\vec{r}_s) + \text{c.c.}, \]

where \( \vec{J}^\mu \) are four complex functions of position \( \vec{r}_s \) only. Then

\[ A^\mu(t, \vec{r}) = \frac{1}{2} \frac{\mu_0}{4\pi r} e^{-i\omega t} \int d^3r_s e^{-i\omega \hat{r} \cdot \vec{r}_s/c} \vec{J}^\mu(\vec{r}_s) + \text{c.c.} \]  

small source (43.7)

Nothing inside the integral depends on the observer’s distance \( r \). However, the observer’s direction \( \vec{r} \) is still present inside the integral.

### 43.2.3 In many applications, the multipole parameter is small

Equation 43.7 is still a bit complicated, but fortunately another approximation is often justified: Often the quantity \( c/\omega \) is much bigger than the source size \( a \) (the multipole...
regime, Equation 43.3). That is, the dimensionless quantity

\[ \epsilon_{\text{multi}} = \omega a/c \]  

is much smaller than 1.

In that case, we may replace the exponential inside the integral by its Taylor series:

\[ 1 - i \epsilon_{\text{multi}} (r \cdot \vec{r_s}) / a + \cdots \]. Making this approximation, and truncating after a finite number of terms, is called multipole expansion. Keeping only the first term (that is, 1) is called electric dipole approximation, for reasons that will be clear soon.

To summarize, we are making a double power series expansion in both \( r_s \) and \( \epsilon_{\text{multi}} \), and from now on in this chapter will keep only the leading nonzero term, which for a generic source is electric dipole (Chapter 44 will explore the next terms).

### 43.3 ELECTRIC DIPOLE RADIATION

#### 43.3.1 A time-varying ED moment leads to 1/r potentials

Equation 43.7 has become

\[ \hat{A}^\mu(t, \vec{r}) = \frac{1}{2} \frac{\mu_0}{4\pi r} e^{-i\omega(t-r/c)} \int d^3 r_s \hat{J}^\mu(v_s) + \text{c.c.} = \frac{\mu_0}{4\pi r} \int d^3 r_s \hat{J}^\mu(t_c, \vec{r}_s). \]  

(43.9)

In this expression, \( t_c \) is shorthand for \( t - r/c \). This quantity is simpler than the retarded time from Chapter 39,\(^5\) because we dial back the time based on the center of the distribution, not the location of particular charges. In particular, \( t_c \) does not depend on \( \vec{r}_s \).

We can now get an even simpler formula\(^6\) for the spatial components \( \hat{A} \). First, the divergence theorem implies

\[ \int d^3 r_s \nabla \cdot \hat{V}_m(\vec{r}_m \tilde{j})|_{\vec{r}_s} = 0 \]

for each of \( m = 1, 2, 3 \). (Remember that \( \tilde{j} \rightarrow 0 \) outside the finite region where the source is located.) Use the product rule for derivatives to find that

\[ \int d^3 r_s \delta_{lm} \tilde{j}(\vec{r}_s) = - \int d^3 r_s \tilde{r}_s \cdot \nabla \cdot \tilde{j} |_{\vec{r}_s} = + \int d^3 r_s \tilde{r}_s \cdot \frac{\partial}{\partial t} \rho(\vec{r}_s) = \frac{d}{dt} \tilde{D}_{e,m}. \]  

(43.10)

The final step made use of the definition of electric dipole moment \( \tilde{D}_e \).

Overall, Equation 43.10 says that \( \int d^3 r_s \tilde{j}(\vec{r}_s) = \frac{d}{dt} \tilde{D}_{e,m} \). So the three spatial components of Equation 43.9 reduce to

\[ \hat{A}^{[\text{ED}]}(t, \vec{r}) = \frac{\mu_0}{4\pi r} \frac{d\tilde{D}_e}{dt} \bigg|_{t-r/c}. \]  

ED approximation, small source  

(43.11)

Again, note that the derivative is to be evaluated at time \( t_c = t - r/c \), not the retarded time from Chapter 39.

---

\(^5\)Section 39.5.1 (page 596).

\(^6\)The following derivation should be familiar from magnetostatics (Chapter 17). What’s different is that this time, time derivatives are not zero in Equation 43.10.
43.3.2 Pure dipole limit

Chapter 38 pulled a spherical wave solution out of a hat and then showed it was an exact solution. Here, we obtained it as an approximate solution to a real physical problem. We can consider the pure-dipole limit, in which \( a \rightarrow 0 \) holding fixed the amplitude \( \tilde{D}_E \). In this limit, the ED approximation really does become exact, and we recover the form found in Chapter 38.

Your Turn 43A

Show that Equation 43.11 agrees with the exact spherical wave solution that we found previously (Equation 38.1, page 584). Relate the constant \( \xi \) used then to \( \mathcal{D}_E \) used in this chapter.

Your Turn 43B

Evaluate \( A^0 \) using Equation 43.9 and check that it agrees with Your Turn 38Aa (page 585).

43.4 THE ELECTRIC AND MAGNETIC FIELDS FALL SLOWLY WITH DISTANCE

We now need a physical interpretation of our answer, Equation 43.11. One good step would be to find the physical fields \( \tilde{E} \) and \( \tilde{B} \). You’ll explore this in detail in Problem 43.1, but in this section we’ll obtain more compact formulas by looking only at the far-field regime (Equation 43.1 and Figure 43.1).

In this limiting case, when taking derivatives we never need to differentiate the \( 1/r \) factor, because that would give \( 1/r^2 \), which we will see is not leading order. The Chain Rule gives

\[
\nabla \times \tilde{A} = \frac{\mu_0}{4\pi} \hat{r} \times \left( -r^{-2} \frac{d\tilde{D}_E}{dt} - (rc)^{-1} \frac{d^2\tilde{D}_E}{dt^2} \right)
\]

\[
= \frac{\mu_0}{4\pi rc} \hat{r} \times (ir^{-1} \omega c + \omega^2) \frac{1}{2} \tilde{D}_E e^{-i\omega c} + \text{c.c.}
\]

In far-field approximation we may drop the first term in parentheses, obtaining

\[
\tilde{B} \approx \tilde{B}^{[ED]} \quad \text{where} \quad \tilde{B}^{[ED]} = -\frac{\mu_0}{4\pi rc} \hat{r} \times \frac{d^2\tilde{D}_E}{dt^2} \bigg|_{t \to r/c} \quad \text{ED approx., far-field}
\]

(43.12)

We can see that:

- Indeed, the only aspect of the source that matters in this approximation is its time-varying electric dipole moment, which explains our name “electric dipole approximation.”

---

7The following derivation is essentially a solution to Your Turn 38B.
Specifically, the $\vec{B}$ field is proportional to the acceleration of the charge.

- The far field wavecrests are spherical and move radially outward at speed $c$, because $\vec{B}$ depends on observer’s distance and time only through the combination $r - ct$.
- The far field is everywhere transverse ($\vec{B}$ points perpendicular to its direction of propagation $\vec{r}$).
- The far field falls off with distance like $r^{-1}$.

We could now obtain $\vec{E}$ by returning to Equation 43.9, this time working out $A^0$, and using the formula for $\vec{E}$ in terms of the vector and scalar potential. But there’s an easier way. Recall that Ampère’s law says $\frac{d}{dt}(\vec{E} / c^2) = c^2 \vec{V} \times \vec{B}$, and we just found $\vec{B}$. Again use the fact that derivatives of $r^{-1}$ will be subleading and may be dropped in far-field approximation. Furthermore, derivatives of $\vec{r}$ fall with distance like $r^{-1}$, and hence will also generate subleading terms. The leading contribution to $\vec{E}$ therefore comes once again from the retardation factor: $\vec{V}(t - r/c) = -\vec{r}/c$. So

$$\frac{d\vec{E}}{dt} \approx c^2 \left( -\frac{\mu_0}{4\pi r^3} \right) \left( \frac{\vec{r}}{c} \right) \times \left( \hat{r} \times \frac{\partial^2 \vec{D}_i}{\partial t^2} \right)_{|t-r/c|}.$$  

Because everything is harmonic in time, we can just drop one time derivative from both sides of this equation:

$$\vec{E} \cong \vec{E}^{[ED]} \text{, where } \vec{E}^{[ED]} = \frac{\mu_0}{4\pi r} \frac{1}{c^2} \hat{r} \times \left( \hat{r} \times \frac{\partial^2 \vec{D}_i}{\partial t^2} \right)_{|t-r/c|}. \text{ ED approx., far-field}$$

Like $\vec{B}$, the electric field is transverse to the line of sight $\vec{r}$, falls like $r^{-1}$, and involves acceleration of the charge. Moreover, $\vec{E}$ is also perpendicular to $\vec{B}$, a property that we observed some time ago for plane waves. What’s new is that now we know the quantitative relations between the charge’s motion and the amplitude and polarization of the wave.

### 43.5 CONCRETE EXAMPLES

#### 43.5.1 Electric dipole antenna

Usually when we introduce “wires,” we implicitly assume an approximation in which no charge builds up anywhere. That is, usually we ignore the capacitance of a system of
“wires”; for example, if the wires do not form a closed circuit, we assume that no current flows.

However, if we attach an alternating potential source to two diverging, finite-length wires (Figure 43.2), then some current really will flow into and out of them, particularly at high frequency. That current alternately builds up charge along the wires, which in turn creates an oscillating electric dipole moment, which we now know can radiate.

The exact theory of such an “electric dipole antenna” is complicated and involves self-consistently solving for the fields, currents, and charges. Instead of doing this, we now assume a simple form for the currents and charges that is at least consistent with the continuity equation. Suppose that one wire segment stretches from the origin along the $z$ axis to $z = a/2$. Another wire segment stretches the other direction to $z = -a/2$. Alternating current is fed into the top wire at the origin; we will suppose that its amplitude falls linearly to zero at the end of the wire. An equal and opposite current is fed into the lower wire at the origin, so that overall the antenna is always net neutral. Moreover, because the wires run in opposite directions, their respective currents are always parallel.

In a formula, the current in each wire is

$$I(t, z) = \dot{I} \cos(\omega t)(1 - |z|/(a/2)) \quad \text{for} \quad |z| < a/2.$$ 

Current is 1D charge flux, so the 1D continuity equation says

$$\frac{d\rho_q^{[1D]}}{dt} = -\frac{dI}{dz} = -(\dot{I} \cos \omega t)(\pm 2/a)$$

for the upper and lower wires respectively. Thus, $\rho_q^{[1D]} = \pm \frac{2I}{a\omega} \sin \omega t$.

We can now find the dipole moment:

$$\vec{D}_H = \dot{z} \sin \omega t \left[ \int_{-a/2}^{0} zdz \frac{-2I}{a\omega} \right] + \int_{0}^{a/2} zdz \frac{2I}{a\omega} = \dot{z} \frac{Ia}{4\omega} \sin \omega t.$$ 

Substituting into the general dipole radiation formulas then gives the radiation created by this antenna. A distant observer in the $xy$ plane will see radiation linearly polarized along $\hat{z}$. A distant observer along the $z$ axis will see nothing. A distant observer along any other direction will see radiation linearly polarized along the direction obtained by projecting $\hat{z}$ to the plane perpendicular to the line of sight.

43.5.2 IR-active gases absorb and radiate via molecular dipole moments

Absorption and emission by single molecules should properly be treated quantum mechanically (Chapter 55); however, some qualitative features can be understood in our classical picture, starting with ideas introduced in Section 3.7.3 (page 47).

Earth’s surface is kept considerably warmer than would otherwise be the case by its atmosphere. Our atmosphere is largely transparent to visible light from the Sun, yet it intercepts infrared radiation and impedes its escape back out into space. Different gas molecules have very different abilities to absorb and reemit infrared photons, however.

Diatomic, homonuclear molecules have little infrared activity, but more complex ones can be strong greenhouse gases.
Figure 43.3: [Infrared photographs.] Energy absorption by an IR-active gas. Two identical, cylindrical chambers with transparent ends, viewed in the wavelength band 7.5–14 μm. False color indicates radiance in this band (reds are higher than blues); the scale bar is labeled with inferred temperature values in degrees Celsius. The chamber on the right contains dry air. The one on the left contains CO₂. Both have axial length 23 cm and are viewed end-on. (a) Both chambers started at room temperature. False color when looking into each chamber matches the backdrop. (b) The chambers were briefly exposed to infrared light. After irradiation was stopped, the one containing IR-active gas was observed to be slightly warmer for about one minute. That is, more infrared light was observed coming out of this chamber than was the case for either the backdrop or the other chamber. [See also Media 18 and Sieg et al., 2019.]

Optical absorption by a molecule involves its distribution of charge and current. Similarly to what we have seen in this chapter, the most important term is controlled by the “transition dipole,” which is the matrix element of the electric dipole moment operator between the ground and excited molecular states.\(^8\)

The molecules O₂ and N₂, which constitute the bulk of Earth’s atmosphere, are called homonuclear, because they contain two identical nuclei. A homonuclear diatomic molecule is symmetric under inversion, even when strained away from its normal chemical bond length, and hence can have no dipole moment. Thus, the transition dipole between the ground state and either a rotational or vibrational excited state must equal zero. Such excited states are typically separated from the ground state by an energy gap corresponding to infrared light. However, a homonuclear molecule cannot use dipole radiation to leave (nor enter) those states, and hence is a poor absorber of infrared light.

Non-homonuclear diatomic molecules, notably nitric oxide (NO), have nonzero dipole moment in their ground state, which changes when the molecule is set into rotational motion. Moreover, the vibrational modes of such a molecule change its dipole moment. The transition dipoles between the ground state and the rotational and vibrational excited states are therefore nonzero, making NO a strong absorber in the infrared. It is therefore referred to as an infrared-active, or greenhouse gas.

A bent triatomic molecule, such as water (H₂O), also has a permanent dipole moment; water vapor is also a potent infrared-active gas. The carbon dioxide molecule (CO₂) has three nuclei in a linear arrangement, and hence zero dipole moment in its ground state. Thus, its transition dipoles between ground and rotationally excited states vanish. However, it develops a dipole moment in some of its vibrational states; transition dipoles

\(^{8}\)See Chapter 55.
Therefore exist for these and also for mixed rotation-vibration states, making CO_2 another infrared-active gas (Figure 43.3).

### 43.6 ENERGY FLUX AND TOTAL POWER SCALE AS $\omega^4$

In this section, we’ll sometimes drop the suffix “$|_{l-r/c}$” for brevity. We continue to work in the far field, in electric dipole approximation.

Now at last we can see how energy is transported: Its flux is

\[
\mathbf{S}^{[ED]} = \frac{1}{\mu_0} \mathbf{E} \times \mathbf{B} = -\mu_0^{-1} \left( \frac{\mu_0}{4\pi r} \right)^2 \frac{1}{c} \left[ \mathbf{\hat{r}} \times \left( \mathbf{\hat{r}} \times \frac{d^2 \mathbf{D}_E}{dt^2} \right) \right] \times \left[ \frac{d^2 \mathbf{D}_E}{dt^2} \right].
\]

The factor in the brace is $\mathbf{\hat{r}} \left( \mathbf{\hat{r}} \cdot \frac{d^2 \mathbf{D}_E}{dt^2} \right) - \frac{d^2 \mathbf{D}_E}{dt^2} \cdot \mathbf{\hat{r}}$. Now use the triple cross product formula again:

\[
\mathbf{S} = -\mu_0^{-1} \left( \frac{\mu_0}{4\pi r} \right)^2 \frac{1}{c} \left[ \left( \mathbf{\hat{r}} \left( \mathbf{\hat{r}} \cdot \frac{d^2 \mathbf{D}_E}{dt^2} \right) - \frac{d^2 \mathbf{D}_E}{dt^2} \cdot \mathbf{\hat{r}} \right) \mathbf{\hat{r}} \right] - \frac{d^2 \mathbf{D}_E}{dt^2} \left( \mathbf{\hat{r}} \left( \mathbf{\hat{r}} \cdot \frac{d^2 \mathbf{D}_E}{dt^2} \right) - \frac{d^2 \mathbf{D}_E}{dt^2} \cdot \mathbf{\hat{r}} \right)
\]

\[
\mathbf{S}^{[ED]} = \frac{\mu_0}{(4\pi r)^2 c} \left( \left\| \frac{d^2 \mathbf{D}_E}{dt^2} \right\|_{l-r/c} \right)^2 - \left( \left\| \frac{d^2 \mathbf{D}_E}{dt^2} \right\|_{l-r/c} \right)^2 \cdot \text{far-field}
\]

(43.13)

Thus, the energy flux vector always points radially outward. It’s not spherically symmetric, however, because its magnitude depends on the direction $\mathbf{\hat{r}}$ to the observer.

The total power output is the rate at which energy passes through a large spherical shell:9

\[
\mathcal{P}^{[ED]} = \lim_{B \to \infty} \int_{r=B} d^2 \mathbf{S} \cdot \mathbf{S}^{[ED]} = \frac{\mu_0}{(4\pi)^2 c} \frac{d^2 \mathbf{\hat{r}}}{dt^2} \left| \left| \left( \left\| \mathbf{\hat{r}} \right\| - \mathbf{\hat{r}} \right) \mathbf{\hat{r}} \right\| \right| \cdot \frac{d^2 \mathbf{D}_E}{dt^2} \left| \left| \left( \left\| \mathbf{\hat{r}} \right\| - \mathbf{\hat{r}} \right) \mathbf{\hat{r}} \right\| \right|\).
\]

The first term inside the square brackets is the integral over all directions of a constant tensor, that is, $4\pi \mathbf{\hat{1}}$. The second term is $-4\pi$ times the average over all directions of $\mathbf{\hat{r}} \cdot \mathbf{\hat{r}}$. It has no dependence on the observer’s position. Thus, it must be a rotationally-invariant, yet constant, 3-tensor of rank 2. There is only one possibility: This term must be a constant times the identity tensor.10 Moreover, its trace must be $-\int d^2 \mathbf{\hat{r}} = -4\pi$, which fixes the constant to be $-1/3$. All together, then, the factor in square brackets is $4\pi(1 - \frac{1}{3}) \mathbf{\hat{1}}$, and we have

\[
\mathcal{P}^{[ED]} = \frac{\mu_0}{4\pi c} \frac{2 \left| \left| \frac{d^2 \mathbf{\hat{r}}}{dt^2} \right| \right|_{l-r/c} \right|}{3} \cdot \text{total power output, ED approximation}
\]

(43.14)

---

9 Because we only want energy that makes it all the way out to infinity, the far-field approximation is automatically satisfied.

10 You showed this in Problem 14.2.
43.7 LINEAR POLARIZATION GIVES RISE TO A DIPOLE DOUGHNUT PATTERN

Consider the case in which \( \vec{D}_E \) is always directed along a single direction, as in Figure 43.2. We can choose coordinates to make that direction be the \( z \)-axis: \( \vec{D}_E = D_z(t) \hat{z} \). First note a relation between the spherical directions:

\[
\hat{z} = \hat{r} \cos \theta - \hat{\theta} \sin \theta.
\]

**Your Turn 43C**

Show that

\[
\vec{B}^{(ED)} = \frac{\mu_0}{4\pi r c} \left( \frac{d^2 \vec{D}_E}{dt^2} \right)_{t \rightarrow r/c} \sin \theta,
\]

\[
\vec{E}^{(ED)} = \frac{\hat{\theta} \mu_0}{4\pi r} \left( \frac{d^2 \vec{D}_E}{dt^2} \right)_{t \rightarrow r/c} \sin \theta.
\]

Hence, in any direction, a distant observer sees a linearly polarized plane wave.

Turning now to the energy flux,

\[
\left\| \frac{d^2 \vec{D}_E}{dt^2} \right\| = \left\| \frac{d^2 \vec{D}_E}{dt^2} \right\|,
\]

\[
\left( \hat{r} \cdot \frac{d^2 \vec{D}_E}{dt^2} \right)^2 = \left( \frac{d^2 \vec{D}_E}{dt^2} \hat{\theta} \cdot \hat{z} \right)^2 = \left[ \frac{d^2 \vec{D}_E}{dt^2} \right]^2 \cos^2 \theta,
\]

\[
\vec{S}^{(ED)} = \mu_0^{-1} \vec{E} \times \vec{B} = \hat{r} \frac{\mu_0}{(4\pi r)^2 c} \left( \frac{d^2 \vec{D}_E}{dt^2} \right)_{t \rightarrow r/c} \sin^2 \theta.
\]

Equation 43.15 shows the angular dependence explicitly: Energy mostly comes out near the equatorial plane (here the \( xy \) plane).

We can now get the total power output from Equation 43.14. If the dipole varies harmonically in time, then we can write \( \vec{D}_E(t) \) in terms of the amplitude (maximum value) \( \vec{D}_E \) as \( \vec{D}_E(t) = \frac{1}{2} \epsilon^{-i\omega t} \vec{D}_E + \text{c.c.} \). Then the time-averaged power output is

\[
\langle \mathcal{P}^{(ED)} \rangle = \frac{\mu_0}{12\pi c} \omega^4 | \vec{D}_E |^2, \quad \text{total power output, harmonic source}
\]

a famous result.

**Your Turn 43D**

Repeat the exercise, but with \( \vec{D}_E(t) = \vec{D}_E \left[ \frac{\cos \omega t}{0} \right] \) and interpret the result.
FURTHER READING

Intermediate:

Greenhouse gases: Bohren & Clothiaux, 2006, chap. 2. Historical: E. Newton Foote published her results on absorption of sunlight by various gases, and drew the conclusion that increasing CO$_2$ in Earth’s atmosphere would raise surface temperature, well before Tyndall’s better-known work: Foote, 1856.

43.1 Beyond far-field approximation

Background: The main text took the exact vector potential outside an arbitrary localized charge/current distribution, Equation 39.6 (page 595), then simplified by assuming harmonic time dependence of the sources and far field regime (blue triangular region in Figure 43.1, page 642), so that we could discard all $O(r^{-2})$ terms in the fields as well as all but the leading term in an expansion in powers of $a/\lambda$, where $a$ is the source size.

Let’s now consider relaxing the far-field assumption, while continuing to assume that $a$ is much smaller than either $r$ or $c/\omega$ (region marked “near and intermediate” in Figure 43.1); for example, we may consider the limit of an oscillating pure (point) dipole ($\mathbf{D}_0 \rightarrow 0$), with

$$\mathbf{D}_e(t) = \dot{Z}\mathbf{D}_0 \cos(\omega t).$$

We have already found in Chapter 38 that analytic results are still possible, although a bit more complicated than the ones in Section 43.3.1.

Before calculating, let’s frame some expectations. Close to the source, at each instant of time it seems reasonable to expect that the electric field will look like the field around a static dipole. Thus, each field line starts on a $+$ charge and terminates on a $-$ charge. We will soon see, however, that far from the source, this expectation won’t be satisfied.

Do:

a. Use a computer to make an arrow plot of the electric field, Equation 38.4 (page 588), in the limit $k \rightarrow 0$ (static case).

Notes:

- For observation points in the $xz$ plane, the electric field also lies in the $xz$ plane. Thus, a streamline that starts in this plane stays there, so throughout this problem you can restrict to making 2D plots of the field in this plane. (The rest is determined by axial symmetry.)
- It will take some effort to make your plot look nice (that is, physically informative). For example, it may be hard to visualize the answer because the arrows are of such differing lengths. One approach is to instead make an arrow plot of the direction $\mathbf{E} = \mathbf{E}/||\mathbf{E}||$ and superimpose a heat map of $||\mathbf{E}||$ as in Figure 42.2 (page 635). The normalized vector field has the same streamlines as $\mathbf{E}$.

b. Next, try nonzero frequency: Assume $\omega = 2\pi c/(1.5 \text{ cm})$. Again plot $\mathbf{E}$ at time $t = 0$.

c. Continuing (b), make another plot with a lot of representative streamlines of $\mathbf{E}$. Perhaps the region $0 < r < 5(2\pi c/\omega)$, in the quadrant $0 < \theta < \pi/2$, will be a nice region to plot your answer; you decide. Note: In the near field, all the field lines terminate on one of the two charges, but as you move outward, you’ll find some integral curves that do something else. Make some comments about the physics of what you see.

Show some initiative. Imagine that these are figures in a paper you’re trying to publish—figure out some improvements in presentation, informative labels, and so on. If you think...
that the range suggested in (f) doesn’t show the physics optimally, choose some better range. Maybe you’d like to make movies for (e–f) to show the time dependence. Play.

43.2 *Angular momentum of fields II*

**Background:** Problem 38.4 (page 588) described how EM waves can carry angular momentum: The density of angular momentum \( \hat{L}_z \), computed using the origin as reference point, is \( \hat{z} \cdot \frac{1}{\mu_0} \left[ \hat{r} \times (\hat{E} \times \hat{B}) \right] \). As usual, we will suppose that the fields are harmonically varying in time and consider only the time average of our answers.

**Do:**

a. Suppose that we have two oscillating dipoles of strength \( \mathcal{D}_i \) at the origin, pointing at right angles to each other and both in the \( xy \) plane. The dipoles oscillate at the same angular frequency \( \omega \) but 90 deg out of phase. Compute the density of the \( z \) component of angular momentum far away from the origin, to leading order in powers of \( 1/r \). Because everything moves radially outward, the radial component of the flux of \( \hat{L}_z \) is then your answer divided by \( c \).

b. A sphere of large radius surrounds the dipoles and absorbs all the radiation. Before you compute anything: Will the sphere gain any net angular momentum \( \hat{L}_z \)? Why/why not? Now do the calculation using (a), to get the rate at which \( \hat{L}_z \) is transferred to the sphere.

c. Also find the power absorbed by the sphere.

d. Divide your answers to (b,c) and comment.
CHAPTER 44

Higher-Multipole Radiation

44.1 FRAMING: SUPPRESSION

Chapter 43 found a solution to the Maxwell equations that, in the far-field region, becomes approximately a spherical wave potential with amplitude proportional to the time derivative of the electric dipole moment (compare Equations 38.1 and 43.11). Does that mean that a charge and current distribution with electric dipole moment equal to zero (or a constant) cannot radiate? No, we already found in Chapter 25 that a purely magnetic dipole also creates far fields that fall like $r^{-1}$, indeed as a different sort of spherical wave.

To see what’s going on, recall a second approximation made in Section 43.2.3: The electric dipole approximation retained only the first term of a multipole expansion. If that term vanishes, then the leading behavior may involve some higher term. In this chapter we’ll pursue such terms, while still making the far-field approximation.

Electromagnetic phenomenon: Some radiative nuclear transitions are much faster than others.

Physical idea: Quadrupole radiation is suppressed relative to dipole by the multipole parameter squared.

44.2 NEXT-ORDER TERMS

Let’s take a second look at the solution to the Maxwell equations in the small-source regime, again supposing that the current and charge distributions are harmonic in time with angular frequency $\omega$:

$$
\mathcal{A}^\mu(t, \vec{r}) = \frac{\mathcal{B}_0}{4\pi r} e^{-i\omega(t - r/c)} \int d^3 r_s \, e^{-i\epsilon_{\text{multi}} \vec{r}_s / a} \frac{1}{2} \epsilon^{\mu\nu\rho\sigma} \epsilon_{\rho\sigma} \, + \, \text{c.c.} \quad [43.7, \text{page 643}]
$$

Recall that in this formula, $t$ and $\vec{r}$ (and hence also $\hat{r}$) refer to the observation, whereas $\vec{r}_s$ is a source point. Again, we restrict to the multipole regime: $a$ is the overall source size (upper bound on $||\vec{r}_s||$), and $\epsilon_{\text{multi}} = \omega a / c$ is the small quantity controlling the multipole expansion. Now, however, we will retain higher terms in the expansion in $\epsilon_{\text{multi}}$ that were dropped in Chapter 43.

44.2.1 Order-one terms in $\epsilon_{\text{multi}}$ can be divided into two tensor structures

Proceeding as before, we now expand the exponential factor inside the integral in Equation 43.7. Chapter 43 evaluated the zeroth-order term, which we’ll now call $\mathcal{A}^{[0]}$; instead,
now we focus on first order in $\varepsilon_{\text{multi}}$. We’ll call the three spatial components of that term $\vec{A}^{[1]}$: \[
\vec{A}^{[1]}(t, \vec{r}) = \frac{\mu_0}{4\pi r} e^{-i\omega(t-r/c)} \int d^3 r_+ (-i\varepsilon_{\text{multi}} \hat{\beta} \cdot \vec{r}_+ / a) \frac{\vec{r}_+}{2} j(t, \vec{r}_+) + \text{c.c.}
\]

We can write $-i\varepsilon_{\text{multi}} / a$ as \(c^{-1} \frac{d}{dt}\): \[
= \frac{\mu_0}{4\pi r c} \hat{\beta} \cdot \frac{d}{dt} \left[ \int d^3 r_+ \vec{r}_+ \otimes j(t - r/c, \vec{r}_+) \right].
\]

The expression in the brace is a 3-tensor of rank two that depends on the observer’s position only via $t_c = t - r/c$. We’ll call it $\vec{\Gamma}(t_c)$; it is a kind of moment.

Like any second-rank tensor, $\vec{\Gamma}$ can be written as the sum of its symmetric and anti-symmetric pieces, which we’ll call $\vec{\Gamma}^{[\text{EQ}]}$ and $\vec{\Gamma}^{[\text{MD}]}$, respectively.

### 44.2.2 Antisymmetric part of the moment: magnetic dipole radiation

$\vec{\Gamma}^{[\text{MD}]}$ already arose in Chapter 17, where we re-expressed its three independent entries as a vector:
\[
\vec{\Gamma}^{[\text{MD}]} = \varepsilon_{npl} \vec{D}_{M,j} \quad \text{where} \quad \vec{D}_{M,j} = \frac{1}{2} \varepsilon_{klm} \int d^3 r_+ \vec{F}_k \vec{\dot{j}}_l.
\]

**Your Turn 44A**

Show that $\vec{\Gamma}^{[\text{MD}]}$’s contribution to $\vec{A}^{[1]}$ is
\[
\vec{A}^{[\text{MD}]} = -\frac{\mu_0}{4\pi rc} \hat{\beta} \times \left( \frac{d}{dt} \vec{D}_{M} \right)_{t=r/c} \quad \text{MD approximation, small source.}
\]

Your result implies that

- Once again, this part of the field is a spherical wave (because the wave crests of $\vec{A}^{[\text{MD}]}$ lie on the spherical shells $ct - r = \text{constant}$).
- $\vec{A}^{[\text{MD}]}$ falls like $r^{-1}$, and hence can potentially transport energy to infinity.

To get simple formulas for the fields, we again specialize to the far field (Equation 43.1, page 641):

**Your Turn 44B**

a. Do a calculation similar to the one in Section 43.4 to show that
\[
\vec{B}^{[\text{MD}]} = \frac{\mu_0}{4\pi r c^2} \hat{\beta} \times \left( \hat{\beta} \times \left( \frac{d^2}{dt^2} \vec{D}_{M} \right)_{t=r/c} \right) \quad \text{far field} \quad (44.1)
\]
b. Then use the Ampère law to find $\vec{E}^{[\text{MD}]}$. 

Contents  Index  Notation  Glossary  November 14, 2023
Remarkably,

*The MD contribution to the magnetic far field looks just like the ED contribution to the electric far field. The MD contribution to the electric far field looks just like the ED contribution to the magnetic far field.*

Consider a circular loop of wire in the xy plane, with area $\Sigma$ and carrying current with amplitude $I$ and angular frequency $\omega$. It has no net charge anywhere, and hence vanishing electric dipole and quadrupole moments. But you found in Your Turn 17A (page 254) that the magnetic dipole moment is nonzero:

$$D_M = (2\Sigma)(I \cos \omega t).$$

**Your Turn 44C**

a. Find the far electric and magnetic fields and compare to your earlier result obtained in Coulomb gauge (Your Turn 25E, page 397).

b. Find the Poynting vector (Equation 35.12, page 561) and compare with the result in Coulomb gauge (Your Turn 25E).

c. Integrate the Poynting vector over all directions $\hat{r}$ to obtain the power $\mathcal{P}$.

The time-average of your result gives

$$\langle \mathcal{P}[\text{MD}] \rangle = \frac{\mu_0}{12\pi c} \omega^4 c^{-2} |\mathcal{D}_m|^2,$$

where

$$|\mathcal{D}_m|^2$$

is the total power output, harmonic source (44.2)

That result has the same frequency dependence as the ED result (Equation 43.16, page 650).

**44.2.3 Symmetric part of the moment: electric quadrupole radiation**

Next we turn to the symmetric part of $\vec{\tau}$. To simplify $\vec{\tau}^{[EQ]}$, use a trick introduced in magnetostatics (Section 17.2): The divergence theorem gives that

$$0 = \int d^3 r^* \nabla \cdot (r^*_s \tilde{r}^{m*} \tilde{j}_m(r^*_s)),$$

where $\vec{V}_s$ denotes partial derivatives with respect to $r^*_s$. Thus,

$$\tau^{[EQ]}_{mk} = \frac{1}{2} \int d^3 r^* \left( \tilde{r}^{m*} \tilde{j}_k(r^*_s) + \tilde{r}^{k*} \tilde{j}_m(r^*_s) \right) = \frac{1}{2} \int d^3 r^* \tilde{r}^{m*} r^*_s \tilde{r}^k \cdot \vec{j}.$$

For a static current distribution, this quantity would be zero by the continuity equation. More generally, however, we get

$$\frac{1}{2} \frac{d}{dt} \int d^3 r^* \tilde{r}^{m*} r^*_s \tilde{r}^k \rho_q \bigg|_{t-r/c}.$$

That is, this term involves the second moment of electric charge. We can write that moment as its traceless part plus the rest, by using Equation 3.3 (page 38):

$$\frac{1}{3} \tilde{Q}^{EM}_j \bigg|_{t-r/c} + \frac{1}{3} \tilde{P}^{EM}_j \int d^3 r^* r^*_s \rho_q \bigg|_{t-r/c}.$$
44.3 Higher Order Terms can Contribute Even if Lower Ones are Zero

So the contribution of $\mathcal{F}_1^{[E]Q}$ to the first-order term of the vector potential, $\vec{A}^{[1]}$, can be written as

$$\vec{A}^{[E]Q} = \frac{1}{64\pi r c} \frac{d}{dt} \mathbf{r} \cdot \left( \frac{d}{dt} \mathbf{Q}_e \right)_{|t-r/c} + \frac{1}{c} \int d^3r e^{2 \frac{d}{dt} \mathbf{r} \cdot \mathbf{E}(t-r/c, \vec{r}_c)}$$

$$= \frac{\mu_0}{24\pi c} \left[ r^{-1} \mathbf{r} \cdot \frac{d^2}{dt^2} \mathbf{Q}_e \right]_{|t-r/c} + \hat{r} r^{-1} \int d^3r e^{2 \frac{d}{dt} \mathbf{r} \cdot \mathbf{E}(t-r/c, \vec{r}_c)}.$$

The second term of this expression looks complicated, but it’s purely a gradient, and hence cannot contribute to the magnetic field. Equivalently, it can be removed by an appropriate gauge transformation, leaving

$$\vec{A}^{[E]Q} = \frac{\mu_0}{24\pi cr} \mathbf{r} \cdot \frac{d^2}{dt^2} \mathbf{Q}_e \big|_{|t-r/c}.$$

Once again, we have found an outgoing spherical wave (the potential depends harmonically on $t-r/c$), falling in the far field region like $r^{-1}$. Compared with electric dipole radiation, EQ radiation is suppressed by an extra factor of $\epsilon_{\text{quadrupole}} = \omega \alpha_c$, but it can be the leading term for a source whose dipole moment equals zero.

44.3 HIGHER ORDER TERMS CAN CONTRIBUTE EVEN IF LOWER ONES ARE ZERO

44.3.1 Higher multipole fields can again transport energy to infinity

Clearly we could carry out the expansion to next order in $\epsilon_{\text{mult}}$ to find $\vec{A}^{[2]}$, with contributions from magnetic quadrupole and other terms. In the electrostatic and magnetostatic multipole expansions, we found that each successive order gave fields falling off with distance faster than the previous one. In contrast, for time-varying sources

*Every order of the multipole expansion gives a contribution whose leading far-field behavior is always $1/r$. Each order is suppressed relative to the previous one by an additional factor of frequency.*

Thus, all of the orders create outgoing spherical waves, so they can all transport energy to infinity.

In greater detail, we have in the far field approximation

$$\frac{d}{dt} \mathbf{E} = c^2 \nabla \times \mathbf{B} \approx -c \hat{r} \times \mathbf{B}$$

so the energy flux is

$$\vec{S} = \frac{1}{\mu_0} \mathbf{E} \times \mathbf{B} = \frac{c}{\mu_0} \hat{B} \times (\hat{r} \times \mathbf{B}) = \frac{c}{\mu_0} \hat{r} ||\mathbf{B}||^2,$$

and each nonzero term of

$$||\mathbf{B}^{[0]} + \mathbf{B}^{[1]} + \mathbf{B}^{[2]} + \cdots||^2$$

(44.4)
has a far-field part that falls with distance as $r^{-2}$.

Let’s consider the various contributions according to their order in the multipole expansion parameter. Equation 43.13 (page 649) gave the $\|B^{[0]}\|^2$ term (electric dipole), and Equation 43.16 (page 650) gave its integral over all directions. If this term is nonzero, it’s likely the most important one.

The cross term $2\vec{B}^{[0]} \cdot \vec{B}^{[1]}$ integrated over angles gives zero. So the next most important terms involve $\|B^{[1]}\|^2$ (magnetic dipole and electric quadrupole, Problem 44.6) and $2\vec{B}^{[0]} \cdot \vec{B}^{[2]}$ (the “anapole” term).

44.3.2 Qualitative approach to nuclear radiative transitions

Some atomic nuclei make transitions that result in the emission of light. The lifetimes of these transitions vary widely; for example, $^{13}$N has a state that emit a photon with half-life $T_N \approx 10^{-15}$ s; in contrast, $^{197}$Hg has a state with half-life $T_{197\text{Hg}} \approx 7 \cdot 10^{-9}$ s.

Although the emitted light is in the gamma ray part of the spectrum, and hence far shorter wavelength than atomic transitions, it is also the case that nuclei are far smaller than atoms, and in fact the ratio $\epsilon_{\text{multi}} = a/(c/\omega)$ is again small. Thus, we expect that a multipole expansion will again be useful for reconciling these observations. Indeed, angular momentum conservation and facts about the initial and final state imply that the excited state of $^{13}$N may transition via its electric dipole moment, whereas $^{197}$Hg can only emit via electric quadrupole radiation.

Gamma ray emission is quantum mechanical in character (usually just one photon is emitted). Nevertheless, we can make a semiclassical estimate to understand the wide range of emission lifetimes. Thus, we will estimate the mean rate of photon emission as our classical estimate of the total energy emission rate, divided by the energy of one emitted photon, which is $\hbar \omega$. For ED radiation, Equation 43.14 (page 649) gives

$$\text{transition rate} = \frac{c^{-1} \mu \omega^2 a^2 \omega^4}{\hbar \omega}.$$  

The present chapter has shown that EQ radiation is suppressed by an additional factor of $(a/\omega)^2$. We thus have

$$\text{transition rate} \approx \frac{a_N^2 \omega_N^4}{\hbar \omega_N} / \frac{a_{197\text{Hg}}^4 \omega_{197\text{Hg}}^6 c^{-2}}{\hbar \omega_{197\text{Hg}}}.$$  

In addition, the two nuclei under study have different overall size; empirically, nuclei are essentially close-packed spheres of nucleons, and so $a \approx (1.4 \cdot 10^{-15} \text{ m}) A^{1/3}$, where $A = 13$ or 197, respectively.

Your Turn 44D

Evaluate the dimensionless ratio Equation 44.5, using observed energies for these transitions: $\hbar \omega_N = 2.38 \text{ MeV}$ and $\hbar \omega_{197\text{Hg}} = 0.13 \text{ MeV}$. (It’s useful to recall $\hbar \approx 6.5 \cdot 10^{-22} \text{ MeV s}$.) Compare to the experimental values at the start of this section.

---

2See Problem 44.5.
3See Chapter 55.
44.4 PLUS ULTRA

A spherically symmetric charge distribution will not radiate, no matter how it depends on time. Indeed, its monopole moment is fixed by charge conservation, and hence has no time dependence; we saw above how the first orders of the expansion involve $\vec{D}_E, \vec{D}_M, \vec{Q}_E$, and so on, all of which are zero for a spherically symmetric distribution.

FURTHER READING

Intermediate:

PROBLEMS

44.1 Pure MD antenna
Chapter 25 discussed the radiation we see when standing far away from an oscillating magnetic dipole. Specifically, the dipole was oriented with its moment in the $\pm \hat{z}$ direction, we imagined measuring the fields at $\vec{r} = (r, 0, 0)$, and we only asked for the leading order term in powers of $1/r$. You found a formula for the vector potential (Your Turn 25E), but even with the far-field limit it still involved a complicated integral. In this problem, you’ll find a simplified expression in a special limiting case.

Consider a series of loops with smaller and smaller radii $a$. However, each loop also has a larger current than the previous one, in such a way that the magnetic dipole moment $\mathcal{D}_M(t) = \mathcal{D}_M \cos(\omega t)$ is the same for all. In this small-source limit (and still large $r$), find a simplified form for the vector potential, magnetic field, and electric field observed far from the source along the $x$ axis. If the outgoing wave is polarized, describe its polarization. Also characterize how the energy density falls with distance.

44.2 Double-loop antenna
Chapter 25 considered an antenna consisting of a circular loop of wire driven by an oscillator. In this problem, consider an antenna consisting of two circular loops, each of radius $a$ and parallel to the $xy$ plane, and centered on the $z$ axis at heights $z = \pm a$. We observe at a distance $r$, and $a \ll c/\omega \ll r$. The currents in the loops are $\pm \frac{1}{2} (I\text{e}^{i\omega t} + \text{c.c.})$ respectively.

Find the lowest-order multipole radiation fields produced by this system. [Hint: You could invent the magnetic quadrupole radiation formula for this purpose. But this is not an arbitrary quadrupole, so a simpler procedure works. Write the far fields of a single oscillating dipole in the $xy$ plane. Shift them along $\pm \hat{z}$ by $a$. Subtract those two expressions, simplify, and find the far-field part.]

---

4This “pure dipole” limit is similar to the one in Section 43.3.2 (page 645).
44.3 Pulsar

A pulsar is a compact star (specifically a neutron star), with a large magnetic dipole moment $\mathcal{D}_m$. As with an ordinary permanent magnet, the moment is frozen into the star. That is, $\mathcal{D}_m$ also rotates with the star’s angular velocity $\omega$. Moreover, a neutron star is likely to be spinning rapidly, by conservation of angular momentum during the collapse that formed it.

The dipole moment is located at the center and oriented at some fixed angle $\alpha$ relative to the rotation axis. The moment $\mathcal{D}_m$ is not directly observable, but it is related to the strength $B_{\text{pole}}$ of the magnetic field at the star’s surface, which is indirectly observable via the Zeeman splitting of spectral lines (Problem 18.7, page 293). So let $\|\mathcal{D}_m\| = \kappa B_{\text{pole}}$, where $\kappa$ is a constant.

a. Find the rate at which the pulsar radiates electromagnetic energy, in terms of $\kappa$, $B_{\text{pole}}$, $\omega$, and $\alpha$.

b. If the source of the energy is the pulsar’s rotational kinetic energy, $E = \frac{1}{2}I\omega^2$ with $I =$ pulsar’s moment of inertia, find the characteristic slowdown time scale $T \equiv (\omega^{-1}d\omega/dt)^{-1}$ at time zero, as a function of $\kappa$, $B_{\text{pole}}$, $\omega$, and $\alpha$, and $J$.

c. Suppose that the pulsar has radius $R$, and derive a formula for $\kappa$ in terms of $R$.

d. Thus, we get a prediction of the slowdown in terms of $B_{\text{pole}}, R, \omega, \alpha$, and $J$. Moreover, $J$ is related to $R$ and the pulsar’s mass in the usual way, because a neutron star is essentially a rigid sphere of uniform mass density. Use typical numbers $M = 1$ solar mass $= 2 \cdot 10^{30}$ kg, $R = 10$ km, $B_0 = 10^8$ T, and assume $\alpha = 90$ deg. Evaluate $\mathcal{P}$ and $\tau$ for $\omega(0) = 10^4$ s$^{-1}$, a frequency thought to be typical of newly formed pulsars.

44.4 $T_2$ Exact MD wave

First do Problem 44.1. In this problem, however, find the $E$ and $B$ fields without assuming that $r \gg c/\omega$. That is, work out the near and far fields of an oscillating pure magnetic dipole, that is, one for which $a \to 0$ and hence $a \ll c/\omega$ and $a \ll r$. [Remark: The answer is a new kind of exact, spherical-wave solution to the Maxwell equations, different from the one in Chapter 38 (Problem 38.1, page 588).]

44.5 $T_2$ No cross-term

Using Equations 43.12 (page 645), 44.1, and 44.3, show that the cross term $2B^{[0]} \cdot B^{[1]}$ integrated over angles gives zero. Thus, there is no term in the total radiated power that is first order in the small parameter $\varepsilon_{\text{multi}}$. [Hint: You will encounter the angular average of $\hat{r}_i \hat{r}_j \hat{r}_k$. It must be a rotationally invariant, symmetric, 3-tensor of rank 3. Is such an object possible?]

44.6 $T_2$ Another cross-term

When we expand Equation 44.4, the term $\|\tilde{B}^{[1]}\|^2$ includes the cross term $2\tilde{B}^{[EQ]} \cdot \tilde{B}^{[MD]}$. Show that in the far-field approximation, this term gives zero when integrated over outgoing directions $\hat{r}$, leaving only the contribution already found in Equation 44.2 (page 656), plus one other subterm that you are to find.

44.7 $T_2$ Anapole term
CHAPTER 45

Synchrotron Radiation

45.1 FRAMING: BEAMING

This chapter will explore one particular example of radiation by an accelerating charge: uniform circular motion of a point charge. The resulting synchrotron radiation deserves an entire chapter, because its applications span many fields:

- In astrophysics, energetic charged particles trapped in strong magnetic fields will execute such motion, leading to a prevalent form of radiation, for example, in an accretion disk surrounding a black hole or neutron star ([Figure 45.1]).
- On Earth, a particle accelerator can play a similar role. The resulting energy loss can be an inconvenience (sapping kinetic energy from the particles that we were trying to accelerate).
- But it can also be a boon, because as we’ll see, the resulting radiation can be intense and directional, making it ideal for x-ray crystallography. Much of structural biology now relies on “synchrotron light sources.”

Figure 45.1: [Computer reconstruction.] The black hole at the center of the galaxy M87. The color scale represents the inferred pattern of total intensity at wavelength 1.3 mm in all polarizations. Superimposed dark lines are streamlines of the electric field polarization direction, in regions where the degree of polarization is large enough for this to be meaningful. (Spatial patterns of polarization can be washed out by the finite resolution of the instrument.) The spatial modulation of polarization results from the initial synchrotron radiation of matter falling into the black hole and subsequent Faraday rotation by passage through other plasma (Chapter 54). Both of those phenomena are dependent on magnetic fields, so the pattern gives clues to the strength and direction of those fields. [From Akiyama & others (Event Horizon Telescope collaboration), 2021a.]
The analysis in Chapter 43 may seem to lead to some qualitative conclusions: A charge executing a circular orbit of radius $b$ has a rotating electric dipole moment, so we might expect the emitted radiation to be predominantly circularly polarized and mainly directed perpendicular to the charge’s orbital plane, but with a broad angular distribution characteristic of electric dipole radiation. Moreover, the centripetal acceleration $a^2 b = (\omega b)^2 / b$ may never exceed $c^2 / b$, so we might expect a limiting amount of radiation as energy increases.

Actually, both of the predictions just made fail as we push the charge up to relativistic speeds. Perhaps this should not be surprising—the multipole expansion breaks down when the source is moving relativistically.\(^1\) Instead, we’ll see that the intensity of synchrotron radiation has no upper limit, and that in the relativistic case the radiation is emitted in a tight beam whose direction sweeps around the plane of the orbit. It’s hard to focus x rays, so this beaming effect is useful for technological applications like crystallography.

### 45.2 THE LIÉNARD–WEICHERT FIELDS

We wish to find the electromagnetic fields far from a point charge. For this, we must compute derivatives of the Liénard–Weichert potentials. Section 39.5.2 (page 598) did this in a special case (charged particle in uniform, straight-line motion). We will now start over for a general specified trajectory, then later specialize to the case of uniform circular motion.

At first, the potentials look straightforward:

$$A(X) = \frac{\mu_0 q c}{4\pi} \frac{1}{(X - \Gamma(\tau_r)) \cdot U_{\tau_r}}, \quad \text{where} \quad \|X - \Gamma(\tau_r)\|^2 = 0. \quad [39.13, \text{page 598}]$$

But recall that in addition to their explicit dependence on the observation point $X$, the potentials involve the retarded proper time $\tau_r$, which also implicitly depends on $X$. Fortunately, we only need the variation of $\tau_r$ in order to take the first derivatives needed for the Faraday tensor. So let’s vary the defining relation just given, which says that $\Gamma(\tau_r)$ must lie on the past light cone of $X$:

$$0 = \|X - \Gamma(\tau_r)\|^2 = \|X + dX - \Gamma(\tau_r + d\tau_r)\|^2.$$

Define $\Gamma_{\tau_r} = \Gamma(\tau_r)$ and collect the first-order terms:

$$0 = (X - \Gamma_{\tau_r}) \cdot (dX - d\tau_r U_{\tau_r})$$

$$d\tau_r = \frac{(X - \Gamma_{\tau_r}) \cdot dX}{(X - \Gamma_{\tau_r}) \cdot U_{\tau_r}} \quad \text{or} \quad \frac{d\tau_r}{\delta X^\mu} = \frac{(X - \Gamma_{\tau_r})^\mu}{(X - \Gamma_{\tau_r}) \cdot U_{\tau_r}}. \quad (45.1)$$

Introduce the abbreviations

$$\alpha(X, \tau) = - (X - \Gamma(\tau)) \cdot U(\tau) \quad \text{and} \quad \zeta(X) = \alpha(X, \tau_r(X)). \quad (45.2)$$

\(^1\)See Section 43.2.3.
Then the derivatives of Equation 39.13 become

\[
\frac{4\pi}{\mu_0 q c^2} \partial_{\mu} A^\nu = \frac{\partial}{\partial X^\mu} \left( \frac{U^\nu(\tau_{r}(X))}{\zeta} \right).
\]

In the preceding formula, the derivative acts on every \( X \) dependence. We now reformulate by using the Chain Rule:

\[
= -\frac{U^\nu}{\xi^2} \frac{\partial \alpha}{\partial X^\mu} r + \frac{\partial \tau_{r}}{\partial X^\mu} \frac{\partial}{\partial \tau_{r}} \left( \frac{U^\nu}{\alpha} \right).
\]

Evaluate the second term via Equation 45.1 and use the abbreviations Equation 45.2:

\[
= \left[ \zeta^{-2} U^\nu U^\mu_{r} - \zeta^{-1}(X - \Gamma_{r})^\mu \frac{\partial}{\partial \tau_{r}} \left( \frac{U^\nu}{\alpha} \right) \right]. \tag{45.3}
\]

The Faraday tensor (in this context called the Liénard–Weichert fields) involves this expression minus \((\mu = \nu)\), so we need not keep the symmetric first term. The remaining term involves the derivative of 4-velocity, in keeping with our expectation that acceleration is required for a charge to radiate.²

Next, recall that Section 39.5.1 (page 596) made \( \alpha \) more explicit. As usual, let \( \vec{\beta} = \frac{d\vec{r}}{d(ct)} \) and \( \gamma = (1 - \beta^2)^{-1/2} \). Also let \( \tau_{r} = \Gamma^0(\tau), \vec{R}_{r} = \vec{r} - \vec{\Gamma}_{r} \), and so on. Thus, \( \tau_{r} \) is the lab time at which a disturbance destined to be observed at \( \vec{X} \) is created by the charge. Then Equation 39.12 (page 598) gave that

\[
\zeta = c \gamma r_{l}(1 - \vec{R}_{r} \cdot \vec{\beta}_{r}). \tag{45.4}
\]

Two qualitative features of this formula will be important:

- For a fast particle \((\beta_{r} \approx 1)\), \(1/\zeta\) is sharply peaked in the direction of the particle’s retarded velocity.
- \(1/\zeta\) falls with distance as \(1/R_{r}\).

The first point suggests that for a highly relativistic particle, those fields will be concentrated on a direction parallel to the charged particle’s motion at the retarded time: a “searchlight” pattern. The following analysis will confirm that expectations. The second point may be cause for concern: If the vector potential falls like \(1/R_{r}\), won’t its derivatives fall as \(1/R_{r}^2\) ? But we have already met a similar situation with the loop antenna (Section 25.5.2, page 394). There we found that the time dependence of the source allowed a slowly falling field. Indeed, we see this again in the second term of Equation 45.3, which involves \(\zeta^{-1}\), not \(\zeta^{-2}\), and hence will have the slow falloff characteristic of radiation.

### 45.3 UNIFORM CIRCULAR MOTION

²See Figure 42.1 (page 634) and Section 43.4.
Chapter 45  Synchrotron Radiation

Figure 45.2: Variables defined in the text. The observation point \( \vec{r} \) (top) sits in the \( xy \) plane, offset horizontally by the radius \( b \) of the circular orbit. The asterisk on the orbit denotes the position of a charged particle at the moment when its lab time \( t_* \), and its proper time \( \tau \), are both zero. The observation is made later, at lab times close to \( t_i = D/c \).

45.3.1 Kinematics

Rather than continuing to develop Equation 45.3 in general, we now specialize to the situation of interest for this chapter: a point charge in uniform circular motion in the \( xy \) plane (Figure 45.2).

The trajectory has angular frequency \( \omega \) and radius \( b \):

\[
\vec{r}(t_*) = b(\dot{x}\cos(\omega t_*) + \dot{y}\sin(\omega t_*)).
\]

This motion has the nice feature that \( \beta = \dot{b}\omega/c, \gamma = (1 - \beta^2)^{-1/2} \), and \( U^0 = \gamma c \) are all constant in time; hence \( t_* = \gamma \tau \) is a linear relation. Also,

\[
\vec{\beta}(\tau) = \beta(\dot{x}\sin(\omega\gamma \tau) + \dot{y}\cos(\omega\gamma \tau)) \quad \text{and} \quad \vec{U} = \gamma c \vec{\beta}.
\]

In a typical application, a specimen to be probed by x-ray crystallography sits at a fixed, distant point in the \( xy \) plane. We then ask for the time course of the fields received at that point. By circular symmetry, it doesn’t matter where the observer sits in the plane; a convenient choice is \( \vec{r} = b\dot{x} + D\dot{y} \), where \( D \) is some large, constant distance (Figure 45.2). Ultimately, we will look specifically at the leading behavior in an expansion in \( b/D \) (the far fields).

Our problem has only one length scale, the radius \( b \) (\( D \) will later be sent to infinity), so it will streamline our formulas to express lengths as dimensionless quantities times \( b \). There is also only one time scale, the inverse orbital angular frequency \( \omega^{-1} \). We have some freedom in how we nondimensionalize time; it will be most convenient to represent times as multiples of \( (\omega\gamma)^{-1} \):

\[
\vec{D} = D/b, \quad \vec{i} = \omega\gamma t, \quad \vec{\tau} = \omega\gamma \tau.
\]

Introduce abbreviations for some unit vectors in the \( xy \) plane:

\[
\vec{G}(\vec{\tau}) = \begin{bmatrix} \cos \vec{\tau} \\ \sin \vec{\tau} \end{bmatrix}, \quad \vec{H}(\vec{\tau}) = \begin{bmatrix} -\sin \vec{\tau} \\ \cos \vec{\tau} \end{bmatrix}, \quad \text{so} \quad \frac{d\vec{G}}{d\tau} = \vec{H} \quad \text{and} \quad \frac{d\vec{H}}{d\tau} = -\vec{G}.
\]

To summarize, we then have

\[
\Gamma(\vec{\tau}) = \frac{c\beta}{\omega} \begin{bmatrix} t/\beta \\ \vec{H} \end{bmatrix}, \quad \vec{U}(\vec{\tau}) = \gamma c \begin{bmatrix} 1/\beta H \end{bmatrix}, \quad \text{and} \quad X = \frac{c\beta}{\omega} \begin{bmatrix} 1/\beta H \end{bmatrix}.
\]
where $\bar{t}$ is dimensionless proper time for the particle and $\bar{\tau}$ is dimensionless observation time. Equation 45.7 has abbreviated by dropping the fourth entry in each 4-vector, because in each case it is the constant zero.

For circular motion, Equations 45.2, 45.5, and 45.7 give

$$\alpha(X, \tau) = -b\left[\gamma/\beta, 1, 1, \beta \sin \bar{t}, \beta \cos \bar{t} \right].$$

The velocity is always perpendicular to the displacement, and we may replace $b\omega/c$ by $\beta$:

$$\alpha(X, \tau) = (c^2\gamma/\omega)(\gamma \tau - \bar{\tau} + \beta^2(\sin \bar{t} - D \cos \bar{t})).$$

Finally, it will be convenient to nondimensionalize $\xi$:

$$\xi(X) = \xi(X_0\omega/(\gamma c^2)) = \alpha(X, \tau_0(X))\omega/(\gamma c^2).$$

### 45.3.2 High speed limit

Section 45.2 suggested that for a fast particle, the observer would see short pulses of radiation, as the retarded charge velocity (“searchlight”) periodically aligns with the line of sight. So we will examine a small range of lab times close to an initial time $t_i$ and later confirm that the radiation is confined to such a small range.

The velocity $\hat{\beta}$ aligns with the line of sight when the retarded proper time is $\tau_{r,i} = 0$ (Figure 45.2). Then the light-cone condition says $||X_0 - \Gamma_{r,i}|| = 0$ or

$$c(t_0 - \gamma \tau_{r,i}) = ||b\hat{x} + D\hat{y} - b\hat{x}|| = D,$$

so

$$t_i = D/c \quad \text{or} \quad t_i = \beta \gamma D.$$

We will therefore consider only observations at lab times close to this value.

### 45.3.3 Far-field limit

The formulas are getting long, so it is time to simplify by looking only at the far fields. First recall that $t$ is close to $t_i = D/c$, so dropping terms without any $D$ factor in Equations 45.8 and 45.9 gives

$$\xi \approx D\beta(1 - \beta \cos \bar{t}_r).$$

We will abbreviate by denoting the factor in parentheses by $\eta$, a dimensionless, $D$-independent quantity that we will call the “beaming factor.” Thus,

$$\xi \rightarrow D\beta \eta, \quad \text{where} \quad \eta = 1 - \beta \cos \bar{t}_r, \quad \text{far field}$$

Again we see that factors of $\xi^{-1}$ will be strongly peaked around $\bar{t} \approx 0$ if $\beta$ is close to 1.

### 45.3.4 Electric field

We can now evaluate the derivatives needed in Equation 45.3: Use $U^0 = \gamma c$ and Equation 45.8 to find

$$\frac{\partial}{\partial \tau} \left( \frac{U^0}{\alpha} \right) = \frac{\gamma \alpha^2}{c^2 \xi^2} (1 - \beta^2(\cos \bar{t}_r + D \sin \bar{t}_r)).$$
Next, Equations 45.5, 45.6, and 45.8 give

\[
\frac{\partial}{\partial \tilde{r}} \left( \left. \left( \frac{\mu_0 q c^2 \tilde{E}}{\gamma \omega} \right) \right|_{\tilde{r}} \right) = -\gamma \beta \omega^2 c^{-1} \left[ \tilde{\xi}^{-1} \dot{G} + \tilde{\xi}^{-2} \dot{H}(-1 + \beta^2 (\cos \tilde{r} + \dot{D} \sin \tilde{r}))) \right] \\
= -\frac{\gamma \omega^2}{c D \eta} \left( \dot{G} + \beta \eta^{-1} \dot{H} \sin \tilde{r} \right), \quad \text{far field (45.13)}
\]

Your Turn 45A

Substitute Equations 45.12–45.13 into Equation 45.3 to find the electric field \( \tilde{E}_i = c E^{(0)} \):

\[
\frac{4 \pi \mu_0 q c^2 \tilde{E}}{\gamma \omega} \rightarrow (\omega/c)^2 (\eta^{-2} D^{-1} \dot{G} + \eta^{-3} D^{-1} \sin \tilde{r} (\dot{H} \beta - \dot{y})) \\
= (\omega/c)^2 D^{-1} \left[ \dot{G} \eta^{-2} + (\beta \dot{H} - \dot{y}) \sin \tilde{r} \eta^{-3} \right], \quad \text{far field (45.14)}
\]

At last, we have found a compact, explicit formula for the electric far field, and indeed it has a radiative component, that is, it falls as \( 1/D \). We will soon confirm that the peak radiation intensity occurs at \( \tilde{r} = 0 \); here Equation 45.14 is polarized transversely to the line of sight (along \( \hat{x} \), as befits a radiation field.

45.3.5 Magnetic fields

We can calculate \( \tilde{B}_k = \frac{1}{2} \varepsilon_{k \ell m} F_{\ell m} \) by using similar steps.\(^3\)

Your Turn 45B

Show that the far fields are:

\[
\frac{4 \pi \mu_0 q c^2 \tilde{B}}{\gamma \omega} \rightarrow D^{-1} \frac{\omega^2}{c^2 \eta^2} \times (\dot{G} + (\beta \eta)^{-1} \dot{H} \beta^2 \sin \tilde{r}) \\
= D^{-1} \frac{\omega^2}{c^2 \eta^2} (-\cos \tilde{r} + \beta \eta^{-1} \sin^2 \tilde{r} \dot{z}), \quad \text{far field (45.15)}
\]

Again, the leading behavior is \( \propto D^{-1} \). Also, the magnetic far field is polarized perpendicular to the orbital plane, and hence, transverse both to \( \tilde{E} \) and to the line of sight.

\(^3\)Oliver Heaviside gave a derivation of this result in 1904.
45.3.6 Energy flux

The energy flux\(^4\) observed at \(X\) is given by the Poynting vector:

\[
\vec{S}_i = c \vec{E}_i = -c \mu_0^3 \epsilon_0^3 \epsilon_j \frac{\partial}{\partial \tau} E_j = c \mu_0^3 (\vec{E} \times \vec{B})_i
\]

Using our far field formulas,

\[
\vec{S} = c \mu_0^3 \left( \frac{\mu_0 q \omega}{4 \pi c} \right)^2 D^{-3} \left[ \eta^{-2} \dot{G} + \eta^{-3} (\beta \dot{H} - \dot{y}) \sin \tau_r \right] \times \frac{\omega^2}{\eta^2} \hat{z} \left( -\cos \tau_r + \beta \eta^{-1} \sin^2 \tau_r \right).
\]

Use the identities \(\dot{G} \times \dot{z} = -\dot{H}, \dot{H} \times \dot{z} = \dot{G}\), and \(\hat{y} \times \hat{z} = \hat{x}\) to get

\[
\vec{S} = \frac{\mu_0 q^2 \omega^4}{(4 \pi D)^2 c} \left[ \dot{H} \eta^{-4} - (\dot{G} \beta - \dot{x}) \eta^{-5} \sin \tau_r \right] (\cos \tau_r - \beta \eta^{-1} \sin^2 \tau_r).
\]

A detector facing toward \(-\hat{y}\) then captures energy flux

\[
\vec{S} \cdot \hat{y} = \text{const} \times \eta^{-4} \left[ \cos \tau_r - \beta \eta^{-1} \sin^2 \tau_r \right]^2.
\]

Multiplying this formula by \(D^2\) gives the outward power per solid angle received at a fixed location in the orbital plane.

To use Equation 45.16, first compute \(\beta = b \omega / c\). We must then find the retarded proper times \(\tau_r\), corresponding to lab times \(t\). Although the relation is complicated, we previously argued that at high speeds, we will only need to consider lab times very close to \(t\). So expand to first order about that moment and use Equation 45.1 (page 664):

\[
\tau_r = \tau_{r'} + c(t - t_i) \frac{\partial \tau_r}{\partial x^0} \bigg|_{t_i} + \cdots \approx -c(t - t_i) \frac{c t_i}{-c t_i U^0_{\tau_i} + (b \dot{x} + D \dot{y} - b \dot{\xi}) \cdot \vec{U}_{\tau_i}}.
\]

In dimensionless variables,

\[
\tau_r \approx (t / \gamma - D \beta)(1 - \beta)^{-1}.
\]

Reassuringly, in the limit \(\omega \to 0\) this formula says that retarded proper time advances as a constant plus observation time. Equation 45.18 lets us translate from lab time to \(\tau_r\), substitute into \(\eta = (1 - \beta \cos \tau_r)\) and Equation 45.16, and at last evaluate Equation 45.16.

Let us make the beaming factor more explicit for an ultrarelativistic particle. In this case, \(\beta \approx 1\) and we have \(\gamma = ((1 - \beta)(1 + \beta))^{-1/2} \approx (2(1 - \beta))^{-1/2}\). Then

\[
\eta = 1 - \beta \cos \tau_r \approx 1 - (1 - 2 \gamma^{-1}) (1 - \frac{1}{2} \dot{\tau}_r^2 + \cdots) \approx \frac{1}{2} \gamma^{-1} + \frac{1}{2} \dot{\tau}_r^2.
\]

Recall that the relativistic energy is\(^5\) \(E = \gamma mc^2\), and so

\[
\eta^{-1} = \frac{2E / (mc^2)}{1 + \frac{1}{2} \dot{\tau}_r^2 E / (mc^2)}.
\]

\(^4\)Beware that many authors instead compute a related but physically different quantity, the energy loss from the particle per emission time \(t_i\).

\(^5\)See Equation 32.26 (page 502).
Indeed, this factor is narrowly peaked in time, and its peak value is enhanced when $\mathcal{E} \gg mc^2$. Conversely, suppose we want a certain energy $\mathcal{E}$ but are considering various kinds of charged particle. A low-mass choice, like an electron, will give off far more synchrotron radiation than a heavier choice, like a proton.

Figure 45.3 shows our result: the time course of outward energy flux at a fixed location for the relativistic case $\beta = 0.95$. Most of the energy is indeed received in one burst per orbital period. The burst duration is just 0.004 times the period, justifying our expansion in Equation 45.17. Equivalently, we can think of a snapshot at one instant of time; power delivered is sharply peaked in angle, to an angular range 0.004 times the full circle. It is straightforward to extend our calculation and find that energy emission is also highly confined in the other direction, transverse to the plane of the orbit.\(^6\)

### 45.4 PLUS ULTRA

[[Not ready]]

[[A key property of synchrotron emission is that the radiation should be linearly polarized. Following a 1953 suggestion by Shklovsky that both the optical and radio emission from the Crab Nebula should be linearly polarized, a number of astronomers in the USSR, the Netherlands, and the United States showed that the observed optical radiation from the Crab Nebula was indeed linearly polarized, and in 1957 scientists at the US Naval Research Laboratory succeeded in detecting 7.5 percent linearly polarized radio emission at 3.15 cm (9.5 GHz). [Mayer et al., 1957] – "Star Noise" book]]

\(^6\)You’ll pursue this point in Problem 45.2.
FURTHER READING

Intermediate:

Technical:
[“The theory of radiation from electrons in betatrons was worked out by J. Schwinger in 1945, although not published until 1949 [J. Schwinger, Phys. Rev. 70, 798 (1949)]; an earlier unpublished manuscript by Schwinger has now been transcribed by M. Furman [LBNL-39088/CBP Note-179]. Independently, D. Ivanenko and A. A. Sokolov worked out the spectral and angular properties of the radiation [D. Ivanenko and A. A. Sokolov, Dokl. Akad. Nauk SSSR [Sov. Phys. Dokl.] 59; 1551 (1948)]. For the history of the development of this subject and many further details, with a particularly Russian perspective, see A. A. Sokolov and I. M. Ternov, Synchrotron Radiation (Akademie-Verlag, Berlin; Pergamon Press, Oxford, 1968).” – Schwinger et al., 1998, p. 426]
45.1  Moderately-relativistic case

Equation 45.17 gave an approximate formula for retarded proper time, valid for lab time very close a radiation pulse. For ultrarelativistic particles (velocity very close to c), the pulse is very short and this approximation was adequate. Write a short computer code that finds retarded proper time numerically, substitute into Equation 45.16, and plot the time course of field strength squared for $b\omega/c = 0.8$.

45.2  Polar beaming

Generalize the derivation in Section 45.3 to the case where the observer is tilted out of the xy plane, that is, at position $b\hat{x} + D\hat{K}$, where $\hat{K} = \hat{y}\cos\theta + \hat{z}\sin\theta$. The main text claimed that at high $\beta \approx 1$, the energy flux will mostly be confined to small values of $\theta$, so check that.
CHAPTER 46

The Microwave Polarizer

46.1 FRAMING: JONES TENSOR

Media 1 shows demonstration experiments done with a microwave generator and receiver. We now know how its center-fed antenna emits linearly polarized radiation. The receiver also starts with a similar antenna, which is therefore sensitive to one polarization. Finally, in the video a polarizer is introduced (a planar array of long, thin, parallel copper wires). Interesting Electromagnetic Phenomena ensue. This chapter will discuss them, and introduce the Jones tensor to summarize the polarizer’s effect.

Electromagnetic phenomenon: An array of aligned, linear conductors can act as a polarizing filter, and can even regenerate a missing polarization in an incoming beam.

Physical idea: Reradiation by the anisotropically polarizable array can reduce the intensity of the incoming polarization, and create another that was not initially present.

46.2 IDEALIZATION AS A CONTINUOUS, ANISOTROPIC CONDUCTING SHEET

In the demonstration, the microwave polarizer consisted of parallel wires with spacing smaller than the wavelength. We will therefore approximate it as a thin, planar conducting sheet. Unlike a sheet of ordinary metal foil, however, we will keep the property of anisotropy, imagining that the sheet conducts easily in one direction but not the other. Thus, the 2D charge flux at the surface, \( \vec{j}^{[2D]} \), is related to the field \( \vec{E} \) by a 2D tensor, the surface conductivity:

\[
\vec{j}^{[2D]} = \vec{\kappa}_s \cdot \vec{E}(z = 0), \quad \text{where} \quad \vec{\kappa}_s = \kappa_s \hat{x} \otimes \hat{x}.
\]  

(46.1)

We take the incoming fields far from the source to be a plane wave traveling along \( \hat{z} \), and begin by supposing that it is linearly polarized along the conducting direction:

\[
\vec{E} = \frac{1}{2} \vec{E}_\text{in} e^{-i(\omega t - k x)} + \text{c.c.}, \quad \text{where} \quad k = \omega/c.
\]

We will also simplify by considering a poor conductor, that is, \( \kappa_s \) is small. Then each surface element will have little influence on the others; each just responds to the incoming plane wave \( \vec{E}_\text{in} \) via our ohmic hypothesis. Each surface element responds in phase with the others. Each in turn radiates according to the Green function solution.\(^1\) For example, at a

\(^1\)See Equation 39.5 (page 595).
46.2 Idealization as a continuous, anisotropic conducting sheet

Point along the +\( \hat{z} \) axis we have a total radiation field from all surface elements given by

\[
\vec{A}_{\text{rad}} = \hat{x} \frac{\mu_0}{4\pi} \int d^2r_s \frac{1}{R} \frac{1}{2} e^{-i\omega(t-R/c)} + \text{c.c.}
\]

The integral runs over the whole plane \( \phi = 0 \). Let \( k = \omega/c \).

This kind of integral comes up in many contexts, and it has a surprising feature, so let’s pause to consider it carefully. We switch to plane polar coordinates; the integral over \( \phi \) just gives \( 2\pi \) and we are left with

\[
\int d^2r_s \frac{1}{R} e^{ikR}.
\]

The integrand, \( e^{ikR} \), is a messy function of \( r_s \), but we can do it by change of variable in the integral. At an observation point along the +\( \hat{z} \) axis (\( z > 0 \)), we have \( R^2 = r_s^2 + z^2 \), so \( RdR = r_s dr_s \). Thus, we can change variables to get

\[
\int_0^\infty d^2r_s \frac{1}{R} e^{ikR} = 2\pi \int_z^\infty dR \ e^{ikR}.
\]

That integral is easy! But it’s confusing:

\[
\text{“} = \frac{c}{i\omega} \left[ e^{ikz} - e^{ikL} \right].
\]

To understand that first term, suppose that our plane had a large, but finite, extent \( L \). Then this term would give a contribution to the potential that oscillates as we increase the cutoff \( L \). But consider the magnetic field, a physical quantity:

\[
\vec{\nabla} \times \vec{A} = \frac{\mu_0}{4\pi} \frac{2\pi k_s E_{\text{inc}}}{2ik} e^{-i\omega t} (ik) (-\hat{x}) \times \hat{z} \left( -e^{ikz} + \frac{z}{\sqrt{L^2 + z^2}} e^{ik\sqrt{L^2 + z^2}} \right) + \text{c.c.}
\]

Taking \( L \to \infty \) at fixed \( z \), we see the second term may be dropped:

\[
\vec{B} = -\hat{y} \frac{k_x \mu_0 E_{\text{inc}}}{4} e^{-i(\omega t - kz)} + \text{c.c.}
\]

**Your Turn 46A**

Compute the electric field as usual, obtaining

\[
\vec{E}_{\text{rad}} = -\hat{x} \frac{k_x \mu_0 E_{\text{inc}}}{4} e^{-i\omega t - kz} + \text{c.c.}
\]

Remarkably, the forward scattered field is again a plane wave traveling along \( \hat{z} \), but **180 degrees out of phase with the incoming wave**. The total forward wave is then

\[
\vec{E}_{\text{tot}} = \hat{x} \frac{1}{2} \left( 1 - \frac{1}{2} k_x \mu_0 c \right) e^{-i\omega t - kz} + \text{c.c.}
\]

The transmitted wave has lost some of its amplitude.

Where did that energy go? Its flux decreased by the square of the factor in parentheses, or \( \approx (1 - \frac{1}{2} k_x \mu_0 c) \) (remember that we work only to lowest order in \( k_x \)). You should work out the radiated wave \( \vec{E}_{\text{rad}} \) in the backward (reflected) direction, along \(-\hat{z}\), but clearly its
energy flux will be proportional to $(\kappa_e)^2$, and so cannot fully account for the effect that we found. Instead, we must look for the culprit elsewhere.

A conductor with finite conductance \textit{dissipates} energy as heat. The total loss is

$$\int d^2 r \, \mathbf{E} \cdot \mathbf{j}^{[2D]} ,$$

(46.3)

where the 2D charge flux on the surface, $\mathbf{j}^{[2D]}$, is given by Equation 46.1. The loss per unit area is just the integrand of Equation 46.3.

\textbf{Your Turn 46B}

Add it to the energy flux from Equation 46.2 and compare to the incoming energy flux.

\section*{46.3 Effect on an Arbitrarily Polarized Wave: Jones Tensor}

Equation 46.1 says that the conductivity tensor’s principal directions are $\hat{x}$ (eigenvalue $\kappa_x$) and $\hat{y}$ (eigenvalue 0). Equation 46.2 says that an incoming waves polarized along $\hat{x}$ will excite currents, and hence will be attenuated. However, a wave polarized along $\hat{y}$ will excite no currents and hence will be \textit{unaffected}—as seen in Media 1.

We can succinctly combine those results by saying that the outgoing complex polarization (Jones vector) $\mathbf{E}^{\text{out}}$ is related to the incoming by a linear transformation:

$$\mathbf{E}_{\perp}^{\text{out}} = \mathbf{J} \cdot \mathbf{E}_{\perp}^{\text{in}} ,$$

where $\mathbf{J}$ is the called the \textbf{Jones tensor}; it acts on the complex 2D space of transverse directions:

$$J_{ij} = \begin{bmatrix} 1 - \frac{1}{2} \kappa_x \mu_0 c & \kappa_x \\ \kappa_x & 1 \end{bmatrix} .$$

Because the two eigenvalues are unequal, and the two eigenvectors are real, this matrix suppresses one linear polarization relative to the other, and hence (partially) polarizes incoming light.

\section*{46.4 A Tilted Polarizer Can Regenerate a Missing Polarization}

Finally, suppose that the incoming wave polarization is linear but tilted by 45 deg: $\mathbf{E}_{\text{in}} = E(\hat{x} + \hat{y})/\sqrt{2}$. Now we find the forward wave to be

$$\mathbf{E}_{\text{tot}} = \frac{1}{2} E \left( \hat{x} + \hat{y} \right)/\sqrt{2} - \frac{\kappa_x \mu_0 c \hat{x}}{2 \sqrt{2}} e^{-i \omega (t - z/c^2)} + \text{c.c.}$$
We can re-express the second term in the tilted basis by noting that \( \hat{x} = (\hat{x} + \hat{y})/2 + (\hat{x} - \hat{y})/2 \). The first of these terms destructively interferes with the incoming beam as before. The other one, however, generates a “transmitted” wave with a polarization not present in the incoming wave, another Electromagnetic Phenomenon seen in Media 1.

### 46.5 EXTENSION TO OPTICAL POLARIZERS

Section 52.3.2 (page 726) described how polymer chains can create a nanoscale version of the device in the preceding section. If the incoming light encounters many consecutive, aligned layers of such chains, the Jones matrix will be raised to a high power, the small eigenvalue will shrink to nearly zero, and we will get essentially a projection operator. In Problem 46.3 you’ll explore phenomena involving such a filter.

### PROBLEMS

#### 46.1 Another integral

Another situation of interest involves a plane wave that impinges on a dielectric (non-conducting but polarizable) sheet. We then need an integral of the form

\[
\int_0^{\infty} (2\pi \rho d\rho) (1 - \cos^2 \alpha) e^{ikr}/(4\pi r). \quad \text{Here} \quad r = \sqrt{\rho^2 + (z_\ast)'^2} \quad \text{and} \quad \cos \alpha = \rho/r. \quad \text{Here} \quad k, z_\ast \quad \text{are constants.}
\]

This integral looks like it might be approximately equal to \( \frac{1}{2k} e^{ikz_\ast} \), but it’s good to check. Unfortunately it is probably not an integral you have met in calculus. Fortunately, however, we can simplify it to the point where a computer can help us. Notice that the problem contains two parameters, \( z_\ast \) and \( k \). There is only one dimensionless combination of these parameters; call it \( M = k z_\ast \).

a. Change variables in the integral from \( \rho \) to \( r \). Define dimensionless variable \( u = kr \), and express the assertion to be shown in terms of it. Express it in the form (a certain integral) \( \approx 1 \).

b. Figure out how to get your favorite mathematical software to do this integral numerically. Evaluate it for various values of \( M \) and check our expectation. [One visually appealing way could be to graph the real and imaginary parts of the quantity you found in (a) as functions of \( M \).]

c. Are there some values of \( M \) for which our expectation is more, or less, accurate?

#### 46.2 [Not ready]

#### 46.3 Circular polarizer revisited

A perfect linearly polarizing filter would have Jones tensor with eigenvalues 0 and 1, for example, \( \frac{1}{2} (\hat{x} + \hat{y}) \otimes (\hat{x} + \hat{y}) \). Rederive your results from Problem 52.1 (page 734) more elegantly as consequences of multiplying the corresponding matrix either from the left or the right by another one that represents a quarterwave plate.
CHAPTER 47

Scattering by Free and Bound Charges

47.1 FRAMING: RERADIATION

[Not ready]

Electromagnetic phenomenon: The angular modulation of the cosmic microwave background radiation’s polarization tells us about inhomogeneity of the early Universe.

Physical idea: Light reradiated by a free charge has a characteristic pattern of polarization versus direction.

47.2 WEAK-FIELD LIMIT

When an EM wave encounters a charged particle, we’ve seen that it shakes the particle. Chapter 20 considered the rather fanciful situation of a particle subject to “viscous friction.” The opposite extreme is a free charged particle. For example, in a plasma like the early Universe just prior to recombination, atoms are dissociated into nuclei and electrons, each of which feels an overall potential due to all the others but is not bound to any specific partner.

Let’s investigate the simplest case, with a single free charge $q$, of mass $m$. We will assume that the charge’s motion is always nonrelativistic (and later justify that assumption, in a limit that we will make precise). Write an incident plane wave as

$$E(t, \vec{r}) = \frac{1}{2} \vec{E} e^{-i(\omega t - \vec{k} \cdot \vec{r})} + \text{c.c.}$$

(and the associated $\vec{B}$ field). The charge feels an electric force $\vec{F} = q\vec{E}$. The magnetic force is negligible because $||\vec{E}|| = c||\vec{B}||$ so $||q\vec{v} \times \vec{B}|| \sim q(v/c)||\vec{E}||$. Our assumption of nonrelativistic motion, $v/c \ll 1$, means that we can neglect this part of the force.

Write the charge’s trajectory as $\vec{r}(t) = \vec{r}_0 e^{-i\omega t} + \text{c.c.}$ Then Newton’s law gives the amplitude of the motion as $\vec{p} = -(q\vec{E})/(m\omega^2)$, whose velocity will be $\ll c$ if

$$||q\vec{E}|| \ll m\omega c. \quad \text{condition for nonrelativistic motion} \quad (47.1)$$

So our assumption is justified for weak enough fields. In practice, this condition is nearly always well satisfied.

---

1 Chapter 54 will study plasmas. A situation like this one also holds for some of the electrons in a metal.
2 Chapter 20 studied the longitudinal force, for which the magnetic part was the leading term and so could not be dropped.
3 See Problem 47.2. The approximation may break down in a free electron laser, however.
47.3 SCATTERING CROSS SECTION AND THE THOMSON FORMULA

Our shaking charge gives rise to a time-dependent dipole moment \( \vec{D}_e(t) = q \vec{T}(t) \), so it will radiate at the same frequency. The charge’s motion remains confined to a region of size \( ||\vec{r}|| \). The criterion for the electric-dipole approximation\(^4\) is met by virtue of Equation 47.1:

\[
||\vec{T}|| \omega/c = (q \vec{E})/(m \omega^2 c) \omega \ll 1.
\]

We can therefore use the ED radiation formulas to find the energy flux in any direction.

Chapter 43 gave the energy flux for a time-dependent, linear dipole as

\[
\vec{S} = \vec{r} \frac{\mu_0}{(4 \pi \epsilon)^2} \frac{1}{c} \frac{d^2}{dt^2} \vec{D}_e \sin^2 \theta,
\]

where \( \theta \) is the angle between the dipole moment and the direction of observation \( \vec{r} \). In our case, suppose that the incoming wave is polarized along \( \hat{z} \); then \( \vec{D}_e(t) = \hat{z} \frac{1}{2} \vec{D}_e e^{-iat} + \text{c.c.} \), with \( \vec{D}_e = -q^2 \vec{E} / (m \omega^2) \). The power output per solid angle is then

\[
\frac{d\mathcal{P}}{d\Omega} = \frac{1}{4\pi^2} \frac{1}{\epsilon_0 c^3 m^2} \frac{q^4}{2} ||\vec{E}||^2 \sin^2 \theta.
\]

Remarkably, the angular frequency \( \omega \) drops out of this formula. Note, too, that the incident wave’s direction \( \hat{k} \) is irrelevant, other than that it defines the plane of allowed directions for \( \vec{D}_e \). Finally, note that a free proton is much less effective at scattering than a free electron, due to the \( 1/m^2 \) factor.

Equation 47.2 tells us something about how good our charge is at scattering radiation, but it’s not intrinsic to the charge—it also depends on \( \vec{E} \). To get something intrinsic, we need to normalize by some measure of the incoming wave’s amplitude. How should we do that? The total power transported by a plane wave is infinite, because of its infinite extent in the transverse directions. But most of that extent is irrelevant—bits of the wave that never come near the charge just cruise by without scattering.

The key insight is that the energy flux (power per unit area) is finite. Think about holding a penny in the sunlight. The energy removed from the incoming beam (reflected or absorbed) equals the solar energy flux times the cross-sectional area of the penny, or

\[
\text{cross section} = \text{energy removed from beam}/\text{energy flux incoming}.
\]

Note how the units work out: energy and time cancel, leaving behind \( 1/(1/L^2) \), or area. The infinite transverse extent of the incoming beam is irrelevant, as desired.

We can similarly characterize how good a single electron is at scattering light by forming the same quotient; the amplitude of the incoming beam cancels from numerator and denominator, leaving behind a quantity with units of area, which we will again call “cross section” by analogy to the macroscopic situation. We just need a formula for the denominator:

\[
\langle ||\vec{S}_{\text{in}}|| \rangle = \frac{1}{\mu_0} \langle ||\vec{E} \times \vec{B}|| \rangle = \frac{1}{2} \epsilon_0 c ||\vec{E}||^2.
\]

\(^4\)Section 43.2.3 (page 643).
The cross-section is traditionally denoted \( \sigma \). Extending our original thought experiment, we can subdivide this scattering cross section into bits attributable to scattering into particular angular bins \( d\Omega \), or:

\[
\frac{d\sigma}{d\Omega} = \left\langle \frac{dP}{d\Omega} \right\rangle \left\langle \|\vec{S}_{\text{in}}\| \right\rangle.
\]

This quantity is generically called the \textbf{differential scattering cross-section}.

For the case of scattering from a single electron, in classical electrodynamics, combining the preceding generic formula with Equation 47.2 gives

\[
\frac{d\sigma}{d\Omega} = \left( \frac{1}{4\pi \varepsilon_0 c^2 m} \right) q^2 \sin^2 \theta. \quad \text{Thomson scattering cross-section} \quad (47.3)
\]

\textbf{Your Turn 47A}

Confirm that the constants in parentheses in Equation 47.3 really do combine into a quantity with dimensions of length, and evaluate it for \( q \) and \( m \) appropriate for an electron. This quantity is called the \textbf{classical electron radius}, or \( r_c \).

Often we don’t care about angular dependence; we only want to know how much energy the electron scatters out of the beam. For this, we can integrate the Thomson formula over all directions, using

\[
\int d\varphi d(\cos \theta) \sin^2 \theta = 8\pi/3.
\]

The total scattering cross-section obtained in this way is \( \sigma = (8\pi/3)r_c^2 \), a useful number you should evaluate for electrons.

\textbf{47.4 LIGHT PROPAGATES DIFFUSIVELY IN A STELLAR INTERIOR}

The Sun’s interior is hot. There’s a lot of light in there. And yet, that light takes a long time to make its way to the surface of the Sun. One way to think about this is to imagine the light constantly scattering, changing direction. Although any one electron in this plasma isn’t very effective at scattering light, there are quite a lot of electrons. So the light must take a zigzag path; even though it’s traveling at \( c \) between collisions, nevertheless that path will be much longer than the Sun’s diameter, so traversing it takes a lot of time.

The quantity that characterizes the tortuous light trajectories is a “mean free path.” Dimensional analysis suggests that, to get dimensions of length, we need to form the quantity \( 1/(r_c^2 n_e) \), where \( r_e \) is the volume density of free electrons. The mean free path for light is this quantity times some geometrical constants of order one.\(^5\)

\(^5\)You’ll work out details in Problem 47.3.
47.5 POLARIZED INCOMING LIGHT RETAINS ITS POLARIZATION UPON SCATTERING

Suppose that the incoming light travels along \( \hat{x} \), with polarization along \( \hat{x} \). Then \( \vec{E} \parallel \hat{x} \). The electric far field points along \( \hat{x} - \hat{r} (\hat{r} \cdot \hat{x}) \); that is, it lies in the plane spanned by \( \hat{r} \) and \( \hat{x} \) and (as always) transverse to \( \hat{r} \).

Linearly polarized light always scatters to some kind of linearly polarized light, regardless of the scattering direction (or to nothing if we observe along the direction of polarization, \( \hat{r} \parallel \hat{x} \)).

47.6 UNPOLARIZED INCOMING LIGHT ACQUIRES PARTIAL POLARIZATION

47.6.1 Selective scattering can create polarization

So far, we have been considering an incoming wave that is monochromatic and polarized. Section 24.3.2 (page 388) argued that we can treat unpolarized light as an incoherent superposition of many pure waves. Scattering can impose partial polarization on such light. For example, when viewed at 90° to the original wave’s direction, the scattered light will be 100% linearly polarized: One component of the incoming light shakes electrons longitudinally to that viewing direction, so there is no reradiation in that direction at all. At other scattering angles, the scattered light’s degree of polarization interpolates between that extreme value and 0% for the forward and backward directions.

47.6.2 Polarization of the cosmic microwave background as a reporter for early-Universe conditions

[[Please read the posted pages from Dodelson’s book about how we can use these observations to learn about the early Universe from the faint polarization pattern in the cosmic microwave background radiation. – Dodelson & Schmidt, 2021, pp310–319.]]

[[Not ready]]

47.7 THE CASE OF BOUND CHARGES

47.7.1 Rayleigh scattering cross section

Next suppose that the charge is bound, for example, to an atomic nucleus. The simplest classical model we can make of that situation is to suppose that the charge gets a linear restoring force with isotropic spring constant \( k \). As usual with harmonic oscillators, it is convenient to introduce \( \omega_0 = \sqrt{k/m} \). Then Newton’s law becomes

\[
-m \omega^2 \vec{r} = -\omega_0^2 \vec{r} + q \vec{E}, \quad \text{so} \quad \vec{r} = \frac{q \vec{E}}{m (\omega_0^2 - \omega^2)},
\]
Substituting this expression into earlier results then gives the Thomson expression for differential and total cross-sections, each now multiplied by $(1 - (\omega_0/\omega)^2)^2$. These results are called the **Rayleigh cross-section** formulas. In particular, the differential cross section has the same polarization behavior as what we already observed for free charges.

Two limiting cases are noteworthy: At high frequency $\omega \gg \omega_0$, our results reduce to the Thomson formulas. In this regime, the fact that the charge is bound is immaterial to its response. In the opposite limit, we get the Thomson formulas multiplied by $(\omega/\omega_0)^4$. The cross-section is now strongly frequency dependent.

### 47.7.2 The blue, polarized sky

Earth’s upper atmosphere consists of polarizable objects (molecules) that are much smaller than the wavelength of visible light, at low enough density that we may neglect their mutual interactions and treat them as independently scattering sunlight to our eyes. They are also randomly placed in space, which eliminates any coherent effects. In such a situation, the fact that there are many such molecules just amplifies the scattering without changing its character. Indeed, we know that:

- The scattered light is polarized in a way that depends on the direction of the line of sight relative to the incoming beam (Section 47.6.1).
- The scattered light is bluer than the incoming sunlight, because higher frequencies scatter more strongly (the $(1 - (\omega_0/\omega)^2)^2$ factor).
- At sunset, we observe sunlight through a thicker layer of air than at noon, and direct (unscattered) light is redder (more depleted of high frequencies) than at noon.

### 47.7.3 Blue, polarized scattering from colloidal suspensions

It’s easy to send a beam of white light from a projector into a dilute suspension of nonfat milk. Milk contains dissolved substances such as lactose and salts, but that just gives a solution that’s homogeneous on the scale of the wavelength of light (it alters the refractive index), and so is irrelevant for scattering. Nonfat milk is also a colloidal suspension of protein micelles, which

- are well separated compared to light wavelength;
- are themselves much smaller than wavelength of light (nanometer scale);
- move randomly and independently; and
- have permittivity different from that of the surrounding water.

Thus, the system is similar in some relevant respects to that of sunlight on the upper atmosphere. And indeed, the light scattered at 90 deg is strongly polarized and more blue than the incoming light, while the light transmitted has been depleted of blue and is visibly redder than the incoming light.

---

6See Media 19.

7Whole milk additionally contains fat globules, which are larger than protein micelles and would complicate the discussion.
47.3’ The transition to Compton scattering

At high frequencies, the quantum character of light starts to matter. Dimensional analysis gives us a clue: We can form another length scale, the Compton wavelength $\lambda_C = \frac{2\pi \hbar}{mc}$, by using Planck’s constant. If the incoming light’s wavelength is shorter than this, then we start to get billiard-ball collisions of electrons and single photons, the Thomson formula is no longer valid, and weirder still, the outgoing photon won’t have the same frequency as the incoming one (Compton scattering, Section 31.4.1, page 481).
47.1 Justify an approximation
A red laser gives a 100 mW beam that is approximately a plane wave with cross-sectional area 1 mm².

a. Find the electric field strength in this beam.
b. Estimate the fractional deformation of a hydrogen atom placed in this beam, due to the electric field. Is it reasonable to make the approximation of working to first order in this deformation when we study polarizability? [Hint: Let \(a\) denote the atom radius. Model the charge cloud as a sphere of this radius, filled with uniform charge density. Suppose that the charge cloud remains spherical but displaces relative to the nucleus by distance \(b\). Recall that the restoring force is then the same as if the part of the cloud with radius less than \(b\) were all concentrated at the center of the sphere; the rest of the charge cancels (Problem 1.5 (page 25)).]

47.2 Justify another approximation
Suppose that the beam in Problem 47.1 encounters a free electron. The electron responds by oscillating. Justify our use of the nonrelativistic approximation for that motion.

47.3 Diffusion of light
Idealize the Sun as a highly ionized plasma with average free electron density \(n_e \approx 10^{24} \text{ cm}^{-3}\).

a. Dimensional analysis suggests that the mean free path for electromagnetic radiation in the Sun is \(\approx (n_e \sigma)^{-1}\), where \(\sigma\) is the scattering cross-section. Evaluate the MFP as a function of wavelength by using the Thomson formula.
b. Over lengths longer than the MFP, radiation takes a random-walk path out of the Sun. Estimate the time required for EM radiation to diffuse from the core to the outside, a distance of \(7 \cdot 10^8 \text{ m}\). [Hint: Take the time between collisions to be \(\approx \text{MFP}/c\).]
Light in Linear Media

Michael Faraday’s sketch of his setup for the discovery of magneto-optical rotation. A piece of glass (right) is placed across the pole tips of electromagnets. When (i) polarized light passed through the glass, (ii) the light traveled parallel to the magnetic field, and (iii) the magnets were turned on, then the light’s polarization was affected. [From the margins of Faraday’s diary, 1845.]
49.1 Framing: Evanescent Wave

We now return to the study of nonconducting, polarizable media, in greater detail than in Chapter 6. Thus, charges are not free to travel throughout the material; however, individual molecules can deform slightly.

This chapter will consider an approximation in which:

- The medium consists of polarizable objects (or permanently polarized, unoriented objects that can become oriented by an external field). We will acknowledge only the dipole fields created by those objects.
- All forms of energy dissipation may be neglected. Thus, we exclude ohmic materials (conductors).\(^1\)
- External fields vary over length scales much longer than the spacing between the polarizable constituents, which can therefore be treated as a continuous density of dipole moment.\(^2\)

All formulas in this chapter are understood to be subject to the limitations of these approximations, whose domain of validity we won’t explore.

Electromagnetic phenomenon: How can light “tunnel” into a zone forbidden by geometrical optics?

Physical idea: An evanescent wave extends into such zones and can transmit energy to inhomogeneities, for example, a fluorophore.

49.2 Polarizable Media

49.2.1 The electric displacement combines \(\vec{E}\) with a contribution from induced electric dipole moment

First we review the discussion of dielectric materials from Chapter 6. Figure 49.1 recalls the argument for why a bound charge density can arise with

\[
\rho_b = -\nabla \cdot \vec{P}. \tag{49.1}
\]

Now imagine cutting the sample along the \(yz\) plane. The newly exposed surface of the right side will be an interface between medium and vacuum with a bound areal charge

---

\(^1\) Problem 8.5 studied the generalization to dissipative media.

\(^2\) We average the effects of finite-size molecules over a length scale smaller than the one of interest, but much bigger than the molecular spacing. The quantities \(\vec{E}, \vec{P}, \vec{B},\) and \(\vec{M}\) below are all such averages.
49.2 Polarizable Media

Figure 49.1: [Duplicate of Figure 6.3.] Origin of bound charge density. A collection of electrically polarizable “molecules” with static but nonuniform electric dipole moment density \( \vec{P} \) (magnitude increasing from left to right). Net bound charge appears that is minus the divergence of the polarization density, which in this case is \(-\delta \vec{P}_x / \delta x < 0\).

\[
\sigma_b = \vec{n} \cdot \vec{P}, \tag{6.4, page 80}
\]

where \( \vec{n} \) is the perpendicular unit vector pointing out into the vacuum.

Even though bound charges are confined to specific regions, nevertheless they can move slightly, and in fact they must move if the polarization is time-dependent. That localized motion will in turn give rise to a bound charge flux \( \vec{j}_{b,P} \). To make these ideas more precise, suppose that \( \vec{P} \) is initially zero, then switches on to the form shown in Figure 49.1. Creation of the internal layer of negative bound charge requires net flow of charge to the right. Mathematically, the continuity equation says that \( \partial \rho_b / \partial t = -\vec{V} \cdot (\vec{j}_{b,P} - \delta \vec{P} / \delta t) \). Substituting that result into Equation 49.1 gives \( 0 = \vec{V} \cdot (\vec{j}_{b,P} - \delta \vec{P} / \delta t) \), which will be satisfied if

\[
\vec{j}_{b,P} = \partial \vec{P} / \partial t. \quad \text{electric dipole contribution to bound charge flux} \tag{49.2}
\]

Equation 49.7 below will cast the electric Gauss law in a form that only explicitly mentions free charge. To this end, define the electric displacement as

\[
\vec{D} = \varepsilon_0 \vec{E} + \vec{P}. \tag{49.3}
\]

(We’ll just call it “the \( \vec{D} \) field.”)

Section 49.2.1’ (page 701) introduces dissipation.

49.2.2 The \( \vec{H} \) field combines \( \vec{B} \) with a contribution from induced magnetic dipole moment

Let \( \vec{M} \) denote the net magnetic dipole moment density created by bound currents (the motions of bound charges in individual polarizable objects). If \( \vec{M} \) is spatially nonuniform, it can give rise to a second contribution to the bound charge flux (in addition to

---

\(^3\)Equation 6.8 (page 81) already introduced the \( \vec{D} \) field.
Figure 49.2: A second source of bound charge flux. A collection of magnetically polarizable "molecules" is represented as current loops parallel to the xz plane (black rings). Their magnetic dipole moment density has magnitude increasing from left to right. Net bound charge flux appears that is proportional to the curl of magnetic moment density, in this case equal to $2(\partial\vec{M}_y/\partial x)$, which is directed out of the page.

Equation 49.2). The general formula

$$\vec{j}_{b,M} = \vec{\nabla} \times \vec{M}$$  \hspace{1cm} \text{(magnetic dipole contribution to bound charge flux)} \hspace{1cm} (49.4)$$

is rotationally invariant and agrees with the special case shown in Figure 49.2. Now imagine cutting the sample along the xz plane. The newly exposed surface of the right side will be an interface between medium and vacuum with a bound surface charge flux

$$\vec{j}_{b,[2D]} = \vec{M} \times \hat{n},$$  \hspace{1cm} \text{(49.5)}$$

where $\hat{n}$ is the outward-directed perpendicular to the interface.

Equation 49.8 below will cast Ampère's law in a form that only explicitly mentions free charge. To this end, define the magnetic field intensity as

$$\vec{H} = \mu_0^{-1} \vec{B} - \vec{M},$$  \hspace{1cm} \text{(49.6)}$$

(We'll just call it "the $\vec{H}$ field.")

49.2.3 The new fields summarize the effects of bound charge

Chapter 6 began a program of eliminating explicit mention of the bound charges, expressing the equations of electrostatics solely in terms of the remaining "free" (not bound) charges. We can now extend that program to nonstatic situations by defining free charge flux as well:

$$\rho_f = \rho_q - \rho_b, \quad \vec{j}_f = \vec{j} - \vec{j}_b = \vec{j} - \vec{j}_{b,p} - \vec{j}_{b,M}.$$  \hspace{1cm} (49.7)$$

**Your Turn 49A**

Use these definitions, and other preceding formulas, to show that free and bound charges each obey separate continuity equations.

Charges that macroscopically violate neutrality are considered free, for example, those placed on a capacitor. Currents carried by charges that can move macroscopic lengths are also considered free, including ionic currents in solution.
Your Turn 49B

Using Equations 49.1, 49.2, and 49.4, show that

\[ \vec{\nabla} \cdot \vec{D} = \rho_f \quad \text{electric Gauss (49.7)} \]
\[ \vec{\nabla} \times \vec{H} - \frac{\partial \vec{D}}{\partial t} = \vec{j}_f. \quad \text{Ampère (49.8)} \]

Equation 49.7 extends the validity of Equation 6.9 (page 81) to situations where the polarization is nonuniform. (The magnetic Gauss law and the Faraday law are unmodified, because they do not involve charges or currents.)

49.2.4 Boundary conditions

Chapter 15 showed that the perpendicular component of the \( \vec{B} \) field must be continuous across a boundary between media:

\[ \Delta \vec{B}_\perp = 0, \quad \text{from magnetic Gauss law.} \quad [15.22, \text{page 232}] \]

Also, Chapter 6 showed that

\[ \Delta \vec{E}_\parallel = 0, \quad \text{from Faraday law.} \quad [6.23, \text{page 90}] \]

But at a dielectric/vacuum interface, even if there is no free surface charge or current, bound charges and currents can create jumps in the other field components. Letting \( \hat{n} \) point outward from medium 1 (toward the vacuum)

\[ \hat{n} \cdot (\vec{E}_{\parallel}^{[\text{vac}]} - \vec{E}_{\parallel}^{[1]}) = \sigma_b/\varepsilon_0, \quad \text{from electric Gauss law.} \quad [6.21, \text{page 90}] \]
\[ \Delta \vec{B}_\parallel = \mu_0 \beta_b^{[2D]} \times \hat{n}, \quad \text{from Ampère law.} \quad [15.23, \text{page 233}] \]

At an interface between two media, or one medium and vacuum, Equation 6.21 can be written compactly in terms of \( \vec{D} \) and Equation 15.23 can be written in terms of \( \vec{H} \):

Your Turn 49C

Allow also for free areal charge density and flux. That is, using a strategy similar to the one that yielded Equations 49.7–49.8, show that

\[ \Delta \vec{D}_\perp = \sigma_f; \quad \Delta \vec{H}_\parallel = \vec{j}_f^{[2D]} \times \hat{n}. \]

Here \( \Delta \vec{D}_\perp = (\vec{D}_{\perp}^{[\text{2}]} - \vec{D}_{\perp}^{[1]}) \cdot \hat{n} \), where \( \hat{n} \) is the unit perpendicular vector pointing from medium 1 to medium 2 and similarly, \( \Delta \vec{H}_\parallel = \vec{H}_{\parallel}^{[2]} - \vec{H}_{\parallel}^{[1]} \).

These results are particularly useful when an interface has zero free areal charge density and zero free surface current.

Section 49.2 (page 701) mentions more realistic ways to think about bound charge and current.
49.3 LINEAR REGIME

We are partway to our goal of eliminating explicit mention of bound charges and currents from the Maxwell equations, but Equations 49.7 and 49.8 have doubled the unknown fields, adding $\vec{D}$ and $\vec{H}$ to $\vec{E}$ and $\vec{B}$. It is true that the new quantities are determined by the old ones, but in a way that buries the bound charges and currents without actually eliminating them (Equations 49.3 and 49.6). We now introduce a further level of approximation that, when justified, will finish our job in a simple way.

49.3.1 In linear media, electric polarization effectively modifies $\varepsilon_0$.

Many dielectric media are approximately linear. That is, induced dipole moment density $\vec{P}$ is a linear function of $\vec{E}$, described by a dielectric susceptibility tensor $\chi_\varepsilon$ via the response function

$$\vec{P} = \varepsilon_0 \chi_\varepsilon \cdot \vec{E}.$$ 

The dielectric susceptibility describes how much $\vec{P}$ (deformation times charge per volume) arises per applied electric field (force per charge). That is, $\chi_\varepsilon$ is essentially a spring tensor, multiplied by density and charge squared. Like any spring tensor, it is symmetric. This chapter will restrict to the situation of a medium that is isotropic (rotationally invariant at each point). Thus, $\chi_\varepsilon$ is a scalar constant $\chi$ times the identity tensor. A medium can be isotropic if its constituent polarizable objects are themselves spherical (like helium atoms), or if they are arranged with random orientations (like water molecules in liquid or vapor phase). Define the permittivity $\varepsilon = \varepsilon_0(1 + \chi)$. Then

$$\vec{D} = \varepsilon \vec{E}.$$ 

consstitutive relation for uniform, linear, isotropic, lossless, nonchiral dielectric  

[6.11, page 81]

The constitutive relation takes us partway to our goal, because the permittivity (or equivalently, the susceptibility) is a material property, characteristic of the dielectric but independent of applied field. We may look it up in a table, then use it to eliminate $\vec{D}$ in favor of $\vec{E}$ in Equation 49.7.

More general forms of the constitutive relation include dissipation (complex $\varepsilon$), anisotropy ($\varepsilon$ with tensor structure), nonlinearity, and chirality.

---

4 For nonlinear media, see Section 49.3’ (page 702).
5 Susceptibility is in general a tensor, because a medium’s response to applied fields need not be isotropic (page 200). This chapter restricts to the isotropic case, but Chapter 52 will relax that assumption.
6 See Your Turn 13E (page 202).
7 Problem 14.2 (page 220) showed that a rotationally invariant, symmetric, rank-2 tensor must be a constant times the identity.
8 Section 6.5.1 (page 79) introduced this quantity.
9 Chapter 52 studies anisotropy. Chapter 51 studies chirality. Chapter 54 discusses the meaning of complex response functions.
49.3.2 In linear media, magnetization effectively modifies $\mu_0$

In many media, induced magnetic dipole moment density $\vec{M}$ is also approximately linear in the applied field (or zero). That is, each of $\vec{B}$, $\vec{H}$, and $\vec{M}$ is linearly related to any of the others. We will find it convenient to quantify the relation (the response function) by introducing another tensor property of the medium:

$$\vec{M} = \mu_0^{-1}\chi_m \cdot \vec{B}.$$ 

Again, this chapter will restrict to isotropic media. Thus, $\chi_m$ will be a scalar constant $\chi_m$ times the identity tensor. Define the permeability $\mu = \mu_0 / (1 - \chi_m)$. Then

$$\vec{H} = \mu^{-1}\vec{B}.$$  

**constitutive relation** for uniform, linear, isotropic, lossless, nonchiral magnetic material

Equation (49.9)

Later we’ll study more general forms of the constitutive relations including dissipation (complex $\mu$), anisotropy, and chirality.

49.3.3 Maxwell plus constitutive relations form a closed system

Equations 49.7 and 49.8 are general. For the special case of linear media, they can be combined with Equations 6.11 and 49.9, and the boundary conditions, to form a system that can be solved to give all fields in terms of free charges and currents.

That is, we can forget about the medium if it’s linear; the electric Gauss law (Equation 49.7) retains its vacuum form but with a modified value of the permittivity that we can look up in a table. The Ampère law (Equation 49.8) also retains its vacuum form, but with a modified value of the permeability. Only the free charge density and flux enter these equations. You also found in Your Turn 49C that the same is true for the boundary conditions.

For example,

*In a bulk, linear, isotropic medium, there will be the same wave solutions as in vacuum, except that the phase velocity is $(\varepsilon\mu)^{-1/2}$ instead of $c$. In particular, there are still two transverse polarizations.*

For example, nonzero dielectric susceptibility ($\varepsilon > \varepsilon_0$) leads to a slowdown, that is, to a value of the refractive index that is larger than the vacuum value of 1.

49.3.4 Macroscopic physical realizations

Consider a medium consisting of (or containing):

- A jumble of long, thin, straight strands of wire, oriented randomly. This medium is electrically polarizable and isotropic.

---

10 Some authors instead define a different quantity $\chi_m$, the **magnetic susceptibility**, via $\vec{M} = \chi_m \vec{H}$. The two descriptions are equivalent: The relation between the two quantities is $\chi_m = \chi_m / (1 + \chi_m)$.

11 Exotic “metamaterials” exist with $\varepsilon$ and/or $\mu$ that are not positive in certain frequency ranges, requiring special interpretation.
Figure 49.3: **Permittivity of a medium** (fused silica glass) as a function of frequency. Although this material has been engineered to minimize variation in the visible range, it shows complex behavior near resonances. [Data from Philipp, 1998.]

- A jumble of circular rings of conductor, oriented randomly. This medium is magnetically polarizable and isotropic.

[[Not ready]]

### 49.3.5 Remarks and further examples

The preceding section imagined macroscopic polarizable objects, which could be relevant for radio or microwave propagation. But individual molecules are also polarizable. The details involve quantum mechanics, which lies outside the scope of this book; nevertheless, for many purposes those details can be incorporated into phenomenological values of the susceptibilities.

Bound charges and currents cannot really respond instantaneously to applied fields, as we implicitly assumed in our constitutive relations (Equations 6.11 and 49.9). However, in a time-independent material we can at least say that polarization is the convolution of the applied field with a time-dependent response function. Equivalently, in the linear regime a disturbance with a single frequency engenders a polarization response with that same frequency, but with a frequency-dependent susceptibility. Hence also $\epsilon$ and $\mu$ will in general be frequency dependent, leading to **dispersion**, that is, nonconstant wave velocity $(\epsilon\mu)^{-1/2}$ (Figure 49.3). Dispersion, in turn, implies that refraction of light will be frequency dependent. Thus, a glass lens will have slightly different focal lengths for each color, making it unable to simultaneously focus them all to a common plane.\(^1\)

Here are some examples of dispersive media:

- A conducting ring with a gap acts as an inductor in series with a capacitance (the gap), that is, as an RC circuit. When it is externally driven, its response will display resonance. A jumble of such split rings will therefore have a frequency-dependent permittivity, offering a macroscopic model for behavior like that seen in Figure 49.3.\(^2\)

\(^1\)This is the origin of chromatic aberration (Section 21.3.7, page 314).

\(^2\)See Further Reading.
49.4 "Total" Internal Reflection and its Evanescent Fields

49.4.1 At high angle of incidence, the fields emerging from a dielectric have complex wavenumber

Let’s revisit a phenomenon that we first examined in the geometrical-optics approximation (Figure 21.3, page 310), this time keeping track of polarization. We are seeking a solution to the Maxwell equations for a plane wave that travels through an isotropic dielectric medium with angular frequency $\omega$, encounters a planar interface at the $xy$ plane, and possibly exits into vacuum or air. Section 21.3.8 (page 314) claimed that no such transmission is possible in some cases, and called that situation total internal reflection. Let’s look more closely, and this time not make the geometrical-optics approximation.

Figure 49.4 shows the boundary of the medium as the $xy$ plane, with vacuum in the region $z > 0$. Axial symmetry about the $z$ axis means that we lose no generality by assuming the incident light has wavevector in the $yz$ plane: $\mathbf{k} = k(\sin \theta \hat{y} + \cos \theta \hat{z})$. We will, however, simplify by considering only one of its two possible polarizations: We’ll take $\mathbf{E}$ to point along $\hat{x}$ (“s wave” polarization). This choice is indeed perpendicular to $\mathbf{k}$, and it is also parallel to the boundary. The Maxwell equations then imply the dispersion relation $k^2 = \frac{\omega^2}{\epsilon_0} \mu_0$.

There is another possible plane wave that also satisfies the dispersion relation for the given $\omega$, however: $\mathbf{k} = k(\sin \theta \hat{y} - \cos \theta \hat{z})$ with the same $k$, $\theta$, and polarization direction. Indeed, we expect that light hitting a boundary is generally at least partly reflected, returning to the medium with the same angle to the perpendicular as the incoming light, so we expect that such a wave will be needed for a complete solution. All told, then, we

- $\epsilon \approx 81 \epsilon_0$ for water at zero frequency; it’s highly polarizable. But the alignment of permanent dipoles in water is slow compared to visible-light frequencies; in this range, $\epsilon$ is only $\approx (4/3)^2 \epsilon_0$.
- Section 54.3.2 will discuss dispersion in a cold plasma, where $\epsilon = \epsilon_0 \left(1 - \left(\frac{\omega_p}{\omega}\right)^2\right)$.

Section 49.3 has advanced our program of summarizing the generation of bound charges and their net motion with a few parameters that characterize a medium. Section 49.3' (page 702) will discuss more general response functions.

Figure 49.4: **Refraction of light** upon exit from a material. For illustration, the text studies the case in which the incoming light is polarized in the $\hat{x}$ direction (perpendicular to the page, the “s wave” or “s polarization”).
will consider a trial solution that superposes incident and reflected plane waves:

$$
\vec{E} = \frac{1}{2} \hat{x} \left[ E_i e^{i(\omega t + \vec{k} \cdot \vec{r})} + E_k e^{i(\omega t + \vec{k} \cdot \vec{r})} \right] + \text{c.c.} \quad \text{for } z < 0. \quad (49.10)
$$

Although Equation 49.10 is a solution to the Maxwell equations, it is only valid inside the medium \((z < 0)\). It’s also not fully specified: Although we were given the incoming wave \((\omega, \vec{\theta}, \text{and } E_i)\), we do not yet know the reflected amplitude \(E_k\). To address both of those shortcomings, we must now try to extend Equation 49.10 by proposing a trial solution for \(z > 0\), and show that we can get a global solution of Maxwell.

Time-translation invariance and linearity imply that the trial solution Equation 49.10 must have the same frequency \(\omega\) as the incident wave. The exterior solution must not only solve the Maxwell equations; it must also tie on to the existing solution by satisfying the appropriate boundary conditions. In particular, at the boundary its dependence on \(x\) and \(y\) must match that of Equation 49.10, so we try a transmitted wave of the form

$$
\vec{E} = \frac{1}{2} \hat{x} \hat{E}_t e^{i(\omega t + \vec{k}^\prime \cdot \vec{r})} + \text{c.c.} \quad \text{for } z > 0, \quad (49.11)
$$

where \(\vec{k}^\prime\) has component \(\hat{y} k\) parallel to the surface. The appropriate dispersion relation is now \(||\vec{k}^\prime||^2 = \mu_0 \varepsilon_0 \omega^2\). Since \(k^2 = \mu_0 \varepsilon_0 \omega^2\), we must then have

$$
(k^\prime)^2 + \varepsilon_0 \mu_0 \omega^2 \sin^2 \theta = \varepsilon_0 \mu_0 \omega^2. \quad (49.12)
$$

If \(1 - (\varepsilon / \varepsilon_0) \sin^2 \theta > 0\), then this equation has a solution for \(k^\prime\), and it gives the usual law of refraction (Equation 21.1, page 310). In the opposite case, however, there is no real solution for \(k^\prime\).

We may now be tempted to say that the exterior fields are just zero, because no acceptable plane wave solution exists, but that would not satisfy the required boundary conditions. Instead, let’s take Equation 49.12 seriously, define \(\alpha = \sqrt{k^\prime / \varepsilon_0} / \varepsilon\), and investigate the trial solution Equation 49.11 with \(\alpha\) real. Although it’s not a plane wave, \(^{14}\) you can readily verify that it does solve the Maxwell equations if

$$
\alpha = (\omega / c) \left(1 - (\varepsilon / \varepsilon_0) \sin^2 \theta \right)^{1/2}. \quad (49.13)
$$

We must choose the positive square root, because otherwise the fields would blow up at \(z \to \infty\).

We can now impose the boundary conditions and try to satisfy them by appropriate choices of the unknowns\(^{15}\) \(E_k\) and \(E_t\). Because we have assumed that the electric fields are all parallel to the interface, we need not consider the \(\Delta D_\perp\) condition; all we need is that \(\Delta E_\parallel = 0\) (Equation 6.23, page 90):

$$
0 = E_t + E_R - E_T. \quad (49.14)
$$

\(^{14}\)Section 18.7.5 (page 280) mentioned the use of complex wavevectors when there is dissipative loss; here there is no dissipation.

\(^{15}\)In the case of real \(k^\prime\), this procedure leads to the Fresnel formulas governing reflection and refraction.
Next, because we assumed \( \mu = \mu_0 \) the remaining boundary conditions just say \( \Delta \vec{B} = 0 \). Using the Faraday law to express \( \vec{B} \) in terms of \( \vec{E} \) yields
\[
\vec{B} = \vec{k} \times \vec{x} E_1 + \vec{k} \times \vec{x} E_R - \vec{k}' \times \vec{x} E_T.
\]
The \( \hat{z} \) component of this equation is redundant with Equation 49.14. The \( \hat{y} \) component with Equation 49.14 yields
\[
0 = (\vec{E}_1 - \vec{E}_R)k \cos \theta - (\vec{E}_1 + \vec{E}_R)(i \alpha)
\]
so
\[
\vec{E}_R = \frac{k \cos \theta - i \alpha}{k \cos \theta + i \alpha} \vec{E}_1, \quad \text{s wave TIR} \tag{49.16}
\]
where \( \alpha \) is the real quantity Equation 49.13.

We have succeeded in finding a global solution to Maxwell and its boundary conditions. Note that:

- Despite the name “total” internal reflection, the fields are nonzero outside the medium, because \( E_\perp \neq 0 \).
- However, in this region the fields fall off with distance as \( e^{-\alpha z} \). Such fields are called evanescent.
- Equation 49.16 implies that \( |E_R|^2 = |E_i|^2 \): The reflected wave carries the same energy flux as the incident wave. Section 21.3.1 called \( E_R/E_i \) the reflection factor \( R \).

### 49.4.2 Evanescent fields carry no energy to infinity

The result just obtained raises several issues: What about the energy of the fields \( \vec{E}_T \) and \( \vec{B}_T \), which fall exponentially with \( z \) but which are nevertheless nonzero? Apparently energy leaks out of the interface and hovers near to it, but can electromagnetic energy actually do that? In vacuum, or in a homogeneous medium, the energy of a plane wave is always moving at speed \( c \) or \( 1/\sqrt{\varepsilon_0 \mu_0} \), respectively, as seen from its nonzero flux. In short, how “total” is total internal reflection?

To answer, we now compute the Poynting vector \( \vec{S}_T \) of the transmitted wave:\(^{17}\)
\[
\vec{S}_T = \vec{E}_T \times \vec{B}_T / \mu_0 = \frac{1}{4 \mu_0} \left[ \hat{x} E_T e^{(-i \omega t + \vec{k}_T \cdot \vec{r})} + \text{c.c.} \right] e^{-\alpha z} \times \left[ \omega^{-1} \vec{k}_T \times \hat{x} E_T e^{(-i \omega t + \vec{k}_T \cdot \vec{r})} + \text{c.c.} \right] e^{-\alpha z}.
\]
Next, examine the \( \hat{z} \) component of energy flux. It will be enough to find only its time average:\(^{18}\)
\[
\langle \vec{S}_T \rangle \cdot \hat{z} = \frac{1}{4 \mu_0} e^{-2\alpha z} \left[ \hat{x} E_T \times (\vec{k}_T \times \hat{x} E_T^* \text{c.c.}) + \text{c.c.} \right] \cdot \hat{z}.
\]
The terms in brackets are
\[
[\vec{k}_T \times (\hat{x} E_T \cdot \hat{x} E_T^*) - \hat{x} E_T^* (\vec{k}_T \times \hat{x} E_T) + \text{c.c.}] \cdot \hat{z}.
\]

\(^{16}\)See Equations 15.22 (page 232) and 15.23. In nonconducting media, there are no free surface charges nor charge fluxes.

\(^{17}\)See Equation 35.12 (page 561).

\(^{18}\)See the method in the Example on page 576.
The first term and its complex conjugate give $|\vec{E}_r|^2 k_L^2 + \text{c.c.}$, which is zero. The second term is zero because we assumed s wave polarization. Altogether, $\langle \vec{S} \rangle_2 = 0$.

A similar result can be obtained for the other incoming polarization (called “p wave”).\footnote{You’ll explore the p wave case in Problem 49.5.} Evanescent fields carry no energy from the surface to infinity, consistent with our earlier result that the reflected wave has the same amplitude as the incident wave.

### 49.4.3 Evanescent fields can nevertheless excite fluorescence

For large enough angle of incidence ($\varepsilon \sin^2 \vartheta > \varepsilon_0$), we have found that although there are fields, and hence field energy, in the “forbidden” region, nevertheless:

- Those fields fall off exponentially with $z$; and
- No energy is transported perpendicular to the surface.

However, there is an important caveat to the second conclusion. Our whole analysis has rested on the assumption that two homogeneous dielectric materials are joined at a planar interface. Point impurities in the second medium can actually extract energy from the evanescent wave, for example, via quantum excitation. This observation led to a detailed experimental confirmation of the rather abstract formulas we have derived.

The preceding section studied an interface between a medium and vacuum. But we could equally have studied an interface between two media; a plane wave originating in the medium with larger refractive index $n_1$ will be totally internally reflected for angle of incidence greater than a critical value $\vartheta_c$. Because the speed of light in a medium is $c/n = (\varepsilon \mu)^{-1/2}$, this angle is

$$\vartheta_c = \sin^{-1}(n_2/n_1).$$

C. Carniglia and coauthors fabricated a dielectric slab with an interior layer that contained fluorescent molecules and was located at a precise distance $z = 45 \text{ nm}$ below the surface (Figure 49.5a). They immersed this layered structure in a fluid with refractive index $n_2$ that was smaller by a factor of 1.10. Without the fluorophores, light shone through the fluid would therefore have been totally internally reflected when its angle of incidence exceeded $\vartheta_c = \sin^{-1}(1/1.10)$.

Although $z$ was fixed, the experimenters varied the angle of incidence while observing the fluorescence intensity. To predict its form, they used Equations 49.14 and 49.15 to find the transmitted amplitude $E_T$, then substituted into Equation 49.11. The “Golden Rule” from quantum mechanics) then predicts\footnote{Section 55.5.2 (page 777) will discuss this step in more detail.} that the probability per unit time for fluorophore excitation should be a constant times the maximum value over time of $|\vec{E}_r|^2$. Figure 49.5b shows this prediction along with the experimental data.\footnote{You’ll work out details in Problem 49.3 (page 703).}

### 49.4.4 Total internal reflection fluorescence microscopy

---

\footnotesize

1. You’ll explore the p wave case in Problem 49.5.
2. Section 55.5.2 (page 777) will discuss this step in more detail.
3. You’ll work out details in Problem 49.3 (page 703).
49.4  "Total" Internal Reflection and its Evanescent Fields

Figure 49.5: **Fluorescence induced by evanescent fields.** (a) [Schematics.] Fabrication of a thin film. A layer one molecule thick of the fatty acid $\text{C}_{20}\text{H}_{40}\text{O}_2$ was spread at an air-water interface. A glass slide was repeatedly dipped into the water. The first time the slide is withdrawn (left), the hydrophilic parts of each molecule in the surface layer adhere to the glass and one monomolecular layer is deposited. By another immersion of the now hydrophobic slide (center) a second fatty-acid layer adheres, and by withdrawal (right) a third layer is deposited. The procedure can be repeated again and again to build up a multilayer system. Two special layers containing fluorophore molecules at low areal density (asterisk) were deposited, followed by 16 pure layers, resulting in the fluorophores' being buried at a fixed distance from the surface. (b) [Experimental data and fit.] Dots: the film fabricated as in (a) was immersed in a high-index solution and illuminated by 476 nm light in s wave polarization, and its fluorescence intensity was recorded at various angles of incidence, in order to probe the probability of the fluorophores being excited 45 nm below the surface. The solid curve above the critical angle is the prediction from Equations 49.14 and 49.15 (see Problem 49.3). Below the critical angle, a similar formula was used appropriate for the regime of partial internal reflection. [(a) After Drexhage, 1970. (b) Data from Carniglia et al., 1972.]

![Figure 49.5](image)

Figure 49.6: **One form of total internal reflection fluorescence (TIRF) microscopy.** [[Light enters from lower left, bends upon entering the objective lens, and illuminates an interface between the glass coverslip and a specimen in water with an angle of incidence that is greater than the critical angle. (The oil layer matches the index of the glass, so there is little reflection or refraction at the oil-glass interfaces.) The intensity of illumination decreases exponentially with distance from the interface, with a decay length $d$ that depends on the angle of incidence. The white and gray spheres represent fluorophores that, in TIRF, are illuminated or dark, respectively. Light from these point sources is collected and imaged by the same lens as is used to create the TIRF illumination. Not drawn to scale.]] [Adapted by permission from Macmillan Publishers Ltd: Sako and Yanagida. Single-molecule visualization in cell biology. Nat. Rev. Mol. Cell Biol. (2003) vol. 4(Suppl.) pp. SS1–SS5, ©2003.]

![Figure 49.6](image)
Chapter 49 Light in Simple Media

[A major issue in microscopy is the problem of not seeing the many objects in a sample that are not of interest. Even with fluorescence microscopy, often there is “background” (unwanted) fluorescence degrading the image. [[Not ready]]...
The “forbidden” region nevertheless carries electromagnetic fields, which can impart energy to pointlike inhomogeneities such as a fluorophore. In the situation shown in Figure 49.6, most of the specimen is unilluminated, except for a layer 100–300 nm thick near the glass surface. If we wish to see only the fluorescent molecules in this region (for example, tagged receptor molecules in a cell’s membrane), then TIRF microscopy is a good approach. Stray fluorescent molecules not in this region will not be seen, because they are not illuminated.]

Ex. [[Won’t the same argument imply that the fluorophores’ light will be trapped, and hence not visible in the microscope? Solution: Light from the fluorophores needs to pass from lower to higher index, so total internal reflection does not apply. Moreover, the fluorophores are point sources of light, and not very far from the interface as was assumed in our derivation of the law of refraction.]]

49.4.5 Frustrated total internal reflection

A propagating wave can “tunnel” through a region supposedly forbidden by TIR, into a second dielectric region.

For the Light which falls upon the Farther Surface of the first Glass where the interval between the Glasses is not above the ten hundred thousandth part of an Inch will go through that Surface, and through the Air or Vacuum between the Glasses, and enter into the second Glass.

— Newton, Opticks (1704)

[[Not ready]]... [[Problem 49.6.]]

FURTHER READING

Intermediate:

More realistic treatments of polarizable material

Our pictorial approach to bound charge and current summarizes the results of an analysis that is really only valid for a restricted class of materials, such as dilute gases, nonpolar liquids, and molecular solids with weak interactions between the molecules. If we want to predict bulk material parameters from microscopic details in such situations, we can make a multipole expansion of the fields from each constituent, spatially average over length scales relevant to the problem (but much longer than the size of the constituents), and then find the effective continuous charge density and flux that could have given rise to the same fields. Usually we may stop at dipole order, as in the main text.

For many materials, quantum-mechanical couplings between constituents invalidate even the approach just mentioned. A more general approach appears in Zangwill, 2013, chaps. 6 and 13. However, our concerns were restricted to understanding general properties of linear response; the heuristic approach in this chapter and Chapter 51 are intended to give physical meaning to formulas allowed by general principles such as rotational, time reversal, and (when appropriate) spatial inversion invariance.

Dissipation and frequency dependence

Suppose that an electric field varies harmonically in time: \( \vec{E}(t) = \frac{1}{2} \vec{E} e^{-i \omega t} + \text{c.c.} \). In a medium that is itself time-translation invariant, we will then find that the displacement \( \vec{D}(t) = \frac{1}{2} \vec{D} e^{-i \omega t} + \text{c.c.} \). If the medium is linear, then

\[
\vec{D} = \varepsilon(\omega) \vec{E},
\]

which defines the frequency-dependent permittivity function. We have tacitly assumed that \( \varepsilon \) is real, but this need not be the case if there is dissipation.\(^{22}\)

To model complex permittivity, imagine a material composed of polarizable “molecules” with density \( \rho_{\text{mol}} \), consisting of a pair of charges \( \pm q \) that can separate by \( \Delta x \). Let \( \chi(\omega) = (\varepsilon(\omega)/\varepsilon_0) - 1 \), as usual. In response to the field, charge will separate by \( \Delta x = \frac{1}{2} \Delta \varepsilon e^{-i \omega t} + \text{c.c.} \).

The density of induced dipole moment is then \( \rho = \rho_{\text{mol}} q \Delta x \). That result lets us find the velocity \( \nu(t) = \frac{1}{2} \vec{v} e^{-i \omega t} + \text{c.c.} \), where

\[
\vec{v} = \frac{-i \omega \varepsilon(\omega) E}{\rho_{\text{mol}}},
\]

The rate at which the field does work on the particle is \( q \vec{E} \nu \) times \( \nu \), or

\[
q \vec{E} \nu = q \left( \frac{1}{2} \vec{E} e^{-i \omega t} + \text{c.c.} \right) \left( \frac{-i \omega \varepsilon(\omega) \vec{E}}{2 \rho_{\text{mol}}} e^{-i \omega t} + \text{c.c.} \right).
\]

The time average of that power, per volume, is thus

\[
\frac{1}{2} \left( -i \omega \varepsilon(\omega) \vec{E} \right)^2 + \frac{1}{4} \left( i \omega \varepsilon(\omega) \vec{E} \right)^2 = \frac{1}{2} \rho_{\text{mol}} \varepsilon(\omega) |\vec{E}|^2 \Im \chi.
\]

As claimed, if the permittivity function is complex, then the material dissipates energy (into heat). Similar remarks apply for the magnetic permeability.

\(^{22}\)See also Section 54.3.1 (page 754).
More general response functions

Examples of nonlinear electric response functions include piezoelectric crystals under stress, or ferroelectrics (“electrets”), both of which have nonzero $\vec{P}$ in zero applied field (Section 6.5.1’, page 92). Also, any medium will be linear only in a regime of sufficiently weak applied fields. For example, the orientational ordering of water molecules (Section 6.8, page 87) must eventually saturate (100% alignment) at high applied fields. Much of optics deals with media in their linear regime, but there is also a big field of “nonlinear optics” (Section [[Not ready]])).

Similarly, ferromagnets have nonzero $\vec{M}$ at zero applied magnetic field (Section 17.5.2, page 259), and again any medium is only magnetically linear for sufficiently weak applied fields.
49.1  **Waves in conductive medium II**
First review Problem 18.8 (page 294). This time, an electromagnetic plane wave propagates through vacuum, then enters a conducting medium, obeying an ohmic relation with conductivity \( \kappa \). The medium is not polarizable (\( \varepsilon = \varepsilon_0, \mu = \mu_0 \)). Assume that the medium is everywhere electrically neutral. The wave is initially traveling along a direction perpendicular to the planar surface of the medium, which extends to infinity beyond that surface. Find a solution to the Maxwell equations that accounts for the free charge flux set up in the medium, and that includes the incoming wave, a transmitted wave, and possibly a reflected wave as well. How much of the wave is reflected, and how does that change in the limit \( \kappa \to \infty \)?

49.2  **Electrorotation of living cells**
[[Not ready]]

49.3  **Total internal reflection fluorescence**
A plane wave with frequency \( \omega \) propagating in an isotropic medium with refractive index \( n_1 \) encounters a planar interface to a medium of lower index \( n_2 \), with polarization along \( \hat{x} \) (s wave polarization, Figure 49.4, page 695). The angle of incidence \( \theta \) is greater than the critical angle.

a. Generalize Equations 49.12–49.15 for this situation. Use them to find the intensity of the transmitted wave (maximum over time of \( |\tilde{E}(t,z)|^2 \)) at fixed perpendicular distance \( z \) from the interface and varying \( \theta \).

b. In the experiment shown in Figure 49.5a (page 699), \( n_1 \) was equal to 1.65 and \( n_1/n_2 \) was 1.1. The depth in the sample, times the angular frequency of the light, was \( z\omega = 0.905(c/n_2) \). Graph your answer to (a) and compare to the appropriate region of Figure 49.5b (page 699).

49.4  **Transmission and reflection coefficients**
Apply the analysis similar to the one that led to Equation 49.16 (page 697) to find the complex amplitude of the reflected wave, but for a non-TIR situation (the angle of incidence is less than the critical angle). Also find the complex amplitude of the transmitted wave. You can restrict attention to only the s wave case.

49.5  **Polarization of evanescent field**
This is a continuation of the preceding problem, and an application to a current research technique. First work Problem 49.3.

**Background:** Polarized total internal reflection fluorescence microscopy, or “pol-TIRF,” is an essential experimental technique in many labs. The main points are:

- TIRF excitation improves signal-to-noise in fluorescence microscopy by only creating electric fields in a thin layer next to the floor of the experimental chamber.
- These electric fields retain information about the polarization of the laser beam that gave rise to them, a fact that can be used to learn about the orientation of a single...
fluorescent molecule in the sample.

The first point is discussed in Section 49.4 (page 695). Problem 49.3 explored the second point for one incoming polarization. But unlike that situation, where angle $\theta$ was variable and depth $z$ was fixed, for the present discussion it’s the other way around.

Again the geometry is given in Figure 49.4 (page 695): A linearly polarized, monochromatic wave of angular frequency $\omega$ enters a sample chamber filled with water (refractive index $n_w \approx 1.33$) from a medium that we assume also to be isotropic, nondissipative, and nonchiral, with index $n_1 \approx 1.46$ at frequency $\omega$. If you wish, you may also assume that the media are nonmagnetic: $\mu_1 = \mu_2 \approx \mu_0$.

Again the incoming wave (in the region $z < 0$) has wavevector $\vec{k}$ lying in the $yz$ plane and all fields are independent of $x$. We are interested in angles of incidence $\theta$ that are larger than the critical angle.

The main text and Problem 49.3 considered “s wave” polarization: $\vec{E}$ was parallel to $\hat{x}$. Now consider both that case and “p wave” polarization ($\vec{B}$ parallel to $\hat{x}$).

Do:

a. Consider incoming light with vacuum wavelength 514 nm. Find the critical angle. Find the exponential amplitude falloff length scale for the fields, assuming $\theta = 70$ deg.

b. Start with the s wave case. Using formulas in the main text, characterize in words the type of polarization you get for the evanescent electric field (for example, linear/elliptical/circular). Then substitute the numbers in (a) to get a quantitative prediction for $\vec{E}_T$ in terms of $\vec{E}_i$. (You need not find the overall phase shift.)

c. Repeat for the p wave polarization: Adapt the formulas of Section 49.4.1 (page 695) for this case. Again characterize in words the type of polarization obtained, then substitute the numbers in (a) to get a quantitative prediction for $\vec{E}_T$ in terms of $\vec{E}_i$.

49.6  Frustrated TIR

[[Not ready]]
Čerenkov Radiation

You don’t understand anything until you learn it more than one way.

— Marvin Minsky

50.1 FRAMING: BOW SHOCK

When we think of the generation of radiation, we generally envision a charge that is shaking, braking, circulating, or otherwise accelerating. So it may seem reasonable that a charged particle in uniform, straight-line motion cannot generate radiation, as we indeed found in vacuum in Chapters 33 and 39. Surprisingly, however:

Electromagnetic phenomenon: When a charged particle moves through a medium faster than the local speed of light, it emits radiation even without accelerating.

Physical idea: A stationary observer transitions suddenly from an early-time regime, with no causal contact to the charge, to a late-time regime that does see it, similarly to the bow shock in the water behind a speedboat.

50.2 IN VACUUM, A CHARGED PARTICLE HAS ONE SOURCE POINT IN THE OBSERVER’S PAST LIGHT CONE

For a charge in uniform, straight-line motion in vacuum applying the Liénard–Weichert formula\(^1\) gave that there is no radiation (for example, \(\vec{E}\) and \(\vec{B}\) fall with distance faster than \(1/R\)). The key step was the observation that at any observation place and time, there is always exactly one retarded source point, and hence no sudden jump in the potential. Not surprisingly, the fields were exactly the same as what we found by doing a Lorentz transformation on the fields of a charge at rest (Section 33.4.2). We must reconsider that conclusion if a transparent dielectric medium, such as water, is present, because its presence spoils Lorentz invariance.

50.3 IN A DIELECTRIC MEDIUM, THERE CAN BE TWO SOURCE POINTS, OR NONE, IN THE OBSERVER’S PAST LIGHT CONE

Sections 6.5 and 49.2.3 argued that may summarize a medium’s effect simply by modifying the Maxwell equations, replacing \(\varepsilon_0\) by a larger permittivity \(\varepsilon\). But now an interesting

\(^1\)Section 39.5.2 (page 598)
50.3 In a Dielectric Medium, There Can Be Two Source Points, or None, in the Observer’s Past Light Cone

Figure 50.1: [Spacetime diagram.] **Graphical solution of** \( c_m(t_p - t_i) = R_z \), a modified form of Figure 39.4 (page 598), for the special case in which the field point \( \vec{r} \) lies on the \( z \) axis. The axes are scaled such that trajectories with speed \( \pm c_m \) lie at \( \pm 45 \text{ deg} \). The \( x \) and \( y \) directions have been suppressed for clarity. The two past-directed, light-speed lines from \( (c_m t_F, z_F) \) hit the particle trajectory exactly once each, whereas those from \( (c_m t_F, z_F') \) both miss the trajectory altogether.

Possibility arises: What if the particle moves faster than the speed of light in medium, that is, with speed greater than \( c_m = c/n \), where \( n = \sqrt{\varepsilon / \varepsilon_0} \)? No physical law forbids a particle from moving through water at, say \( 0.9c \), which is \( \approx 1.2c_m \).

It is true that the modified Maxwell equations have a Lorentz-like invariance, with \( c_m \) playing the role of light speed, and we can use that invariance to find the fields if \( v < c_m \). In the contrary case, however, there’s no Lorentz-type transformation that can bring us to the rest frame of the particle, so the method used in Section 33.4.2 is inapplicable.

Luckily, the proof that the radiation Green function solves the Maxwell equations is mathematically just as valid in the medium as it was in vacuum; we need only substitute \( c \rightarrow c_m \) in the derivation of Section 39.5.2 (page 598). However, the geometry is different when \( v/c_m > 1 \). The Green function solution again tells us to look back in time from an observer to the charged particle along lines moving at speed \( c_m \), but now we see there may be two such points, or even none (illustrated Figure 50.1). That latter case implies that unlike the vacuum case, at a given instant of time there are some points in space where the fields are zero. No matter how far back in time we look on the trajectory, these places have not yet come into causal contact with the moving charge, so they don’t yet “know” that it’s coming.

Figure 50.2 establishes more generally that there are always exactly two or zero contributions to the vector potential, and that the crossover between the cases occurs on a cone-shaped locus. In contrast to Figure 39.5b, in this case the stick held fixed on the \( z \) axis is longer than the pivoting stick (Figure 50.2b). You’ll explore the consequences of this difference in Problem 50.1, but the upshot is that:

- Unlike the vacuum case, an observer first “learns” about the oncoming charge via a *singular* field, a “shock wave” analogous to the bow wave of a boat moving through water faster than the speed of water waves. Inside this zone, there are always two points along the trajectory in causal contact with the observer.
- That “shock wave” can carry energy out to infinity, a form of radiation very different from what we found in the multipole expansion.

**Your Turn 50A**

Redraw Figures 50.2a,b for the case where \( z_p > z_Q \) and comment.
Chapter 50 Čerenkov Radiation

Figure 50.2: Modified form of Figures 39.5a,b (page 599). In these figures, the time direction is suppressed but the y direction is shown. An observation is made at P. At time $t_0$, a charge is at Q; at earlier times, the charge sits lower on the z axis. (a) Circles centered on $pc_\omega t_\omega$, of radii $c_\omega(t_0 - t_\omega)$, for four choices of time $t_\omega$ prior to the observation time. The locations of the charge at those times are shown as stars on the z axis, and circles centered on those points are shown with matching colors. The circles only cover a wedge-shaped region of the plane, so none of them hit observation point P′. Moreover, two of the circles hit points like P inside the wedge. (If the x direction had been shown, the circles would instead be spheres filling out a cone.) (b) Two sticks (heavy lines) with lengths in the ratio $w : c_m$ are joined by a hinge. The longer stick is held fixed and vertical, while the shorter one swings around the hinge. The end of the stick hits the ray $QP$ in exactly two points, but never hits $QP'$ at all.

50.4 INTERPRETATION

The phenomenon you’ll find is called Čerenkov radiation\(^2\) It gives rise to the characteristic blue glow emanating from a water-cooled nuclear reactor.\(^3\) Čerenkov light is also essential for particle identification in accelerator physics (via the $\beta$ dependence of the radiation cone) and in searches for exotic particles impinging on Earth.

The result may seem paradoxical: How can a non-accelerating charge radiate? Remember, however, that the one charge we investigated is not the only one in the system. The medium is polarizable because it contains many charges in the deformable molecules that constitute it. As the free charge of the particle flies past one such molecule, it gives each molecule a momentary jolt. The sum of the resulting fields from all of the molecules can and does include a radiation component, if $v > c_m$.\(^{[Not ready]}\)

\(^2\) Or Vavilov–Čerenkov radiation (after S. Vavilov and P. Čerenkov, who observed it experimentally). But it was predicted theoretically by Oliver Heaviside, in papers published in 1888–89.

\(^3\) Legend has it that in the first cyclotrons, beam alignment was achieved by observing Čerenkov light generated in the experimenters’ eyes. Be that as it may, astronauts outside our protective magnetosphere and atmosphere do see flashes of light generated in their eyes by the Čerenkov mechanism from individual cosmic ray particles.
Figure 50.3: [Computer simulation.] **Particle-ID discrimination.** Left: Electrons undergo multiple scattering and bremsstrahlung, which then produce electron-positron pairs and a diffuse Čerenkov cone. Right: Muons scatter little, leading to a sharp Čerenkov cone. [Images by T2K collaboration, courtesy Atsuko K. Ichikawa.]

### 50.5 APPLICATION: PARTICLE IDENTIFICATION IN UNDERGROUND DETECTORS

[[Not ready]]... Figure 50.3.

---

**FURTHER READING**

*Intermediate:*
Smith, 1997; Ginzburg, 1989.

*Technical:*
Historical: Jelley, 1958.

[[Allison, 2023.]]
50.1 What a shock
In this problem, assume \( \mu \approx \mu_0 \). Also neglect dispersion (assume that the permittivity \( \varepsilon \) is constant).

a. Draw a figure analogous to Figure 39.6 (page 600), with appropriate changes for a particle moving with speed \( v > c_m \).

b. Use your figure to derive expressions for the scalar and vector potentials, adapting the scheme in Section 39.5.2 (page 598).

c. Compute appropriate derivatives to find what direction \( \vec{E} \) and \( \vec{B} \), and hence the Poynting vector, poyn. Which way does energy flow? Will it just stay concentrated near the \( z \) axis, or flow outward?

[Hint: The problem has one rotational symmetry axis, so the formula for curl in cylindrical coordinates (\( r \), \( \varphi \), and \( z \)) may be useful:

\[
\hat{\nabla} \times \vec{A} = \hat{r} \left(r^{-1} \frac{\partial \vec{A}_z}{\partial \varphi} - \frac{\partial \vec{A}_r}{\partial z}\right) + \hat{\varphi} \left(\frac{\partial \vec{A}_z}{\partial r} - \frac{\partial \vec{A}_r}{\partial z}\right) + \hat{z} \left(r^{-1} \frac{\partial}{\partial r} (r \vec{A}_r) - r^{-1} \frac{\partial \vec{A}_r}{\partial \varphi}\right).
\]

Here \( r \) is distance from the \( z \) axis; \( \hat{r}, \hat{\varphi}, \) and \( \hat{z} \) are all unit vectors; and \( \vec{A}_\varphi = \hat{\varphi} \cdot \vec{A} \) and so on.]
CHAPTER 51

Chiral Media

The value of a formalism lies in the degree to which it encourages physical intuition in guessing the solution of new problems.
— A.B. Pippard

51.1 FRAMING: CROSS-SUSCEPTIBILITY

Section 49.3.3 argued that plane waves propagate in isotropic, linear, uniform media in much the same way as in vacuum: Changing the value of $\epsilon$, $\mu$, or both can slow the waves down, but there are still two transverse polarizations, and a wave's polarization is constant as it propagates.

Real materials often consist of objects, such as water molecules, that are individually far from being isotropic. Nevertheless, in liquid water many molecules are jumbled together in random orientations. The same holds for a mixture, such as a solution, and even for an amorphous solid material such as glass. In each of these materials, the overall polarizability tensors are therefore averaged over all possible rotations, and hence are proportional to the identity tensor, effectively creating an isotropic medium. So we again predict no effect on the polarization of light.

The prediction just made fails spectacularly, however, even for everyday material like a solution of sugar in water! For example, corn syrup (essentially a concentrated glucose solution) rotates the axis of linearly polarized light in a clockwise direction when viewed along $k$ (Figure 51.1 and Media 20). This Electromagnetic Phenomenon is called circular birefringence. To understand it, we’ll need some to think carefully about symmetry.

Electromagnetic phenomenon: A solution of randomly oriented molecules is completely isotropic, yet nevertheless it can rotate polarized light. Physical idea: Molecular chirality affects light by allowing cross-susceptibility.

51.2 CIRCULAR BIREFRINGENCE SEEMS TO PRESENT A PARADOX

In greater detail, experiments like the one sketched in Figure 51.2 show that:

1 See Problem 14.2 (page 220).
2 Some books use the synonym optical activity; media with this property are sometimes said to possess optical rotatory power.
Figure 51.1: [Photos.] **Circular birefringence.** A dish containing a layer of corn syrup 0.5 cm deep was placed between two polarizing filters and illuminated from below with white light. (a) When the two polarizers are oriented at 90°, the region outside the dish is darkest, but light passing through the syrup is partly transmitted. (b) When the angle between polarizers is increased to about 92°, a little light passes through the outer region, but the central region is darker. Red light is maximally blocked at this angle, giving a bluish tint to the light transmitted through the syrup. (c) At angle about 100°, blue is maximally blocked, giving a reddish tint to the light transmitted through syrup. See also Media 20. [Photos courtesy Le-Qi Tang.]

**Figure 51.2:** [Schematic.] **Measuring circular birefringence with a polarimeter.** The arrows represent the electric field vector in a linearly polarized beam of light. They are shown rotating by an angle $\theta$ as the light passes through the sample; the rotation shown corresponds to the value $\theta = +\pi/2$. In the situation shown, an observer looking into the oncoming beam sees the electric field rotating in the clockwise direction as the beam advances through the medium. Try looking at this figure in a mirror to see that the sense of the rotation changes sign.

- If linearly polarized light enters traveling perpendicular to the surface, then linearly polarized light exits the other end, but the direction of polarization has been rotated about $\hat{k}$.
- The effect is itself rotationally invariant: The angle of rotation does not depend on the initial direction of polarization.
- The angle of rotation is linearly proportional to the depth of the sample (distance traversed).
- The sign of the rotation is a fixed characteristic of the medium.

What property could determine this direction of rotation, a choice that breaks spatial inversion invariance (Figure 51.2)? Because the Maxwell equations are themselves invar-
Figure 51.3: Chirality. (a) A simple chiral molecule can be obtained by bonding four different atoms to a central carbon. The molecule 1 cannot be brought into coincidence with its mirror image (shown as 2) by any rotation; 3,4 show some failed attempts. (b) Chirality of amino acids. Cyclosporin, a cyclic peptide made by fungi, contains a pair of alanine groups that are mirror images. (c) Helical wires as a model for chiral molecules. When $\vec{B}(t)$ is increasing in magnitude in the direction shown, the two loops of wire will create opposite electric polarizations due to their chirality. The case of nonconstant applied $\vec{E}(t)$ involves a similar cartoon. (d) Idealized, macroscopic crystal of tartaric acid and its mirror. [(a) Art by Sarina Bromberg; (b) from Goodsell, 2016.]

The only source of polarization rotation must be a property of the sugar molecules themselves—one not shared by, say, water molecules. Indeed, glucose differs from H$_2$O by a property called chirality. An object that cannot be superimposed on its mirror image by any rotation or translation is called chiral (Figure 51.3a). That is, the mere presence of a chiral object breaks inversion symmetry. In contrast, the oxygen, nitrogen, and argon making up most of our atmosphere are nonchiral, and hence the partial polarization of the blue sky is not washed out by different rotations from the passage through varying thicknesses of air. But we still face a paradox, because the argument about orientational averaging at the start of this section seems to apply to an isotropic

---

3Problems 15.4 and 18.1.

4Objects that are not chiral are called “nonchiral” or “achiral.” The two mirror images of a chiral object are called each other’s enantiomer. Circular birefringence differs from ordinary birefringence, which can happen even in a nonchiral crystal consisting of nonchiral molecules (Chapter 52).

5Section 47.7.2 (page 682).
solution of any kind of molecule—chiral or not.

We must be missing something crucial. Since we calculated that an effect is zero and observed that it’s not, maybe we made a bad approximation. Did we truncate a power series to an order at which the effect does not yet arise? No, that’s not the answer.

51.3 CROSS-SUSCEPTIBILITY

The resolution of our puzzle lies in another possibility that we’ve neglected so far. The most general response function that is uniform, linear, isotropic, and lossless is actually:

\[
\begin{bmatrix}
\vec{P} \\
\vec{M}
\end{bmatrix} = 
\begin{bmatrix}
\varepsilon_0 \chi_e \mathbb{1} & ? \\
? & \frac{1}{\mu_0} \chi_m \mathbb{1}
\end{bmatrix}
\begin{bmatrix}
\vec{E} \\
\vec{B}
\end{bmatrix}.
\] (51.1)

Until now, we have implicitly assumed that the off-diagonal blocks, or cross-susceptibilities, were zero. If that’s not the case, the constitutive relations (Equations 6.11, page 81 and 49.9, page 693) will acquire cross-terms. As long as the cross-susceptibilities are proportional to the identity 3-tensor, they will still be compatible with rotational invariance (isotropy).\(^6\)

51.3.1 Macroscopic physical realizations give intuition about a possible new effect

Are the cross-susceptibilities imagined in Equation 51.1 really allowed? To see, let’s invent another simple physical realization, along lines similar to the model in Section 49.3.4. Consider a helix of wire open at each end (Figure 51.3c). This helix can be left- or right-handed. Its handedness has nothing to do with how it is oriented in space; for example, flipping it end-for-end does not change the handedness. In short, it is a chiral polarizable object, and that property will not be erased by rotational averaging.

Imagine a spatially uniform \(\vec{E}\) field directed along the helical axis. This applied field leads to an electric dipole moment \(\vec{D}_e\) as usual. But suppose that the magnitude \(|\vec{E}|\) is switched on at some moment, so that \(\partial \vec{E}/\partial t\) is also parallel to \(\vec{E}\). Charge must flow to establish the electric polarization. The helical structure then forces that charge flux to have an azimuthal component, so it generates a magnetic dipole moment \(\vec{M}\) proportional to the time derivative. It will be convenient to express the proportionality as \(\vec{D}_m/c = \varepsilon_0 \eta (\partial \vec{E}/\partial t)\); the constant \(\eta\) summarizes this new kind of polarizability.

If \(\partial \vec{E}/\partial t\) points upward, net positive charge flows up, regardless of the handedness of the helix. The direction of the azimuthal current, and hence also the sign of \(\eta\), depend on the handedness of the helix:

**Your Turn 51A**

Show that \(\eta\) is positive for a right-handed helix and negative for a left-handed one.

---

\(^6\)If response is instantaneous, then the matrix entries in Equation 51.1 are constants. If not, then they may depend on the frequency of incoming light (Section 49.3.5, page 694).
Next, imagine a magnetic field directed along the helical axis direction with $||\vec{B}||$ increasing in time, so $\partial \vec{B}/\partial t$ is parallel to $\vec{B}$ (Figure 51.3c). The Faraday law and the ohmic relation imply that this field again induces a current in the wire, creating a cylindrical current sheet that partially cancels the $\vec{B}$ inside the coil. But the helical shape also imposes an axial motion of charge, and hence an electric dipole moment:

**Your Turn 51B**

a. Show that $\vec{D}_n$ can be expressed as $-c\varepsilon_0 \eta' (\partial \vec{B}/\partial t)$, where the constant $\eta'$ is positive for the right-handed helix, or negative for the left-handed one.

b. Check that the units of $\eta'$ and $\eta$ match.

c. Show that the sign of the charge carriers is immaterial.

Both arguments above are for $\vec{E}$ and $\vec{B}$ directed along the helical axis. But even if the medium contains randomly oriented helices, some of them will have their axes along $\vec{E}$ or $\vec{B}$.

51.3.2 The general linear response function has new cross-terms

The preceding discussion suggested that in general, a uniform, linear, isotropic, lossless, medium will have induced polarizations of the form

$$
\begin{bmatrix}
\vec{P} \\
\vec{M}
\end{bmatrix} = \varepsilon_0 \begin{bmatrix}
\chi_e & -\eta' \partial / \partial t \\
\eta \partial / \partial t & \chi_m
\end{bmatrix}
\begin{bmatrix}
\vec{E} \\
\vec{B}
\end{bmatrix},
$$

(51.2)

where the new material parameters $\eta$ and $\eta'$ are nonzero if the medium is chiral. (Here $\vec{M} = \vec{M}/c$ and $\vec{B} = c\vec{B}$. Those definitions simplify our formulas by giving all the entries in the matrix the same dimensions.) Equation 51.2 also abbreviates by omitting the factors of $\vec{F}$.

We conclude that the generalized response function proposed in Equation 51.1 is physically possible, though with the surprise that the cross-susceptibilities both involve time derivatives. In retrospect, however, this is not so surprising. A collection of static molecules may break spatial inversion symmetry, but not time reversal. Making the off-diagonal entries of Equation 51.2 odd in frequency was needed to join quantities with different time-reversal characters.

In addition to possible frequency dependence of $\eta$ and $\eta'$, the new cross-terms have explicit frequency dependence due to the time derivatives in Equation 51.2.

Previously you showed that $\eta$ and $\eta'$ always have the same sign in a macroscopic physical realization of cross-polarization. Replacing the helices by their mirror images

---

7Lenz’s law; see Section 18.3.1.

8There is a slight change of notation here: Now $\eta$ and $\eta'$ include the density of the polarizable molecules. Each now has dimensions of time.

9Your Turns 51A and 51Ba. Indeed, Onsager reciprocity implies quite generally that $\eta = \eta'$. See Landau et al., 1984, Eq. 103.10. (Note that Landau uses spatial derivatives, but these can be converted to time derivatives by using the Maxwell equations.)
reverses the signs of both $\eta$ and $\eta'$. If molecular chirality is indeed the cause of circular birefringence in solutions, then we would similarly predict that two pure solutions of molecules differing by spatial inversion would exhibit equal magnitude, but opposite sign, circular birefringence, and experimentally that prediction is confirmed. Let’s now make the connection explicit.

51.4 THE ORIGIN OF CIRCULAR BIREFRINGENCE

We now set up a plane-wave trial solution for the medium described by Equation 51.2:

$$\vec{E}(t, z) = \frac{1}{2} \vec{E} e^{i(-\omega t + kz)} + \text{c.c.}; \quad \vec{B}(t, z) = \frac{1}{2} \vec{B} e^{i(-\omega t + kz)} + \text{c.c.}$$

and seek complex amplitudes that make it a solution of the Maxwell equations.

Your Turn 51C

a. Show that $\vec{E}$ and $\vec{B}$ must both be perpendicular to $\vec{k}$. Use the Faraday law to get $\vec{B}$ in terms of $\vec{E}$, then substitute that expression into the Ampère law (Equation 49.8 (page 691)).

b. To keep things simpler, you may (unrealistically) set $\chi_\pm = \tilde{\chi}_m = 0$, that is, neglect the ordinary susceptibilities and focus only on the cross-susceptibilities. Show that

$$\vec{E}(-ik^2/\mu_0 \omega + i\omega \varepsilon_0) + \vec{k} \times \vec{E}(-\omega \varepsilon_0 \eta - \omega \varepsilon_0 \eta') = 0.$$ 

c. Show that the last equation simplifies if we expand the polarization vector in the circular polarization basis (helicity basis) $\vec{E}_{(\pm)} = (\hat{x} \pm i\hat{y})/\sqrt{2}$ (Equation 18.34, page 285), and that in a chiral medium, each circular polarization propagates with a different phase velocity $\omega/k_\pm$.

The two wave speeds that you found can as usual be expressed as indices of refraction, $c/n_\pm$, explaining the term “circular birefringence.”

We can now ask what happens to an arbitrary linear combination of the two circularly polarized eigenmodes of propagation. Specifically, if we feed in a linearly polarized plane wave, its frequency will not change, by time-translation invariance. Once the wave enters the medium, however, each circularly polarized component propagates with a different wavenumber $k_\pm$ (the two values you found in Your Turn 51C). Mathematically, we can analyze the incoming linearly polarized light into circularly polarized components, let each propagate, then reassemble the two resulting waves via superposition to see what emerges into vacuum at the other end of a slab of medium.
Your Turn 51D

a. Try the procedure just outlined. To interpret the result, show that the final wave is again linearly polarized, but in a direction rotated relative to the original. Show that the angle of rotation is independent of the original direction of polarization, and explain why this had to be so.
b. Show that the angle of rotation is proportional both to \( n_+ - n_- \) and to the thickness of the slab.
c. In particular, show that \( n_+ - n_- \) is proportional to the density of chiral polarizable objects (for example, concentration of a solution).

Your result (a) justifies the alternate term “optical rotatory power” as a synonym for circular birefringence (Figure 51.2, page 712). Result (c) implies that the total rotation depends on the “chiral optical depth.” That is, it equals a constant (characterizing the chiral molecule in question) times the projected areal density of those molecules encountered by the light during its passage.

In short:

- Cross-susceptibility is possible in a medium that breaks spatial inversion invariance. That could occur because the medium contains chiral molecules (such as most sugars, proteins, DNA, and so on), even if they are arranged isotropically. Indeed, any chiral molecule, whether or not it looks helical, can give rise to circular birefringence. For example, we could take \( \text{CH}_4 \) and substitute three of the hydrogen atoms with distinct things (maybe an OH group for one, a Cl atom for another, and a chain for the third). Even if each group is itself nonchiral, the whole thing will break spatial inversion invariance (Figure 51.3a).
- Alternatively, nonchiral molecules may be arranged in a chiral crystal structure (such as in quartz).
- However, air (\( \text{O}_2, \text{N}_2 \)), liquid water (\( \text{H}_2\text{O} \)), and so on don’t display this phenomenon—they are all disordered arrangements of nonchiral (inversion-invariant) objects.
- The time derivatives in Equation 51.2 predict that the effect will be strongly dependent on frequency.

51.4.1 Historical remarks

- Remarkably, in 1825 (long before Maxwell) A. Fresnel interpreted the polarization rotation observed in optically active liquids as a difference in refractive index for left- and right-circularly polarized light. Fresnel went on to predict that therefore letting a beam of unpolarized light enter at an angle into such a medium would separate it into two circularly polarized components, due to their unequal refraction angles.
- L. Pasteur intuited the connection between chiral molecules and circular birefringence in 1848, also before Maxwell, just by thinking about symmetry. Pasteur noticed that synthetic tartaric acid differed from the natural form in that it lacked circular birefringence. He then crystallized the synthetic version and noticed that the tiny
crystals came in two mirror image forms (Figure 51.3c). He reasoned that the two macroscopic shapes might reflect molecular sorting during crystallization, with each small crystallite consisting exclusively of one version of the underlying molecule.\(^\text{10}\) To investigate, he then painstakingly separated a pile of these tiny crystals into two piles, in this way manually purifying the two enantiomers. Dissolving each one in water indeed yielded two solutions with opposite circular birefringence!

- Later, it became clear more generally that living organisms discriminate between the two enantiomers of each biomolecule and only synthesize the one they need. In contrast, most artificial synthesis techniques make both enantiomers indiscriminately (they create a “racemic mixture”). Most purification techniques are unable to separate enantiomers (apart from Pasteur’s heroic effort).\(^\text{11}\) Thus, the presence of circular birefringence can in principle distinguish artificial from synthetic compounds, a circumstance that provided the crucial plot element in (at least one) novel from the classical era of British crime fiction.

### 51.5 HOW TO OBSERVE CIRCULAR BIREFRINGENCE

To follow up on those questions, Figure 51.1 (page 712) shows a dish with a layer of corn syrup (a concentrated sugar solution) illuminated from below and viewed from above, with polarizers fixed above and below the dish. The \(\pi/2\) filter orientation that blocks light passing outside the dish does not completely block light passing through it. The effect is absent when we substitute water in place of sugar solution.

With a thicker layer of syrup, a greater rotation of the second polarizer relative to the first is required to obtain the same transmission of light. However, adding water has no effect: It dilutes the sugar while increasing layer thickness, with no effect on the projected areal density of sugar molecules.

The figure also shows that blue (higher frequency) light rotates more than red, consistent with the qualitative expectation in Section 51.3.2.

### 51.6 EXTENSIONS

- Chapter 54 will show that rotation of polarization can occur in an astrophysical plasma, if a uniform \(\mathbf{B}\) field is present. Although this medium is very different from sugar solution, nevertheless it breaks spatial inversion symmetry in a way that is mathematically similar to what we have studied, again leading to different phase velocities for the two circular polarizations.
- The math predicted that circular birefringence goes to zero at zero frequency, due to the time derivative in Equation 51.2 (page 715). More generally, the entire spectrum

---

\(^{10}\)Pasteur’s breakthrough is all the more remarkable in that the very existence of molecules, let alone their definite shapes, was controversial in 1848.

\(^{11}\)Even mass spectrometry cannot separate them, because they have the same charge/mass ratio.
of circular birefringence is called the **optical rotatory dispersion**, and it amounts to a fingerprint of the constituent molecules, independent of the ordinary dispersion. ORD is convenient to measure, because the uninteresting water molecules in a solution don’t contribute to it. Note that although the polarization rotation angle is ambiguous by 180°, its differential rate of increase as depth increases is well defined. The value of (rotation angle)/(depth×concentration) as a function of frequency is what characterizes the solute.\(^\text{12}\)

- One could now try to formulate a quantum mechanical calculation that leads from molecular structure to a prediction of the value of \(\eta\). The calculation is hard, and in the end much of the work will be discarded by averaging over random orientations. To a physicist, what’s interesting is how symmetry analysis dictates that there’s *just one* phenomenological parameter \(\eta\) characterizing the effect of chirality of an isotropic medium on light (to leading nontrivial order in frequency).

### 51.7 CIRCULAR DICHLROISM

We have seen how circular birefringence stems from a differential phase velocity between the two circular polarizations of a plane wave. In addition, chiral media can also display differential *absorption* of light between the circular polarizations, a phenomenon called **circular dichroism** (CD). An unexpected structural form of the DNA molecule called “Z-DNA” was first discovered via its nonstandard CD spectrum. Figure 51.4 shows the full frequency dependence of CD for two forms of DNA. Although both contain the same atoms in the same linear sequence, bound by the same covalent chemical bonds, a major difference is visible between the forms. For example, between 220 and 240 nm the CD has the opposite sign for the Z form. CD is a valuable nondestructive diagnostic of molecular conformation.

---

\(^{12}\) Chemists sometimes use the unit \(\text{M}^{-1}\text{dm}^{-1}\) for this quantity; you should convince yourself that it has the same dimensions as area, and indeed is in some sense a cross-section.
To quantify CD, first define absorbance $A = -\log_{10}(I_{\text{out}}/I_{\text{in}})$, where $I$ is power flux of a beam of light. Like circular birefringence, $A$ will be proportional to the path length $\ell$ and to a solution’s concentration $c_{\text{mol}}$ of absorbing molecules. Its difference between the two helicities of light, $\Delta A$, will be proportional to the concentration of absorbing molecules alone, so define the differential absorptivity as

$$\Delta \varepsilon = \Delta A / (\ell c_{\text{mol}}).$$

Thus $\Delta \varepsilon$ has units $L^2$: It is a kind of cross-section.

### 51.8 PLUS ULTRA

Was it worth the effort? It’s not much of an exaggeration to say that this story illustrates in miniature how physicists think about nearly everything. We saw the possibility of a surprising new coupling, we characterized it in terms of symmetry, we looked for what sort of physical setup had the required (lack of) symmetry, and we then confirmed that the math could transmit the key property from the physical setup to observable, quantitative predictions. Then Section 51.6 described some observations.

Although we argued that the imagined structure in Figure 51.3c would exhibit cross-susceptibility, real chiral molecules usually do not have any obviously helical structure (panels a,b). But from the symmetry viewpoint, all three lack invariance under spatial reflections, even when averaged over rotations. And that invariance is the only thing that could forbid cross-susceptibility and its symptom (circular birefringence). So we expect to observe that phenomenon with any chiral molecule—and there it is.

Section 33.3.5 claimed that, to a physicist, “beauty” often means the combined effect of inevitability and surprise. In that sense, circular birefringence and its explanation are beautiful. And then when you see the colors—that’s another level of beauty.

### FURTHER READING

**Intermediate:**
- Historical: Pasteur, 1848 (en.wikipedia.org/wiki/Louis_Pasteur#Molecular_asymmetry).
- Relativistic treatment of media: Landau et al., 1984, §76.

**Technical:**
- Experimental observation of double diffraction from optically active liquids: Ghosh et al., 2007.
51.3’a Just two enantiomers

Why are there just two forms (enantiomers) of a chiral molecule? The point is that electromagnetism, including its quantum version, is invariant under the group O(3) of orthogonal 3 × 3 matrix transformations of space. Any two molecules related by such a transformation will have the same energy, stability, excited states, and so on. And this group is twice as big as the rotation group SO(3).

The coset space of O(3) matrices modulo all rotations is just the group \(\mathbb{Z}_2\) with two elements.

51.3’b Relativistic formulation

Equation 51.2 (page 715) involves a 6 × 6 matrix of susceptibilities, which is not obviously a 4-tensor. But in fact, we can define a response 4-tensor analogously to \(F^{\mu\nu}\), as

\[
\begin{bmatrix}
0 & \vec{P}_x & \vec{P}_y & \vec{P}_z \\
-\vec{P}_x & 0 & -M_z & M_y \\
-\vec{P}_y & M_z & 0 & -M_x \\
-\vec{P}_z & -M_y & M_x & 0
\end{bmatrix}^{\mu\nu},
\]

(51.3)

where \(M_i = c^{-1}\tilde{M}_i\). This big formula can be summarized in the usual way by \(R^{0i} = -R^{i0} = \tilde{P}_i\), and \(R^{ij} = -\varepsilon_{ijk}\tilde{M}_k/c\). Also, let \(\tilde{J}_f\) denote the free charge flux 4-vector field.

In terms of these definitions, four of the Maxwell equations take the form

\[
\tilde{\partial}_\mu H^{\nu\mu} = c^{-1}J_f^\nu,
\]

(51.4)

where

\[
H^{\nu\mu} = c\varepsilon_{\nu\mu\alpha\beta}F^{\alpha\beta} + R^{\nu\mu}.
\]

(51.5)

Thus, \(H^{\nu\alpha} = \tilde{D}_\alpha\) and \(H^{\nu\sigma} = c^{-1}\varepsilon_{\nu\mu\alpha\beta}\tilde{H}_\beta\), in parallel to the naming of elements of \(F\). So \(R\) must be a 4-tensor, because the world is Lorentz invariant, and Equations 51.4–51.5 are only invariant if \(R\) is a tensor.

The remaining four Maxwell equations are unchanged from the case of vacuum, because they have no source terms.

Linear response is the statement that \(R\) is a linear function of \(F\):

\[
R^{\mu\nu} = K^{\mu\nu}_\lambda F^{\lambda\sigma},
\]

(51.6)

where the susceptibility operator \(K\) is antisymmetric on its first two indices, and also on the last two.

Let’s apply “Einstein thinking” to see what structures are allowed for the susceptibility operator. First, \(K_{\mu\rho\lambda\sigma}\) must be a symmetric operator in the sense that exchanging \(\mu\nu\) with \(\lambda\sigma\), and \(\hat{\alpha} \rightarrow -\hat{\alpha}\), must leave it unchanged. Next, note that even an isotropic medium breaks Lorentz symmetry— unlike the vacuum, it can have states of motion. But isotropy and homogeneity do imply that the only quantity describing the state of the medium is its 4-velocity \(\tilde{U}\). Hence, it must be possible to express \(K\) as a combination of \(\tilde{U}\)’s, \(\hat{\alpha}\)’s, and 4-scalar quantities describing the medium. Playing

\[\text{\footnote{See Section 32.3.1}}\]
around shows that there are only three possible forms permitted by the symmetries:

\[
K_{\mu\nu} = \frac{\alpha}{2} (\delta_\mu^\alpha \delta_\nu^\beta - \delta_\mu^\beta \delta_\nu^\alpha) + \frac{\tau}{2} \left( U^\mu U_\mu U_\nu - U^\nu U_\mu U_\mu \right) + \frac{\gamma}{2} \left( \xi^\mu_{\rho\sigma} U^\rho U_\sigma - \xi^\nu_{\rho\sigma} U^\rho U_\sigma \right)
\]

(51.7)

Here the components of the 4-dimensional Levi-Civita tensor are \( \varepsilon_{0123} = +1 \) and so on.\(^{15}\)

**Your Turn 51E**

Specialize this formula to an inertial coordinate system in which the medium is at rest. Show that the constants \( \alpha \), \( \tau \), and \( \gamma \) can be chosen so that Equation 51.7 reproduces Equation 51.2. That is, find the relation between the three phenomenological parameters \( \chi \), \( \tilde{\chi} \), and \( \eta \) and \( \alpha \), \( \tau \), and \( \gamma \).

Then substituting an arbitrary 4-velocity at once tells us the appropriate form of the susceptibility tensor in a moving medium.\(^{16}\)

Every term in Equation 51.7 must be time-reversal invariant, because a static collection of molecules does not break time-reversal invariance.\(^{17}\) (This is why the \( \gamma \) term needs a derivative.) Also, the \( \alpha \) and \( \tau \) terms are invariant under spatial inversions—but not the \( \gamma \) term. Thus, \( \gamma \) must equal zero for a non-chiral medium, as noted for liquid water in Section 51.5 (page 718).

---

\(^{14}\)More precisely, this is the most general structure to leading order in powers of derivatives. The logic is similar to what we’ve done before, for example in Section 35.5 (page 559). Some terms that may seem to be missing from our list are in fact redundant by the Maxwell equations and the constraint that \( U_\mu U^\mu = -c^2 \).

\(^{15}\)See Section 34.5 (page 542).

\(^{16}\)You previously used similar logic in Problem 34.2 (page 551).

\(^{17}\)A more general approach would be needed to include ferromagnetism.
51.1  *Circular birefringence*
Repeat Your Turn 51C, but this time without the unrealistic simplifying assumptions $\chi_e = \chi_m = 0$.

51.2  **Relativistic formulation**

a. Use Equations 51.4–51.7 to derive the plane-wave solutions for light in flowing water, relevant to the Fizeau experiment.

b. Find solutions corresponding to light propagating in an isotropic, chiral medium (such as sugar water) at rest.
CHAPTER 52

Anisotropic Media

52.1 FRAMING: MOLECULAR ORDER

The preceding chapter showed that circular birefringence was possible in an isotropic medium, due to electric/magnetic crossterms in the susceptibility of a chiral medium. We now consider another situation, a homogeneous medium in which chiral effects are negligible, but the dielectric susceptibility is not isotropic. A material may consist of molecules held in a crystal lattice (molecular order), so that their polarizability tensor is not averaged over rotations. We’ll call the ensuing behavior of light ordinary birefringence, often abbreviated to just “birefringence.”

Typical transparent, crystalline solids include quartz and many kinds of hard clear plastic. Even liquid crystals can have at least partial orientational order.

Electromagnetic phenomenon: A transparent, crystalline material can change the polarization of light, without chirality.

Physical idea: In such materials, an anisotropic molecular polarizability is not averaged over rotations.

52.2 THE SUSCEPTIBILITY TENSOR DEFINES PRINCIPAL DIRECTIONS

We are interested in a material whose dielectric susceptibility $\varepsilon_0$ is constant in space and time. If it is not a scalar times the identity tensor, then the material is called birefringent. We then generalize Section 49.3.1 (page 692) by defining

$$\bar{\varepsilon} = (\bar{1} + \bar{\varepsilon}_m)\varepsilon_0 \quad \text{and} \quad \bar{D} = \bar{\varepsilon} \cdot \bar{E}. \quad (52.1)$$

For simplicity, we also assume that the magnetic susceptibility is $\mu_m \approx 0$, so $\bar{H} = \mu_0^{-1}\bar{B}$. Then the Maxwell equations for linear media are again valid. We suppose no free charges nor currents in our medium; for example, it may be an electrical insulator.

Like any real symmetric matrix, $\bar{\varepsilon}_m$ has a basis of three mutually perpendicular, real eigenvectors. Choose coordinates for which those eigenvectors are $\hat{x}$, $\hat{y}$, and $\hat{z}$ and consider a trial solution that is a plane wave propagating along $\hat{z}$:

$$\bar{E}(t, \bar{r}) = \frac{1}{2}\bar{E}e^{-i(\omega t - \bar{k} \cdot \bar{r})} + \text{c.c.} \quad \bar{B}(t, \bar{r}) = \frac{1}{2}\bar{B}e^{-i(\omega t - \bar{k} \cdot \bar{r})} + \text{c.c.}$$

---

1 Recall the discussion on page 200.
2 Equation 6.12 (page 81). See also example 3 on page 199.
3 Equations 49.7–49.8 (page 691) and the unmodified Equations 0.2 and Equation 0.3 (page 2).
The Maxwell equations, with the constitutive relation Equation 52.1, then say
\[ i\mathbf{k} \cdot \mathbf{E} = 0, \quad i\mathbf{k} \cdot \mathbf{B} = 0, \]
\[ i\mathbf{k} \times \mathbf{E} + (-i\omega)\mathbf{B} = 0, \]
\[ i\mathbf{k} \times \mathbf{B} / \mu_0 - (-i\omega)\mathbf{E} = 0. \]

(52.2)
(52.3)
(52.4)

Equation 52.2 tells us that \( \mathbf{E} \) and \( \mathbf{B} \) must both be perpendicular to \( \mathbf{k} \). Equation 52.3 tells us \( \mathbf{B} \) in terms of \( \mathbf{E} \). Substituting into the last equation gives the dispersion relation, which is simple if \( \mathbf{E} \) is directed along either of the two transverse eigenvectors of the permittivity:
\[ k = \omega \sqrt{\varepsilon_{(\alpha)}} \mu_0, \quad \alpha = 1 \text{ or } 2. \]

propagation along a principal axis

Here \( \varepsilon_{(\alpha)} \) denotes one of the eigenvalues.

In other words, in this simple situation (light traveling along a principal axis, and polarized along another principal axis), the light propagates without change. However, those two linear polarizations propagate at different speeds (phase velocities\(^4\)). We’ll call those speeds \( c/n_{(\alpha)} \) where \( n_{(\alpha)} = \sqrt{\varepsilon_{(\alpha)}} / \varepsilon_0 \).

A material with two equal eigenvalues is called uniaxial; if all three eigenvalues are distinct the material is called biaxial. Uniaxial birefringence is called “positive” when the third (special) refractive index is greater than the other two, or “negative” in the opposite situation. For example, quartz has \( \Delta n = +0.009 \) for visible light, whereas for calcite it’s \(-0.172\).\(^5\)

### 52.3 APPLICATIONS OF ANISOTROPY

#### 52.3.1 Half-wave plate

Consider a slab of anisotropic dielectric material, cut so that its two parallel faces are perpendicular to the material’s third principal axis. Suppose that the other two principal axes have different eigenvalues of \( \varepsilon \). Thus, the material could be biaxial, or uniaxial with its special axis lying parallel to the surface. A plane wave of light enters the material propagating perpendicular to the interface (angle of incidence zero), and hence is polarized parallel to the interface. If its polarization aligns with either of the two principal axes in that plane, then it will partly enter the material with no change of polarization (and partly be reflected). We can now ask, what happens to a mixture of those two polarizations?

One particularly useful case is when the slab thickness \( z_{\text{tot}} \) is such that
\[ \omega(n_{(1)} - n_{(2)})z_{\text{tot}} / c = \pi. \]

half-wave plate

(52.5)

---

\(^4\)See Problem 19.1 (page 296).

\(^5\)Because refractive index generally depends on frequency, this qualitative property can also be frequency dependent.

\(^6\)The optical element shown in Figure 37.1c (page 578) had thickness one half as great as in Equation 52.5, but the light beam traversed it twice, so the following discussion applies to it. See Problem 52.1.
Figure 52.1: Change of linear polarization after passage through a half-wave plate. \( \vec{E}_1 \) and \( \vec{E}_2 \) are components of the polarization along principal directions of the permittivity tensor; the third principal direction points out of the page and coincides with \( \vec{k} \). In each panel, the outgoing polarization is obtained by inverting \( \vec{E}_{\text{in}} \) through either the 1 or 2 axis (Equation 52.6). Thus, there can either be no effect (a,b), or a change of polarization (c,d), depending on the angle between \( \vec{x} \) in and the crystal’s principal axes.

Now consider light that at time zero, and \( z = 0 \), is linearly polarized at some arbitrary angle to the \( x \) axis: \( \vec{E}_{\text{in}} = \vec{E}(\hat{x} \cos \alpha + \hat{y} \sin \alpha) \). How will it look when it emerges a distance \( z_{\text{tot}} \) from its entry point?

We may solve the Maxwell equations for each eigenvector component separately, then superpose the answers.\(^7\) Let \( \bar{n} = (n_{(1)} + n_{(2)})/2 \). From previous paragraphs, then,

\[
\vec{E}_{\text{out}} = \vec{E}(t, z_{\text{tot}}) = \frac{1}{2} \vec{E} e^{-i \omega (t - n z_{\text{tot}}/c)} \left( x e^{i \omega \Delta n z_{\text{tot}}/(2c)} \cos \alpha + y e^{-i \omega \Delta n z_{\text{tot}}/(2c)} \sin \alpha \right) + \text{c.c.}
\]

Thus, \( \vec{E} \) gets reflected through a plane while remaining linearly polarized (Figure 52.1). If the wave enters polarized along \( \hat{x} \) or \( \hat{y} \), then it leaves in the same polarization state. But in intermediate cases, its polarization gets rotated by an angle somewhere between zero and 90 degrees, depending on the initial polarization. We can adjust that angle by rotating the half-wave plate in its plane, so it is a useful gadget for optical setups.

### 52.3.2 Linear dichroism

As mentioned in Section 46.5, a dense material in an electromagnetic wave can have dissipative losses that also depend on the polarization of the light. This means that different polarizations can be differentially absorbed. E. Land discovered that when polyvinyl alcohol chains on a plastic substrate were heated and stretched, they became aligned and electrically conducting along their length. The resulting material, which Land named polaroid, has large ohmic dissipation when oscillating an electric field is applied along one axis. It therefore acts as a practical and inexpensive polarizing filter.\(^8\)

### 52.3.3 Imaging with polarized light

Many materials are optically transparent and colorless, including most living cells. Even when there is contrast in the absorption or scattering of light, it is often weak, and hence

---

\(^7\)The following argument is similar to the one in Section 51.4 (page 716), but with a real (cartesian) basis instead of a complex (circular) one.

\(^8\)See also Chapter 46.
52.3 Applications of Anisotropy

Figure 52.2: A thin layer of ice between crossed polarizers.

difficult to observe against the bright background illumination. This situation makes it difficult to obtain informative photographs of the internal structures that interest us.

Those structures may, however, differ from their surroundings in their refractive index, so many optical techniques attempt to convert phase information to visible intensity modulation. The simplest such method is simply to sandwich the observed sample between crossed polarizing filters. Then a sample region with nonzero $\Delta n$ will appear light against a dark background (Figure 52.2). Polarization photography is also useful for revealing structure of macroscopic objects (Figure 52.3).

Similarly, bodily fluids extracted by needle aspiration can be assayed for suspended microcrystals via birefringence. For example, the presence of monosodium urate crystals can be used to diagnose gout.

52.3.4 How to make a circular polarizing filter

[[ED YONG BOOK.]] In Problem 52.1, you’ll see how to construct a circular polarizing filter by combining a wave plate with a linear polarizer. Remarkably, insects have evolved linear and even circular polarizing filters, probably to signal to others of their species (Figure 52.4).

52.3.5 E. Land’s optical ring sight

Suppose that a biaxial material is cut into a flat plate with its faces perpendicular to its special axis (unlike the half-wave plate). Then light that enters perpendicular to the faces will effectively see an isotropic medium (both transverse indices are equal), and hence will suffer no change of polarization. Suppose first that this plate is sandwiched between two linear polarizing filters, and they are aligned with each other. When we look through this gadget, in the forward direction it behaves simply as a polarizing filter; because our
eyes do not detect polarization, we see an undistorted image of the world. However, away from the forward direction light travels to our eye via rays that impinge at nonzero angles of incidence. Those rays do start to be affected by $\Delta n$, leading to an interference-type
pattern as we look out at larger angles.

E. Land invented a related gadget in which the linear polarizers were replaced by two identical circular polarizers. Once again, in the forward direction the net effect of the gadget is the same as a single circular polarizer. This time, however, at larger angles we see a series of concentric rings of interference. Land’s invention was widely used as a gun sight in the second World War. One advantage over other sights is that the rings appear to be located at infinity, and so may be focused simultaneously with the target.

52.3.6 A medical application

Patients with glaucoma may also suffer deterioration of the optic nerve. Even ophthalmically healthy patients with a family history of glaucoma need to be screened for this. A technique called scanning laser polarimetry uses the birefringence of the optic nerve fiber layer as a noninvasive way to quantify its thickness.

52.4 INDUCED BIREFRINGENCE

52.4.1 Electro-optic (Pockels or Kerr) effects

We have based our understanding of non-vacuum permittivity in the fact that media consist of molecules, and molecules may be deformed slightly by applied fields (Chapter 6). So far, we have always considered only the lowest-order (linear) response of a medium. But the effective springs maintaining the structure of molecules, and their packing into a crystal, will be nonlinear at large enough deformation. That observation leads us to expect that the permittivity tensor may itself be a function of applied field strength: an electro-optical effect.

Indeed, many crystals acquire or change their birefringence upon application of an external field. The change may appear at linear order in the field (Pockels effect). In some cases, this response is forbidden by inversion symmetry of the crystal, but nevertheless birefringence can and does appear at quadratic order (quadratic electro-optical or Kerr electro-optical effect). When combined with polarizing filters, these effects can be used for very fast switching of light (as elements in a light modulator or “shutter”).

52.4.2 Liquid crystals

In the liquid crystal states of matter, individual molecules are mobile (unlike a crystal), their locations are disordered (like a liquid), but their orientations in space are partially ordered (like a crystal). One common example is called nematic: Each molecule has one axis of approximate rotational symmetry (its director), as well as approximate reflection symmetry through the plane perpendicular to that axis. Thus, the director can be viewed as “an arrow without a head”; we can represent it by a unit vector \( \vec{n} \), as long as we remember that \(-\vec{n}\) is an equivalent description.
Figure 52.5: One form of liquid crystal display cell. (a) Without external electric field, the nematic director is constrained by boundary conditions to rotate a quarter turn from bottom to top of the cell. The lower molecules have the same shape as the upper ones, but here are being viewed end-on. The diagram is simplified; actually, the molecules are partially disordered. (b) An applied $\vec{E}$ field realigns the molecules. The director must still align with the two boundaries, but in the interior they transition to being parallel to the applied field.

In a certain temperature range, the directors partially align, so that the average $\langle \hat{m} \otimes \hat{m} \rangle$ becomes nonzero and different from $\frac{1}{3}\hat{1}$. The resulting anisotropy gives the material a permittivity tensor like that of a uniaxial crystal.

In the case of nematics, the state of lowest free energy has spatially constant order parameter, and hence homogeneous $\hat{\mathcal{E}}$. If we force the order parameter to be nonconstant, for example by imposing incompatible boundary conditions on the ends of a chamber, then the material responds with the most slowly-varying configuration compatible with those boundary conditions: rotation with a constant spatial gradient from one boundary to the other (Figure 52.5a).

Because molecular alignment can be controlled by external fields, liquid crystals are useful for making computer displays (LCDs). Figure 52.5 shows one version of the idea. The screen is divided into cells (pixels), each a small chamber filled with a nematic liquid crystal medium and sandwiched between crossed polarizers and two transparent electrodes. In each cell, the front and back walls pin the director in incompatible directions, leading to the twisted arrangement shown in Figure 52.5a. The corresponding spatially-varying permittivity tensor in turn can affect the polarization of light.

**Ex.** Obtain an expression for the permittivity tensor in the situation where the director rotates in the $xy$ plane as we move along $z$.

**Solution:** Suppose that at $z = 0$ the special eigenvector of $\hat{\mathcal{E}}$ points along $\pm \hat{x}$. Then elsewhere it will be rotated through angle $\alpha = \pi z/(2L)$, so that at $z = L$ it has completed a quarter turn as required by the boundary conditions. Thus, we have (Equations 14.2, page 210 or 32.8, page 492)

$$
[\hat{\mathcal{E}}] = \begin{bmatrix}
\cos \alpha & -\sin \alpha \\
\sin \alpha & \cos \alpha \\
0 & 0 & 1
\end{bmatrix} \begin{bmatrix}
\varepsilon_\odot & & \\
& \varepsilon_x & \\
& & \varepsilon_x
\end{bmatrix}
\begin{bmatrix}
\cos \alpha & \sin \alpha \\
-\sin \alpha & \cos \alpha \\
0 & 0 & 1
\end{bmatrix}.
$$

9Typically this range lies between a low-temperature phase (solid, completely ordered) and a high-temperature phase (ordinary liquid, completely disordered).

10The difference is a state variable called the order parameter: a traceless, symmetric 3-tensor field of rank 2.
Abbreviate \( P = (\varepsilon_x + \varepsilon_\perp)/2 \) and \( Q = (\varepsilon_x - \varepsilon_\perp)/2 \). Multiplying matrices shows that the permittivity tensor involves the double angle:

\[

\begin{pmatrix}
  P - Q \cos(2\alpha) & Q \sin(2\alpha) & 0 \\
  Q \sin(2\alpha) & P + Q \cos(2\alpha) & 0 \\
  0 & 0 & P + Q
\end{pmatrix}.

\tag{52.7}
\]

As a check on our math, note that the trace and determinant are everywhere the same as they are at \( z = 0 \). Also, the isotropic limit \( Q \to 0 \) behaves as expected.

You’ll work out the propagation of light in such a medium in Problem 52.3. This form of circular birefringence is different from the one discussed in Section 51.3 (page 714), because it does not involve magnetic cross-susceptibility, and because it relies upon spatial patterning of \( \varepsilon \). However, it again leads to the rotation of linearly polarized light: When the cell is illuminated from one side, light passing through the first polarizer can also pass through the second one.

Applying an electric field from back to front aligns the molecules perpendicular to the screen, overriding the twisted ordering (Figure 52.5b). Now light passing through the cell sees a permittivity that is mostly isotropic in the transverse directions. Its polarization is therefore not rotated, it is blocked by the crossed polarizers, and the pixel appears dark. Only a few volts need to be applied across the cell to switch it, and its response time can be as short as 10–20 ms,\(^{11} \) so the mechanism just described is suitable for a computer display. Masking individual pixels with red, green, or blue color filters gives a color display.

Section 52.4.2\(^{11} \) (page 733) contrasts the optical properties of liquid crystals and ordinary polar liquids.
52.5 OPTICAL TORQUE WRENCH

Section 20.5 (page 303) pointed out that a circularly polarized plane wave in vacuum transports angular momentum; a similar derivation applies in a simple isotropic medium like water. If a bead made of anisotropic material lies in the path of the beam, this chapter has noted that it will affect the beam's polarization, in general converting it to a mixture of the two helicities and hence reducing the angular momentum flux that it carries. The conservation of angular momentum then implies that net angular momentum must be transmitted to the particle itself at some rate. In other words, the beam applies a \textit{torque} to the bead. A typical lab laser, applied to a micrometer scale quartz bead, can in this way apply torque of a magnitude that is relevant for affecting the conformation of a macromolecule attached to the bead: an \textbf{optical torque wrench} effect.

FURTHER READING

\textbf{Semipopular:}

\textbf{Intermediate:}
Peatross & Ware, 2015, chap. 5.
Electro-optical effects: Landau et al., 1984, §100.

\textbf{Technical:}
Circular polarizer in beetle carapace: Sharma et al., 2014; Srinivasarao, 1999; Sharma et al., 2009.

---

\textsuperscript{11}In contrast, Kerr and Pockels cells typically require much higher $\vec{E}$ field, but respond much faster.
52.2’ Magnetic anisotropy

Liquid crystals also exhibit magneto-optical effects.

[[Figure 52.7: “Nematic liquid crystal materials generally contain molecules that possess permanent dipole moments and that also exhibit an anisotropic electronic polarizability and an anisotropic diamagnetic susceptibility.” — Smith et al., 2007]]

52.4.2’ Contrast to polar liquids

A polar liquid, such as water, also consists of anisotropic molecules, which can also be partially aligned by an applied electric field, and yet the optical effect of that alignment is far smaller than that of a nematic. The difference is that the nematic has a nonzero net alignment even at zero applied field. Because that preferred alignment is indifferent to its direction (away from boundaries), it is very easy to realign, with a corresponding large change in the permittivity.

When a liquid crystal approaches its phase transition to an ordinary fluid, the free-energy cost of a gradient in orientation also becomes small. Then even spontaneous thermal fluctuations are able to reorient the director, leading to strong light scattering (opalescence).
52.1 Quarter-wave plate
Review Section 52.3.1 (page 725). A quarter-wave plate is a slab of biaxial dielectric material that has been cut so that its planar faces are perpendicular to one of its three principal axes. Light enters along that direction. Unlike Section 52.3.1 (page 725), however, the thickness of the slab is chosen such that, at a particular wavelength, the transit times for light linearly polarized along the other two principal axes differ by 1/4 of a wave period. Assume that \( \mu = \mu_0 \) and the material is nonchiral.

a. The optical setup in Figure 37.1 involves a beam of light that passes twice through a quarter-wave plate. What is the net effect on the light’s polarization? Does the second passage cancel the effect of the first? Why or why not?

b. Suppose that a linearly polarized plane wave enters this material, with polarization vector oriented midway between the two principal directions (Figure 52.1d or its mirror image). What sort of wave emerges from the other side? Write a short formula to justify your answer.

c. Hence, a combination of linear polarizing filter followed by quarter-wave plate, with the angular orientation just described, can be a useful way to obtain the kind of light you found in (b). What is the effect of this gadget on light passing through it in the other direction (first the quarter-wave plate, then the polarizer), and what might that configuration be useful for?

52.2 Optical torque wrench
[[Not ready]]

52.3 Liquid crystal display pixel
Background: Section 52.4.2 (page 729) introduced nematic liquid crystals, and outlined how to produce a state with spatially varying permittivity tensor. Generally it is hard to solve the Maxwell equations in such a situation, but in the application shown in Figure 52.5 (page 730), \( \varepsilon \) varies on the length scale of the entire chamber, which is far larger than the wavelength of visible light. In such situations, there can be a great simplification, analogous to the eikonal approximation (Section 21.5.3, page 323).\(^{12}\)

Let us make a trial solution for light to propagate along \( +\hat{z} \) with transverse polarization in a medium with spatially-varying permittivity \( \varepsilon(z) \) given by Equation 52.7 (page 731): \( \vec{E}(t,z) = \vec{E}(z) \cos(\omega t) \). Note that we have not yet made any assumption about sinusoidal dependence on \( z \); \( \vec{E} \) is an unknown real function of \( z \).

Problem:

a. Find \( \vec{B} \) as usual via the Faraday law.

b. Show that the Gauss laws are satisfied if \( \vec{E} \) is transverse to \( \hat{z} \), because then Equation 52.7 (page 731) shows that \( \varepsilon \cdot \vec{E} \) will also be transverse.

\(^{12}\)This limit is also similar to the adiabatic limit of a quantum mechanical system with time-dependent hamiltonian.
c. Write down the Ampère law in terms of $\vec{E}$. This will be a system of two ordinary differential equations with nonconstant coefficients.

d. Thus, we cannot expect the solutions to be sinusoidal in $z$. Nevertheless, transform to a rotating coordinate system, by letting $\alpha = \pi z/(2L)$ and

$$\vec{E}(z) = \begin{bmatrix} \cos \alpha & \pm \sin \alpha \\ \mp \sin \alpha & \cos \alpha \end{bmatrix} \begin{bmatrix} \xi(z) \\ \eta(z) \end{bmatrix}.$$  \hspace{1cm} (52.8)

Substitute this representation of the field into the Ampère law. Then make the eikonal-type approximation that, when taking $z$ derivatives, we keep only derivatives of the rapidly-varying $\xi$ and $\eta$, because they will be much larger than terms with derivatives of $\pi z/(2L)$.

e. One choice of signs in Equation 52.8 then diagonalizes the problem. Make that choice, find the resulting approximate dispersion relation, and interpret the two solutions. [Hint: The diagonal property means that if a solution starts out linearly polarized in one principal direction, it will remain in that direction in the rotating frame as it propagates.]
CHAPTER 53

Energy in Media

All there is to thinking is seeing something noticeable which makes you see something you weren’t noticing which makes you see something which isn’t even visible.

— Norman Maclean

53.1 FRAMING: EXTRAORDINARY POLARIZATION

Chapter 49 developed a program begun in Chapter 6, of formulating electrodynamics with simple media in a way that makes very little mention of the media themselves. We did this by encapsulating the response of the medium using material parameters, such as permittivity and permeability. We now continue the program by examining the local conservation of energy and momentum in a similar way.

In a sense, our work is already done, because Chapter 35 already proved the Poynting theorem. Recall that there we partitioned energy density and flux each into a part carried by charged material particles, which was not conserved, plus a second part depending only on fields. We then proved a local continuity equation for the sum of these quantities. But applying that result to fields in media would require us to think about every charge in the medium. Moreover, if for example we wanted the mechanical work required to bring two free charges together, in this approach we would need to calculate and add in the energy stored in (or released from) the elastic deformation of molecules in the medium as it accommodates changes in the charges’ separation.

It will be more convenient, and in line with Chapter 49, to consider a different partition, into energy density and flux for free charges and currents, plus a second part depending on fields and the nature of the medium. However, the goal will be similar to that in Chapter 35: We will be satisfied if we find candidate expressions for the joint energy and momentum of fields+medium that are local, and a mathematical result that when we add the contribution from free charges, the result obeys a continuity equation.

Another difference with Chapter 35 is that the presence of a medium breaks invariance under Lorentz boosts. So there is little to gain by maintaining manifest Lorentz invariance; instead, we will work out separate results for energy and momentum.

Electromagnetic phenomenon: Printed text appears doubled when viewed through a crystal of calcite.

Physical idea: Refraction at the interface with an anisotropic material can split a beam of light into ordinary and extraordinary polarizations.
53.2 LOCAL CONSERVATION, ISOTROPIC MEDIUM

Throughout this chapter, we will simplify by neglecting possible dispersion. Thus, $\varepsilon$ will be taken to be a constant tensor (independent of frequency).

53.2.1 Energy

To keep the formulas manageable, we first restrict to the simplest case: an isotropic, nonmagnetic, nonchiral, nondissipative dielectric medium. That is, $\varepsilon$ will be a constant real scalar, $\mu$ will equal $\mu_0$, and there will be no cross-susceptibilities.

Inside some volume $V$, fields deliver work to the free charges at a rate (power) given by

$$p_t = \int_V \mathbf{E} \cdot \mathbf{J}_f. \quad (53.1)$$

Now consider the quantity

$$\mathbf{\varepsilon} \cdot (\mathbf{E} \times \mathbf{H}) = \varepsilon_{ijk} (\mathbf{E}_j \mathbf{H}_k + \varepsilon_{ijk} \mathbf{E}_j \mathbf{H}_k) = \mathbf{H} \cdot (\mathbf{\nabla} \times \mathbf{E}) - \mathbf{E} \cdot (\mathbf{\nabla} \times \mathbf{H}). \quad (53.2)$$

The Faraday law, and the Ampère law with media give

$$= \mathbf{H} \cdot (-\partial \mathbf{B}/\partial t) - \mathbf{E} \cdot (\partial \mathbf{D}/\partial t + \mathbf{J}_f). \quad (53.3)$$

Combine Equations 53.1–53.3 and use the divergence theorem to find

$$p_t = -\int_V \mathbf{D} \cdot \mathbf{E} \, dt - \int_{\partial V} \mathbf{E} \cdot (\mathbf{\nabla} \times \mathbf{H}). \quad (53.4)$$

Chapter 35 showed that all forms of energy combine into a locally conserved quantity. So any net energy delivered to free charges must come from the fields and/or state changes of the medium. Equation 53.4 therefore suggests that we define an energy flux of fields plus medium (Poynting vector), and a corresponding density, as

$$\mathbf{S}_{\text{med}} = \mathbf{E} \times \mathbf{H} \quad \text{and} \quad \rho_{\varepsilon,\text{med}} = \frac{1}{2} (\varepsilon \mathbf{E}^2 + \mu_0^{-1} \mathbf{B}^2). \quad (53.5)$$

These formulas reduce to the usual ones in the absence of any medium ($\varepsilon = \varepsilon_0$). Next, note that

$$p_t = -\frac{d}{dt} \int_V \mathbf{D} \cdot \mathbf{E} \, dt - \int_{\partial V} \mathbf{E} \cdot \mathbf{S}_{\text{med}}. \quad \text{In words: Fields can do work on free charges by releasing stored field energy, or by transport of field energy from elsewhere.}$$

---

1 Real permittivity means no dissipation (Section 49.2.1, page 701). With dissipation, energy will still be conserved, but some of it will be converted to heat.

2 Power is force dot velocity. Magnetic forces do no work on charges. $\mathbf{J}_f \mathbf{d}^3r$ is the free charge times average velocity of a small chunk of a distribution.

3 Equation 49.8 (page 691).
53.2.2 Momentum

Proceeding similarly, we can also work out the momentum transfer rate (force density) to the free charges, \( \vec{f}_f = \rho_f \vec{E} + \vec{j}_f \times \vec{B} \). If we define a momentum flux of fields plus medium as

\[
\vec{T}_{\text{med}} = \vec{D} \cdot \vec{E} + \vec{B} \cdot \vec{H} - \frac{1}{2} (\vec{B} \cdot \vec{H} + \vec{D} \cdot \vec{E}) \vec{n}
\]

and a corresponding density \( \vec{s}_{\text{med}} = \vec{D} \times \vec{B} \), then

\[
\vec{f}_f = -\frac{\partial}{\partial t} \vec{s}_{\text{med}} + \vec{\nabla} \cdot \vec{T}_{\text{med}}.
\]  (53.6)

Minkowski gave this result in 1908.

Although it is mathematically correct, the physical interpretation of Equation 53.6 is not as straightforward as in the case of energy. For example, a wave passing through a nondissipative medium cannot continually deposit energy there in steady state, but it can continually deposit momentum; to which partition should that momentum be attributed?

M. Abraham considered a different partition of momentum density and flux into “particle” and “other,” leading to a different-looking local continuity equation.\(^4\) One advantage of his scheme was that then the resulting energy–momentum flux 4-tensor for field+medium was symmetric.

53.3 ENERGY IN ANISOTROPIC MEDIA

53.3.1 Poynting vector

**Ex.** Generalize the derivation of Equations 53.5 to a medium that is again linear, uniform, and nonmagnetic, but this time anisotropic (Equation 52.1, page 724). What now are the appropriate formulas for \( \rho_{\varepsilon,\text{med}} \) and \( \vec{S}_{\text{med}} \)?

**Solution:** Equations 53.1–53.4 are unchanged. Eliminating \( \vec{D} \) and \( \vec{H} \) now give a continuity equation with \( \rho_{\varepsilon,\text{med}} = \frac{1}{2} \hat{\varepsilon} \cdot \vec{E} + \frac{1}{2} \hat{\mu}^{-1}\varepsilon \vec{B} |\vec{B}|^2 \) and \( \vec{S} = \vec{E} \times \vec{H} \).

For nonmagnetic materials, the Poynting vector has the same form as the one in vacuum.

53.3.2 Double Refraction

Section 52.3.1 (page 725) introduced the half-wave plate, a useful optical device. The analysis involved a special case of light crossing from vacuum into an anisotropic dielectric medium:

- The light arrived traveling perpendicular to the interface.
- The crystalline material was cut in such a way that one of its principal axes was also perpendicular to the interface.

\(^4\)See Further Reading.
When we relax the second of those conditions, a remarkable new phenomenon can appear.

The mineral calcite is a form of calcium carbonate that, in pure form, is transparent to light but has uniaxial anisotropy. E. Bartholin noticed in 1670 that unlike in many transparent materials, images viewed through a large single crystal of calcite appear doubled (Figure 53.1). More precisely:

- The effect requires a slab of calcite cut along planes that are not perpendicular to any principal direction of the permittivity.
- When the object being viewed is illuminated with unpolarized light, the two images are both strongly polarized; one or the other may be eliminated by viewing through an appropriately oriented polarizing filter.\(^5\)
- One image (the “ordinary image”) appears where we would expect to see it through glass; for example, when viewing along the perpendicular to the slab faces, the light does not appear to bend.
- The other image appears to be displaced relative to the first one, but is otherwise undistorted. Rotating the crystal, without moving the viewed object, rotates the displacement of this image while the ordinary image maintains its apparent position (Figure 53.1).

Indeed, we can follow the path of light by shining a laser into the crystal and observing scattering off its imperfections or fluorescence from impurities (Figure 53.2). What we find is that, even when the light arrives perpendicular to its surface, it splits into two beams, *one of which bends* upon entry and subsequent exit. That is extraordinary behavior indeed; in contrast, the usual “law” of refraction claims that angle of incidence equal to zero leads to an angle of refraction also equal to zero, regardless of the material’s refractive index.\(^6\)

To understand this seemingly unlawful behavior, we can do a matching calculation similar to the one we used for total internal reflection,\(^7\) but with some differences:

- We now consider light that enters the material (instead of trying to exit).

---

\(^5\)See Media 22.


\(^7\)See Section 49.4.2 (page 697).
Figure 53.2: Passage of light through birefringent medium. A violet laser pointer (400 nm, left) was shone through a single crystal of calcite. Two spots of violet light emerge (right). The specimen contained fluorescent impurities that allow us to follow the light beam’s progress through the crystal. Fluorescence shifts that wavelength down to the peach color seen here and scatters some of it sideways to the viewer. The path length was 8.5 cm. Nothing was changed in the second photo except to turn out room lights. See also Media 22. [Realization and photographs by William Berner and Jax Gottschalch.]

Figure 53.3: Refraction upon entering an anisotropic medium. A wave travels through vacuum (lower left), arriving perpendicular to the planar surface of an anisotropic crystal. Coordinates have been chosen to simplify the permittivity tensor, not the planar surface. In the “ordinary wave” situation, the incoming light is polarized along \( \hat{x} \), a principal direction of \( \hat{\varepsilon} \) (perpendicular to the page). The “extraordinary wave” situation has the other allowed polarization, perpendicular both to \( \hat{k} \) and to \( \hat{x} \) (and hence not a principal direction).

- The material is now anisotropic. For simplicity, we consider only uniaxial anisotropy (two eigenvectors of the permittivity are equal), which is appropriate for calcite.
- We will choose cartesian coordinates that diagonalize the permittivity tensor:

\[
\hat{\varepsilon} = \begin{bmatrix} \varepsilon_x & 0 \\ 0 & \varepsilon_\perp \end{bmatrix}.
\]

So the boundary of the material will be some arbitrary plane, not the \( xy \) plane (Figure 53.3).
- For simplicity, we will only consider the case of perpendicular incidence (angle of incidence = 0).
- At the end, we will need to consider energy flux (the Example on page 738) to interpret the result.
Suppose that a wave propagates in vacuum with frequency $\omega$. By the axial symmetry about $\hat{z}$, rotate the coordinate axes to arrange that $\vec{k}$ lies in the $yz$ plane (Figure 53.3):

$$\vec{k} = k(\hat{y} \sin \theta + \hat{z} \cos \theta) \quad \text{where} \quad k = \omega/c.$$  

The boundary is perpendicular to $\vec{k}$, so we have vacuum where $\vec{k} \cdot \vec{r} < 0$ and calcite where $\vec{k} \cdot \vec{r} > 0$. We expect partial reflection at the interface. At perpendicular incidence, the reflection will also be perpendicular, that is, wavevector $-\vec{k}$. Thus, we consider the trial solution

$$\vec{E} = \frac{1}{2} \left[ \vec{E}_0 e^{i(-\omega t + \vec{k} \cdot \vec{r})} + \vec{E}_0^* e^{i(-\omega t - \vec{k} \cdot \vec{r})} + \text{c.c.} \right] \quad \text{for} \quad \vec{k} \cdot \vec{r} < 0. \quad (53.7)$$

The formula just given is a solution to Maxwell in vacuum because $k = \omega/c$. We wish to know if we can extend it to a solution everywhere, obeying the relevant boundary conditions, via a transmitted wave of the generic form

$$\vec{E} = \frac{1}{2} \left[ \vec{E}_t e^{i(-\omega t + \vec{k}' \cdot \vec{r})} + \text{c.c.} \right] \quad \text{for} \quad \vec{k} \cdot \vec{r} > 0, \quad (53.8)$$

with appropriate polarization $\vec{E}_t$ and direction $\vec{k}'$. Thus, given $\vec{E}_t$, we will seek $\vec{E}_r$, $\vec{E}_t$, and $\vec{k}'$ that satisfy the boundary conditions.

**Ordinary polarization**

It is not hard to guess that the easiest solution will be the case of incoming polarization along $\hat{x}$, because that direction is transverse to $\vec{k}$ (hence parallel to the interface) and also a principal direction of $\vec{\varepsilon}$. Here the anisotropy will be irrelevant, so prior experience leads us to guess further that

$$\vec{E}_t = \hat{x} E_t, \quad \vec{E}_r = \hat{x} E_r, \quad \vec{E}_t = \hat{x} E_t, \quad \text{and} \quad \vec{k}' = \vec{k} \sqrt{\varepsilon_x/\varepsilon_0}. \quad \text{ordinary polarization} \quad (53.9)$$

Certainly with these choices the transmitted wave is a solution to Maxwell in the medium. But we must also confirm that we can choose $\vec{E}_r$ and $\vec{E}_t$ in a way that satisfies the boundary conditions at the plane $\vec{k} \cdot \vec{r} = 0$, and hence also $\vec{k}' \cdot \vec{r} = 0$. The condition $\Delta \vec{E}_\parallel = 0$ implies that $\vec{E}_t - \vec{E}_r = \vec{E}_t$ and also that $\vec{k}'$ must be parallel to $\vec{k}$ as already assumed (Equation 53.9). The condition $\Delta D_{\perp} = 0$ is automatically satisfied because $\vec{D} = \varepsilon_x \vec{E}$ has no component perpendicular to the boundary.

Next, we get the corresponding magnetic fields via the Faraday law:

$$\vec{B}_t = (k/\omega)(-\hat{z} \sin \theta + \hat{y} \cos \theta) \vec{E}_t \quad (53.10)$$

$$\vec{B}_r = (-k/\omega)(-\hat{z} \sin \theta + \hat{y} \cos \theta) \vec{E}_r \quad (53.11)$$

$$\vec{B}_t = (k'/\omega)(-\hat{z} \sin \theta + \hat{y} \cos \theta) \vec{E}_t. \quad (53.12)$$

Then the condition $\Delta \vec{B} = 0$ amounts to $-\vec{E}_t + \vec{E}_r = -\sqrt{\varepsilon_x/\varepsilon_0} \vec{E}_t$. We see that all boundary

---

8Instead of guessing, we could have used the systematic procedure in the next subsection.
conditions can then be met with the choices\(^9\)
\[
E_x = E_1 \frac{\sqrt{\varepsilon_0 - \varepsilon_x}}{\sqrt{\varepsilon_0} + \sqrt{\varepsilon_x}}, \quad E_T = 2E_1 \left(1 + \sqrt{\varepsilon_x/\varepsilon_0}\right)^{-1}.
\]
Thus, we have a solution with wavevector parallel to the incoming wave and with the same linear polarization (Equation 53.9), called the ordinary wave solution. Note that for the special case where \(\varepsilon_x = \varepsilon_0\) (no medium), we get complete transmission as expected.

**Extraordinary polarization**

The incident and reflected waves both have electric field that is constant over the interface, because we assumed perpendicular incidence. To match them, the transmitted wave must also have that property, so \(\vec{k}'\) must also be perpendicular to the interface. Hence, we must again have that \(\vec{k}'\) is parallel to \(\vec{k}\). We now ask, what plane waves are possible in the medium, more generally than the one already guessed for the ordinary wave?

The transversality condition looks simplest when expressed in terms of the displacement, not the electric field, so let us switch to describing the transmitted wave by the complex amplitude of its displacement: That is, replace Equation 53.8 by a trial solution of the form
\[
\vec{D}_T = \frac{1}{2} \eta \vec{E}_0 (\vec{k}' + \vec{k}'') + \text{c.c.}
\]
We can always recover \(\vec{E}_T\) in this language as \(\vec{\varepsilon}^{-1} \cdot \vec{\eta}\). Then the electric Gauss law \(\vec{V} \cdot \vec{D}_T = 0\)

simple says that the complex amplitude \(\vec{\eta}\) must be perpendicular to \(\vec{k}'\), and hence to \(\vec{k}\) as well.

As usual, we can obtain the dispersion relation by combining the Faraday and Ampère laws.\(^10\) The Faraday law says
\[
-(\omega \vec{B}_T) = i\vec{k}' \times (\vec{\varepsilon}^{-1} \cdot \vec{\eta} - \vec{\varepsilon}^{-1} \cdot \vec{\eta}) \quad \text{(53.13)}
\]
The Ampère law says
\[
-i\omega \mu_0 \vec{\eta} = i\vec{k}' \times \vec{B}_T.
\]
Eliminating \(\vec{B}_T\) then gives
\[
-\frac{\omega^2 \mu_0}{(k')^2} \vec{\eta} = \left[\vec{k} (\vec{k} \cdot (\vec{\varepsilon}^{-1} \cdot \vec{\eta})) - \vec{\varepsilon}^{-1} \cdot \vec{\eta}\right] \quad \text{(53.14)}
\]

Equation 53.14 is an eigenvector problem on the two-dimensional space of \(\vec{\eta}\) vectors transverse to \(\vec{k}\). We already know one solution: the ordinary wave, with \(\vec{\eta}\) along \(\vec{x}\). That suggests that our problem would be simpler if reexpressed in a new orthonormal basis:
\[
\vec{\eta} = A\vec{k} + B\vec{x} + C\vec{k} \times \vec{x}.
\]

---

\(^9\)This relation is another special case of the general refraction formulas (Fresnel formulas) mentioned in Chapter 49. The ratios \(E_R/E_1\) and \(E_T/E_1\) were previously called the reflection and transmission factors (\(R\) and \(T\)), respectively.

\(^10\)See Equations 52.3–52.4 (page 725).
To find the three components of the right side of Equation 53.14 in the new basis, we take dot products:

\[ \hat{k} \cdot \left[ \hat{k} \cdot \left( \hat{\varepsilon}^{-1} \cdot \vec{n} \right) \right] - \hat{\varepsilon}^{-1} \cdot \vec{n} = 0 \]
\[ \hat{x} \cdot \left[ \hat{k} \cdot \left( \hat{\varepsilon}^{-1} \cdot \vec{n} \right) \right] = -\vec{x} \cdot \vec{n} \]
\[ (\hat{k} \times \hat{x}) \cdot \left[ \hat{k} \cdot \left( \hat{\varepsilon}^{-1} \cdot \vec{n} \right) \right] = -(\hat{k} \times \hat{x}) \cdot \hat{\varepsilon}^{-1} \cdot \vec{n} \]

Evaluating then expresses Equation 53.14 as

\[ \frac{\omega^2 \mu_0}{(k')^2} \begin{bmatrix} A \\ B \\ C \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -\varepsilon_x & 0 \\ \text{etc.} & 0 & -((\varepsilon_x)^{-1} \cos^2 \vartheta + (\varepsilon_\Theta)^{-1} \sin^2 \vartheta) \end{bmatrix} \begin{bmatrix} A \\ B \\ C \end{bmatrix} \]

The first row says \( A = 0 \) (no longitudinal polarization). The second reproduces Equation 53.9. The third gives the new dispersion relation we were seeking:

\[ \mu_0 \omega^2 = (k')^2 ((\varepsilon_\Theta)^{-1} \sin^2 \vartheta + (\varepsilon_x)^{-1} \cos^2 \vartheta) \quad \text{extraordinary wave} \quad (53.15) \]

As a check, note that for an isotropic medium, this is a familiar equation. More generally, however,

At a given frequency, the extraordinary wave’s wavelength, \( 2\pi / ||\vec{k}'|| \), depends on direction.

We now ask whether \( \vec{E}_i \) and \( \vec{E}_r \) can be chosen to satisfy the boundary conditions, with \( \vec{n} \) parallel to \( \hat{k} \times \hat{x} \) and \( \hat{k}' \) from Equation 53.15. The condition \( \Delta D_\perp = 0 \) is again automatically satisfied \( (\hat{D} \cdot \hat{k} = 0 \text{ on both sides}) \). The parallel electric field condition says

\[ \vec{E}_i + \vec{E}_r = [\hat{\varepsilon}^{-1} \cdot \vec{n}]_\parallel = \eta((\varepsilon_x)^{-1} \cos^2 \vartheta + (\varepsilon_\Theta)^{-1} \sin^2 \vartheta)(\hat{y} \cos \vartheta - \hat{z} \sin \vartheta) \]

As with the ordinary wave, \( \vec{E}_i \) and \( \vec{E}_r \) must both be perpendicular to \( \vec{k} \). The right side of the preceding equation also has that property, so it amounts to two equations in two unknowns (the transverse components of \( \vec{E}_i + \vec{E}_r \)).

**Your Turn 53A**

Show that similarly, the condition \( \Delta \vec{B} = 0 \) amounts to two equations in the two transverse components of \( \vec{E}_i - \vec{E}_r \), and that therefore our trial solution can be arranged to satisfy all the boundary conditions.

### 53.3.3 Energy flow in double refraction

To summarize, we considered a trial solution that, like the ordinary wave, again has wavevector \( \vec{k}' \) perpendicular to the surface. We found that such a solution (the extraordinary wave) does exist, but that result may seem disappointing, because we were trying to understand double refraction. To get more insight, we now look more closely at the energy flow.
Ordinary wave
The energy flux of the ordinary wave is straightforward: Section 53.2.1 along with Equations 53.9 and 53.12 give
\[
\vec{S}_\tau = \vec{E}_\tau \times \vec{B}_\tau / \mu_0 = \frac{1}{4} (\vec{E}_\tau \cdot \vec{e} \exp(\cdot \cdot \cdot) + \text{c.c.}) \times (k'/\omega)(\vec{E}_\tau (\mathbf{\hat{e}} \sin \theta + \hat{y} \cos \theta) \exp(\cdot \cdot \cdot) + \text{c.c.}) / \mu_0.
\]
The time-average is\(^\dagger\)
\[
\langle \vec{S}_\tau \rangle = \frac{1}{2\mu_0\omega} \vec{k}' |\vec{E}_\tau|^2. \tag{53.16}
\]
Energy flows perpendicular to the surface.

Your Turn 53B
Show that the net energy flux to the interface is zero.

Extraordinary wave
This time, the steps that previously led to Equation 53.16 give
\[
\frac{1}{4\mu_0\omega} \vec{E}_\tau \times (\vec{k}' \times \vec{E}_\tau^*) + \text{c.c.}
\]
\[
= \frac{k'}{4\mu_0\omega} (2\mathbf{\hat{k}}(\vec{E}_\tau \cdot \vec{E}_\tau^*) - \vec{E}_\tau^*(\vec{k} \cdot \vec{E}_\tau) - \vec{E}_\tau(\vec{k} \cdot \vec{E}_\tau^*)).
\]
The first term is parallel to \(\mathbf{\hat{k}}\), but the other two need not be. Note that the factor in parentheses is \(\eta k' \sin \theta \cos \theta((\epsilon_\perp)^{-1} - (\epsilon_\parallel)^{-1})\) or its conjugate, which is generally nonzero unless the medium is isotropic.

Your Turn 53C

a. Use Equation 53.13 to show that
\[
\langle \vec{S}_\tau \rangle = \frac{1}{2\mu_0\epsilon} |\eta|^2 (k'/k)((\epsilon_\perp)^{-1} \sin^2 \theta + (\epsilon_\parallel)^{-1} \cos^2 \theta) \begin{bmatrix} 0 \\ (\epsilon_\parallel)^{-1} \sin \theta \\ (\epsilon_\parallel)^{-1} \cos \theta \end{bmatrix}. \tag{53.17}
\]
(First confirm that the units work out.)

b. Which way does this vector point in the special case \(\theta = 0\)? Discuss why that result is expected.

c. Also discuss the special case where \(\epsilon_\parallel = \epsilon_\perp\).

For isotropic media (glass), we see that the energy transport is again perpendicular to the surface (and along \(\mathbf{\hat{k}}\)). But in general, your result implies that the extraordinary wave carries energy in a different direction from \(\mathbf{\hat{k}}\), an example of a more general phenomenon.

\(^\dagger\)Again, the method of the Example on page 576 is useful.
sometimes called “Poynting vector walkoff.” This result is the ultimate origin of the apparent violation of the “law” of refraction visible in Figure 53.212 (and also Figure 53.1, page 739).

When the extraordinary wave eventually encounters the other parallel face of the crystal, it bends by the opposite amount upon exiting. The net effect of these two bends is to shift the image carried by the extraordinary wave. The amount and direction of the shift depend on the internal structure, cut, thickness, and orientation of the crystal, as is observed.

FURTHER READING

Intermediate:
Double refraction: Peatross & Ware, 2015, chap. 5; Zangwill, 2013, chap. 17.

12The fluorescence seen in the figure marks out the region with nonzero energy density.
CHAPTER 54

Waves in a Cold Plasma and Faraday’s Rotation Effect

Who knows then, but there may be, as the Antients thought, a region of this fire above our atmosphere…. Perhaps the Aurorre Boreales are currents of this fluid in its own region, above our atmosphere, becoming from their motion visible. There is no end to conjectures.

— Benjamin Franklin

54.1 FRAMING: FARADAY ROTATION

Maxwell was not the first to intuit a connection between electromagnetism and light. Faraday and others devoted a lot of experimental effort to seeking an effect of electric fields on light. Those researches were unsuccessful; they required more sensitive instruments, or larger field strengths, than what Faraday possessed. But eventually Faraday turned to looking for an effect of magnetic fields on light, following a prediction by J. Herschel. Here his persistence was rewarded in 1845, near the end of his career, with the discovery of the “magneto-optical Faraday rotation.”

One reason that electromagnetic effects on light are hard to observe is that the Maxwell equations in vacuum are linear: A wave can simply be superposed with a background field as it passes into it from a field-free region, with no change to its character. For that reason, Faraday found nothing until he considered light passing through a transparent medium: There he found that adding a background magnetic field would rotate the polarization of light traveling along $\vec{B}$, even if the medium itself was not chiral. For simplicity, this chapter will mainly study the situation where the medium consists of a plasma, but glass, water, oil, and so on all show the effect to varying degrees.

Far from being a historical curiosity, the Faraday’s effect is used today to give us evidence of strong magnetic fields in astrophysical objects, and in other research areas as well. For example, the synchrotron radiation from an accretion disk is polarized, and so this rotation can be used to disclose strong magnetic fields in plasma regions that intervene between the disk and an observer.

*Electromagnetic phenomenon:* When viewed with a polarizer, the synthesized image of a

---

1The first electro-optical nonlinear effect was ultimately seen by J. Kerr in 1875 (Section 52.4.1, page 729).
2Unlike the Zeeman effect (Problem 18.7, page 293), which involves one atom at a time and requires very large magnetic fields to be visible, we’ll see that the Faraday rotation accumulates over many atoms and hence can be seen with field strengths available in Faraday’s day.
black hole shows fine structure in its polarization.

Physical idea: Strong magnetic fields give rise both to synchrotron radiation, and to Faraday rotation of the resulting polarized light.

54.2 APPROXIMATIONS

A plasma is a partially (or fully) ionized gas, but the ideas in this chapter also apply to any substance in which some charge carriers move freely. In general, such systems are complicated; all of the statistical mechanics of gases gets combined with all the intricacies of long-range electromagnetic interactions. We will study a limiting case called cold plasma, in which all thermal motion of the charges may be neglected. For example, a fluorescent light bulb remains cool to the touch; the passage of an electric arc, not thermal collisions, keeps the atoms ionized. If thermal motion is negligible, then there is no gas pressure, no frictional drag on particles, and so on.

As further approximations, Section 54.3 will consider only low-amplitude electric and magnetic fields (a linearized approximation); Section 54.4 will then consider low-amplitude fluctuations superimposed on a uniform and time-independent magnetic field, whose magnitude may not be small. These disturbances propagate in a medium that consists at least partly of free electrons and a neutralizing background of their partner ions; nonionized atoms, if any, will be neglected. In fact, we will also neglect the dynamics of the sluggish ions, treating them as a uniform neutralizing background and focusing solely on the electrons.

54.3 DISPERSION RELATION FOR TRANSVERSE WAVES

54.3.1 Induced free current obeys a nondissipative response function

Our situation is spatially and temporally translation invariant, and we are linearizing, so we may expect solutions of exponential form as we have encountered many times before:

\[ \vec{E}(t, \vec{r}) = \frac{1}{2} \vec{E} e^{-i(\omega t + \vec{k} \cdot \vec{r})} + \text{c.c.,} \quad \vec{B}(t, \vec{r}) = \frac{1}{2} \vec{B} e^{-i(\omega t + \vec{k} \cdot \vec{r})} + \text{c.c.} \]

Let us suppose further that the complex polarization vectors \( \vec{E} \) and \( \vec{B} \) are both perpendicular to \( \vec{k} \); if no such solutions exist, we will discover that when we try to satisfy the Maxwell equations. These trial solutions have the convenient feature that \( \vec{\nabla} \cdot \vec{E} = 0 \) and \( \vec{\nabla} \cdot \vec{B} = 0 \) automatically.

We must extend our trial solution by accounting for what the electrons are doing. Begin with one representative electron, moving near a position \( \vec{r} \) and subject to the Lorentz force law. Its equation of motion is linear in the electric and magnetic fields, so it’s reasonable to write a trial solution containing a single angular frequency:

\[ \vec{r}_{\text{el}}(t) = \vec{r} + \frac{1}{2} \vec{r} e^{-i(\omega t + \vec{k} \cdot \vec{r})} + \text{c.c.} \]

\(^3\)Longitudinal waves can also propagate, analogously in some ways to ordinary sound.
The electron is assumed to feel others only via a mean field that they create,\(^4\) which is included in \(\vec{E}\) and \(\vec{B}\). Its motion stays nonrelativistic because fields are weak, so we may use newtonian mechanics and also neglect the magnetic force. Thus, the electron moves in the plane of \(\vec{E}.\) \(\vec{E}\) is itself constant over this plane, because we assumed it to be perpendicular to \(\vec{k}\).

The excursion \(\vec{r}\) therefore satisfies
\[
(-i\omega)^2 m_e \vec{r} = q \vec{E},
\]
where \(q = -e\) is the electron charge. Solving for \(\vec{r}\) and taking a time derivative yields the velocity:
\[
\vec{v}(t) = \frac{1}{i \omega} \frac{i q}{m_e} \vec{E} e^{-i \omega t + \vec{k} \cdot \vec{r}} + \text{c.c.}
\]
The electron executes motion in a plane perpendicular to \(\vec{k}\); that is, \(\vec{k} \cdot \vec{r} = 0\). Conveniently, the fields of our assumed plane wave are all constant over any such plane.

We now assume that many electrons, with number density \(n_e\), are all doing this dance, each with the phase appropriate to its plane of motion. They create a charge flux \(\vec{j} = q \vec{v} n_e\), or
\[
\vec{j} = \frac{1}{2} i n_e \left( \frac{q}{m_e \omega} \vec{E} \right) e^{-i \omega t + i \vec{k} \cdot \vec{r}} + \text{c.c.}
\]
This relation superficially resembles that in an ohmic material (Equation 8.7, page 119), but there is a crucial difference:

- **We did not find that charge flux \(\vec{j}\) (a real vector) is a (real) constant times electric field \(\vec{E}\) (a real vector).**
- **Rather, we found that the complex amplitude \(\vec{j}\) is a complex constant times the complex amplitude \(\vec{E}\). Equivalently, the charge flux is phase shifted by \(\pi/2\) relative to the electric field.**

Had the constant of proportionality in the second statement been real, that would have implied the first statement; but in our case, the constant in Equation 54.2 is purely imaginary. People often abbreviate by saying, “The conductivity is pure imaginary,” but that is an abuse of language. Conductivity reflects a dissipative process, whereas our zero-temperature, collisionless plasma has no dissipation:\(^5\)

**Your Turn 54A**

a. Show that, unlike in an ohmic material, the time-averaged power dissipation per volume equals zero.

b. Show that \(\nabla \cdot \vec{j} = 0\), and hence that it was self-consistent for our trial solution to assume that electron density is constant.

(Other waves, involving compression, may also exist. Our transverse trial solution is sometimes called “electromagnetic” to distinguish it from these “acoustic” modes.)

---

\(^4\) Mean field approximation was introduced in Section 10.3.3 (page 146).

\(^5\) Conversely, when a quantity that normally represents a nondissipative effect, for example permittivity, is assigned an imaginary part, that shorthand signals a dissipative process (Section 49.2.1, page 701).
54.3.2 The dispersion relation has a cutoff

The next steps are familiar.6 We already know that our trial solution satisfies the electric Gauss law. Imposing Faraday’s law as usual says that

\[ \vec{B} = (\vec{k} / \omega) \times \vec{E}, \]

which automatically obeys the magnetic Gauss law. Ampère’s law in a conductive medium gives

\[ i\vec{k} \times \vec{B} = \mu_0 (i \frac{q^2 n_e}{m_e \omega} + \epsilon_0 (-i \omega)) \vec{E}. \]

Substituting gives

\[ i\vec{k} \times ((\vec{k} / \omega) \times \vec{E}) = -\mu_0 \omega (\epsilon_0 - \frac{q^2 n_e}{m_e \omega^2}) \vec{E}. \] (54.4)

Expanding the triple cross product and using the assumed transversality gives the dispersion relation: The trial solution indeed solves the Newton+Maxwell equations if

\[ ||\vec{k}||^2 / \omega^2 = c^{-2} \left( 1 - \frac{q^2 n_e}{\epsilon_0 m_e \omega^2} \right). \] (54.5)

Some abbreviations and comments are in order. First, define the plasma frequency as

\[ \omega_p = |q| \sqrt{\frac{n_e}{\epsilon_0 m_e}}. \] (54.6)

Then Equation 54.4 says that for transverse waves, the plasma behaves in some ways as a dielectric medium. Unlike an ordinary dielectric (bound electrons), however, its permittivity is less than \( \epsilon_0 \),\(^7 \)

\[ \epsilon = \epsilon_0 (1 - (\omega_p / \omega)^2). \]

Hence, the phase velocity of our waves is:

\[ v_{ph} = \omega / k = c / \sqrt{1 - (\omega_p / \omega)^2}. \] (54.7)

For frequencies above the plasma frequency, this is greater than \( c \); that is, the refractive index \( v_{ph} / c \) is less than one, unlike any ordinary dielectric. However, in that high-frequency regime the index is real: Waves propagate without loss.8 The dependence on frequency means that propagation is highly dispersive for frequencies close to \( \omega_p \).

The situation gets more interesting at frequencies below the plasma frequency: Here \( e^{i\vec{k} \cdot \vec{r}} \) is exponentially damped. When a wave in this low-frequency regime, traveling in vacuum, impinges on the plasma, it cannot penetrate far. Nor is it converted to heat, because there is no dissipation; instead, it must reflect. Section 21.5.6 (page 327) pointed out that this effect is responsible for “skip” (skywave transmission) of shortwave radio signals.

Velocities greater than \( c \) may make us concerned about causality:

---

6See for example Problem 18.8 (page 294).
7See Equation 6.12 (page 81).
8Recall Your Turn 54Aa.
Your Turn 54B

a. Check how the units work in Equation 54.6.

b. Work out the group velocity $v_g = (d/d\omega)^{-1}$ from the dispersion relation as a function of angular frequency, compare with the phase velocity $\omega/k$, and comment.

54.3.3 Earth’s ionosphere permits the passage of cosmic messengers

Earth’s ionosphere (once called the “Heaviside layer”) contains many ions, mostly created by hard ultraviolet light from the Sun, with number density $n_e \approx 10^{11} \text{ m}^{-3}$ and hence $\omega_p/(2\pi) \approx 3 \text{ MHz}$. So aliens won’t be able to monitor our AM radio broadcasts, probably a good thing. On the other hand, luckily the peak in the cosmic microwave background radiation is safely above the cutoff at $\omega_p$, so we can observe it from Earth.

In addition, the mean electron spacing $n_e^{-1/3}$ is much smaller than wavelength of radio-frequency radiation, supporting our continuum treatment of current in the Maxwell equations.\(^9\)

54.3.4 Scintillation of compact astrophysical radio sources

When we look at stars, they appear to twinkle. The effect is due to space- and time-dependent fluctuations in air density, which refract light passing through Earth’s atmosphere, randomly modifying what our eyes collect from distant objects on time scales that we can notice. A similar phenomenon can occur with radio signals from distant objects as they pass through the turbulent solar wind environment of our solar system (via the density dependence of Equation 54.7), leading to interplanetary scintillation (IPS).

S. J. Bell was assigned to search for IPS signatures of quasars in 1967. She found them, but, alert to the unexpected, she also discovered objects that pulse radiation on the time scale of seconds, now known as pulsars.\(^10\) It took some courage to suggest that such “impossible” signals were anything other than manmade. But Bell noticed that the signals came from a fixed sky location once per sidereal (not solar) day, pointing to an extraterrestrial source.

54.3.5 Interstellar plasmas transmit signals with dispersion

Pulsars send out a narrow, rotating searchlight beam of electromagnetic radiation. We at Earth intercept a tiny angular window, so we might expect to receive crisp, periodic pulses of electromagnetic energy. Instead, we get a “chirp,” with higher Fourier components arriving first, followed by the lower ones (Figure 54.1). This dispersion of the signal serves as a useful diagnostic of the medium intervening between the radiation source and us. Typical plasma frequencies in interstellar space fall below the observed radio frequency

\(^9\) Although the Debye screening length is zero at zero temperature, in a more realistic treatment it is $\approx 2 \text{ mm}$, which is still much bigger than $n_e^{-1/3}$. So if we wished to include screening, we could use the continuum approximation outlined in Section 10.3.4 (page 156).

\(^10\) Later, even much faster pulsars were discovered.
range, so the inverse group velocity given by Equation 54.5 can be written approximately as:

\[
c/\nu_g = c \frac{dk}{d\omega} = \omega(\omega^2 - \omega_p^2)^{-1/2} \approx \left(1 + \frac{\omega_p^2}{2 \omega^2}\right).
\]

Then total transit time at angular frequency \(\omega\) for an object at distance \(L\) is

\[
\frac{L}{c} + \frac{1}{2c\omega^2} \int_0^L dx \frac{n_e(x)e^2}{m_e\varepsilon_0},
\]

whose frequency dependence is related to the optical depth (also called dispersion measure), \(\int dx n_e(x)\).

Conversely, if we have an estimate for electron density along the line of sight, then finding the dispersion measure gives an estimate of the source’s distance.

### 54.3.6 Metals

Although the conduction of electricity through metals is quantum-mechanical in character, metals do have delocalized electrons that freely wander among immobile nuclei. And indeed qualitatively they have the property of completely reflecting light at low frequencies, while becoming partially transparent at high frequency, like a plasma. For the simplest metals, such as lithium or sodium, the vacuum wavelength corresponding to the cutoff is around \(2\pi c/\omega \approx 200\) nm (hard ultraviolet).

### 54.4 FARADAY’S MAGNETO-OPTICAL EFFECT

#### 54.4.1 A plasma becomes a chiral medium in a steady magnetic field

We now repeat the derivations of Section 54.3, but this time add a uniform and time-independent background magnetic field \(\vec{B}_0\) to the trial solution. To simplify the math, we consider only waves propagating along (or opposite to) \(\vec{B}_0\), for example,

\[
\vec{k} = k\hat{z}, \quad \vec{B}_0 = B_0\hat{z},
\]
where the scalar $k$ is positive but $B_0$ may have either sign.

We still consider superposing small wavelike perturbations $\vec{E}$ and $\vec{B}$, but $B_0$ itself may not be small. Hence, although we continue to neglect the effect of $\vec{B}$ on the nonrelativistic electron motion, we must keep $B_0$ in the Lorentz force law: Equation 54.1 becomes

$$(-\omega^2)m_e\vec{\Gamma} = q(\vec{E} + \vec{\sigma} \times \vec{B}_0),$$

so

$$\vec{\Gamma} = -\frac{q}{m_e\omega^2}(\vec{E} - i\omega B_0 \vec{\sigma} \times \vec{A}).$$

(Eq. 54.8)

We are familiar with this equation in the absence of any wave: Electrons undergo cyclotron motion.\(^{11}\) The symmetry of that solution under combined time shift and rotation suggests that it would be fruitful to specialize our trial wave solutions to ones with similar symmetry, that is, to circularly polarized waves. Accordingly, we construct a trial solution in which the complex polarization $\vec{E}$ points along one of the two circular basis vectors,\(^{12}\) for example:

$$\vec{E} = P\hat{\xi}_{(\pm)}$$

where $\hat{\xi}_{(\pm)} = (\hat{x} \pm i\hat{y})/\sqrt{2}$.\(^{[18.34, page 285]}\)

Equivalently, we may state the components of $\vec{E}$: $\vec{E}_+ = P$, $\vec{E}_- = 0$.

In Problem 18.6, you showed the useful identity that the circular basis vectors are eigenvectors of the operation $\vec{z} \times$, that is,

$$\vec{z} \times \hat{\xi}_{(\pm)} = \mp i\hat{\xi}_{(\pm)}.$$  

(Eq. 54.9)

Equation 54.8 then separates into two decoupled equations:

$$-\frac{m_e\omega^2}{q}\vec{\Gamma}_+ = P + i\omega B_0(-i)\vec{\Gamma}_+$$

(Eq. 54.10)

and

$$-\frac{m_e\omega^2}{q}\vec{\Gamma}_- = i\omega B_0(i)\vec{\Gamma}_-.$$  

The second of these equations clearly has $\vec{\Gamma}_- = 0$ as a solution. Equation 54.10 is more interesting:

$$\vec{\Gamma}_+ = -P\left(\frac{m_e\omega^2}{q} + \omega B_0\right)^{-1}.$$  

(Eq. 54.11)

Proceeding as in Section 54.3.1 gives the charge flux. In terms of the plasma frequency,\(^{13}\) we now find

$$\vec{j} = \frac{1}{2}(i\omega p^2\varepsilon_0)(\omega + B_0 q/m_e)^{-1}P\hat{\xi}_{(\pm)} e^{-i\omega t + ik \cdot \vec{r}} + \text{c.c.},$$

(Eq. 54.12)

which reassuringly reduces to Equation 54.2 when $B_0 = 0$.

Again using Equation 54.9, Equation 54.3 becomes

$$\vec{B} = -i(k/\omega)P\hat{\xi}_{(\pm)}.$$  

(Eq. 54.13)

\(^{11}\)See Section 33.3.6 (page 516).

\(^{12}\)You’ll investigate the other circular polarization in Your Turn 54C.

\(^{13}\)Equation 54.6.
and hence Equation 54.4 is modified to

\[ -i k^2 \omega^{-1} p(-i \xi_{(+)}') = -\mu_0 \varepsilon_0 \omega \left( \omega - \omega_p^2 (\omega + q B_0 / m_e)^{-1} \right). \]  

(54.14)

The dispersion relation is now

\[ (ck)^2 = \omega^2 \left( 1 - \omega_p^2 (\omega^2 + q B_0 \omega / m_e)^{-1} \right), \]  

(54.15)

which again reduces to Equation 54.5 when \( B_0 = 0 \).

Your Turn 54C

a. Check how the units work in Equation 54.15.

b. Redo the derivation for the other helicity (circular polarization) of light and note the difference in dispersion relation.

c. Explain how one might have expected the result in (b) on symmetry grounds.

In short, the originally isotropic plasma has acquired circular birefringence.\(^{14}\) From here, the analysis is much the same as in Section 51.4 (page 716); see Problem 54.2.

54.4.2 The accretion disk of M87*

[[Not ready]]

... Figure 45.1 (page 663). [[ “The main finding is that we not only see the magnetic fields near the black hole as expected, but they also appear to be strong. Our results indicate that the magnetic fields can push the gas around and resist being stretched. The result is an interesting clue to how black holes feed on gas and grow,” – Dexter quoted in www.space.com/first-black-hole-image-polarized-m87. “Magnetic fields are theorized to connect black holes to the hot plasma surrounding them,” says Daniel Palumbo, a co-author and researcher at the CfA. “Understanding the structure of these fields is the first step in understanding how energy can be extracted from spinning black holes to produce powerful jets.” – news.harvard.edu/gazette/story/2021/03/for-first-time-images-capture-black-holes-magnetic-fields.]]

54.4.3 Faraday rotation also appears in condensed matter

Looking back, we may interpret our result by saying that a background magnetic field induces cross-susceptibility.\(^{15}\) The electric field of an incoming wave gets transmuted into a current inducing a magnetic response. Viewed that way, we might expect that a magneto-optical effect would occur in nearly any transparent condensed matter, including Faraday’s original choice (glass). Indeed, Herschel’s original prediction of the effect relied solely on noting that circularly birefringent media break reflection invariance in a particular way (imposing a handedness), and so does a uniform magnetic field.

\(^{14}\)See Sections 51.2–51.3.

\(^{15}\)See Section 51.3.
A coil of wire 30 cm long with current $I$ produced an axial magnetic field of strength $(0.005 \text{T/A})I$. The coil was submerged in tap water and laser light was sent down the bore (right to left in the photo). Crossed linear polarizers at each end blocked the light when $I = 0$. With $I = 6.5 \text{A}$, some light was transmitted. Readjusting the second polarizer to block the light required rotation of about 3 deg. Reversing the current led to a rotation of the same magnitude but opposite sign. [Courtesy Peter Hamish.]

FURTHER READING

Semipopular:
J. Bell Burnell gives her own account of the pulsar discovery: Bell Burnell, 1977 = www.bigear.org/volno1/burnell.htm. “I went home that evening very cross — here was I trying to get a Ph.D. out of a new technique, and some silly lot of little green men had to choose my aerial and my frequency to communicate with us.”


M87*:
scitechdaily.com/astronomers-polarized-image-shows-magnetic-fields-at-the-edge-of-m87s-black-hole/.


news.harvard.edu/gazette/story/2021/03/for-first-time-images-capture-black-holes-magnetic-fields/.

Intermediate:
Chapter 54  Waves in a Cold Plasma and Faraday’s Rotation Effect


Technical:
Michilli & others, 2018.
Goddi & others (Event Horizon Telescope collaboration), 2021; Akiyama & others (Event Horizon Telescope collaboration), 2021a; Akiyama & others (Event Horizon Telescope collaboration), 2021b.
Martinot et al., 2018.
Alex Lopatka, “Radio emission confirms that a magnetic field spans intergalactic space,” Physics Today 72(8), 17 (2019); doi.org/10.1063/PT.3.4264

PROBLEMS

54.1 .
54.2  Rotation measure
Suppose that a linearly polarized plane wave with some angular frequency \( \omega \) enters a cold plasma, propagating along the direction of a uniform background magnetic field of strength \( B_0 \).

a. Find the difference in phase velocities for the two circular polarizations and expand to lowest nontrivial order in \( B_0 \).

b. Express the incoming wave in the circular basis. Derive a formula for the complex electric fields after propagating a distance \( L \) in the plasma, recombine them, and show that the result is again linearly polarized, but along a new direction.

c. Derive a formula for the rotation of the polarization vector in terms of \( B_0, L, \) frequency of the incoming wave, and electron density in the plasma. State very carefully the predicted sign of the effect.

d. The result is often expressed in terms of rotation measure, the angle of rotation in radians per vacuum wavelength squared of the incoming light. Why is this a useful quantity? Re-express your answer to (c) by giving a formula for rotation measure.

e. It may seem pointless to have a formula for rotation, when we can’t travel to a distant astrophysical object and measure the original direction of polarization! Explain how on the contrary your formula in (d) can be useful even without that information.

[Hint: Recall Section 54.3.5.]

54.3  A one-way light valve
Faraday rotation combined with polarizing filters creates a one-way valve for light.

Faraday rotation combined with polarizing filters creates a one-way valve for light.

November 14, 2023; Contents  Index  Notation  Glossary
A two-slit diffraction pattern built up from individual pointlike photon absorption events on a camera, in an experiment where no more than one photon is present between source and detector at any moment. Pixels with zero counts are black; pixels that captured 500 or more photons are white. [Image courtesy Joseph Peidle; see also Rueckner & Peidle, 2013.]
Motionless in appearance, matter contained… dramas subjected to implacable fatality; it contained life and death. Such were the facts which the discovery of radioactivity revealed. Philosophers had only to begin their philosophy all over again and physicists their physics.

— Eve Curie

55.1 FRAMING: EX NIHIL

Before we unleash a lot of formulas, let’s first frame the issues with an experimental observation. A concrete example of what we’d like to understand is the pattern of light seen from a single immobilized fluorophore, for example in defocused orientation imaging (Figure 55.1). The angular distribution of photon arrivals resembles the dipole radiation pattern found in Chapter 43, but the emission of single photons by a single molecule is as far from being classical as one can get. Is the observed agreement in radiation patterns just a coincidence? This chapter will argue that in fact, a quantum-mechanical treatment recapitulates the classical distribution of energy flow as a probability density function for photon arrivals. Along the way we’ll understand how the creation of light (and other) particles ex nihilo (from nothing) is possible at all.

Electromagnetic phenomenon: Insects and crustaceans can “see” the polarization state of light.
Physical idea: The probability for a molecule to absorb or emit a photon depends on its orientation relative to the photon’s polarization.

Figure 55.1: [Experimental data and fits.] Defocused orientation imaging. Top: Observed histograms for the arrivals of individual photons, for each of three isolated fluorophores. Lighter colors correspond to pixels with larger photon counts. Bottom: Corresponding theoretical predictions, after finding the best-fit value of the angle between the transition dipole and the microscope’s optical axis. From left to right, the fit values of this angle were 10°, 60°, and 90°. The in-plane orientation (azimuth) was also obtained by fitting. [From Toprak et al., 2006.]
55.2 Maxwell Equations as Decoupled Harmonic Oscillators

Classical electrodynamics describes a system whose states are field configurations. But Nature is described by quantum probability amplitudes, not classical state variables. The goal of this section is therefore to recast the eight Maxwell equations for the electric and magnetic fields in a form that is suitable for quantization. Later, Section 55.3 will find the photon concept as a consequence of field quantization. Although the application in Figure 55.1 occurs in solution, for simplicity we will instead work out the relevant formulas in vacuum.

As usual, we can represent electric and magnetic fields via a scalar potential field, \( \psi(t, \vec{r}) \), and a vector potential field, \( \vec{A}(t, \vec{r}) \):

\[
\vec{E} = -\frac{\partial}{\partial t} \vec{A} = \vec{\nabla} \psi; \quad \vec{B} = \vec{\nabla} \times \vec{A}.
\]  

Chapter 18 showed that in this representation, half of the Maxwell equations are identities (automatically true). We will choose to work in Coulomb gauge, that is, use only vector potentials that obey \( \vec{\nabla} \cdot \vec{A} = 0 \).\(^1\) Section 18.8.3 showed that in a world with no charged particles we can always specialize further, supplementing Coulomb gauge with the extra condition that the scalar potential \( \psi = 0 \) everywhere. (Later sections will reinstate \( \psi \) when we consider coupling of the field to electrons.)

We next show that the Maxwell equations reduce to a set of simple, decoupled dynamical systems. It’s convenient to imagine a finite world of some very large size \( L \), which will ultimately be taken to be infinity, and specifically to take that world to be a cube with periodic boundary conditions. Then the vector potential can be expanded as

\[
\vec{A}(t, \vec{r}) = \frac{1}{2} \sum_k' (\hat{A}_k(t)e^{i\vec{k} \cdot \vec{r}} + \text{c.c.}). \tag{55.1}
\]

In this formula, each coefficient \( \hat{A}_k \) is a complex 3-vector depending on time. There are many such vectors, indexed by a discrete label \( \vec{k} \) with components of the form \( 2\pi \eta_j / L \); the \( \eta_j \) are integers, not all of which are zero. The primed summation means that for each such wavevector \( \vec{k} \), we exclude the redundant \(-\vec{k}\).

The Coulomb gauge condition implies that \( \vec{k} \cdot \hat{A}_k = 0 \), or in other words that the component of each \( \hat{A}_k \) along its \( \vec{k} \) must equal zero. The other two components are unrestricted, so for each \( \vec{k} \), we choose a basis of two real unit vectors perpendicular to it and to each other; we denote these polarization basis vectors by \( \hat{s}_{(\alpha;\vec{k})} \), where the index \( \alpha \) runs from 1 to 2. Then Equation 55.1 becomes

\[
\vec{A}(t, \vec{r}) = \frac{1}{2} \sum_{\alpha, \vec{k}}' (A_{\alpha,\vec{k}}(t)\hat{s}_{(\alpha;\vec{k})}e^{i\vec{k} \cdot \vec{r}} + \text{c.c.}). \tag{55.2}
\]

In each term of the summation, \( A_{\alpha,\vec{k}} \) is now a single function of time. The polarization basis vectors are not dynamical variables. The dynamical variables, whose equations of motion we wish to find and quantize, are the coefficients \( A_{\alpha,\vec{k}}(t) \).

\(^1\)Section 18.8.2 (page 281).
Your Turn 55A

Show that, with these definitions, the Maxwell equations in Coulomb gauge become simple:
\[
\frac{d^2}{dt^2} A_{\alpha k} = -\epsilon_0 c \vec{k} \vec{\nabla} A_{\alpha k}. \tag{55.3}
\]
Here \(\alpha\) runs over 1,2 and \(\vec{k}\) runs over the nonredundant set described earlier.

Equation 55.3 shows that every distinct combination of polarization \(\alpha\) and wavevector \(\vec{k}\) corresponds to an independent dynamical system, decoupled from the others. To make the system more familiar, we now give separate names to the real and imaginary parts of \(A_{\alpha k}\):
\[
A_{\alpha k} = (\epsilon_0 L^3 / 2)^{-1/2} (X_{\alpha k} + i Y_{\alpha k}). \tag{55.4}
\]
The overall rescaling chosen in the definitions of \(X\) and \(Y\) will simplify some later formulas.

The real scalar quantities \(X_{\alpha k}\) and \(Y_{\alpha k}\) separately obey Equation 55.3, so we see that
\[
The vacuum Maxwell equations are mathematically equivalent to a set of decoupled harmonic oscillators. \tag{55.5}
\]
The harmonic oscillator has a well known quantum-mechanical formulation, so Idea 55.5 achieves the first goal of this section.

To understand the meaning of these oscillators better, we now express the electromagnetic field energy \(\mathcal{E}\) and momentum \(\vec{P}_{\text{field}}\) in terms of the new variables \(X\) and \(Y\). Let \(\dot{\vec{A}}\) denote the time derivative \(\partial \vec{A} / \partial t\). Then Your Turn 35Da (page 561) gives
\[
\mathcal{E} = \frac{\epsilon_0}{2} \int d^3 r \left( \vec{E}^2 + c^2 \vec{B}^2 \right) = \frac{\epsilon_0}{2} \int d^3 r \left( (-\dot{\vec{A}})^2 + c^2 (\vec{\nabla} \times \vec{A})^2 \right)
\]
\[
= \frac{\epsilon_0}{2} \sum_{\alpha k_1} \sum'_{\beta k_2} \int d^3 r \left[ \frac{1}{2} (A_{\alpha k_1} \dot{\vec{\xi}}_{(\alpha k_1)} e^{i \vec{k}_1 \cdot \vec{r}} + \text{c.c.}) \cdot \frac{1}{2} (-A_{\beta k_2} \dot{\vec{\xi}}_{(\beta k_2)} e^{i \vec{k}_2 \cdot \vec{r}} + \text{c.c.}) + c^2 \frac{1}{2} (A_{\alpha k_1} \vec{k}_1 \times \dot{\vec{\xi}}_{(\alpha k_1)} e^{i \vec{k}_1 \cdot \vec{r}} + \text{c.c.}) \cdot \frac{1}{2} (A_{\beta k_2} \vec{k}_2 \times \dot{\vec{\xi}}_{(\beta k_2)} e^{i \vec{k}_2 \cdot \vec{r}} + \text{c.c.}) \right]. \tag{55.6}
\]
The integrals are easy to do, because most of them vanish: Only those cross-terms with \(\vec{k}_1 = \vec{k}_2\), and hence involving \(e^{i \vec{k} \cdot \vec{r}} e^{-i \vec{k} \cdot \vec{r}} = 1\), survive. Moreover, we have \(\dot{\vec{\xi}}_{(\alpha k)} : \dot{\vec{\xi}}_{(\beta k)} = \delta_{\alpha \beta}\), leaving
\[
\mathcal{E} = \frac{\epsilon_0 L^3}{4} \sum_{\alpha k} (|A_{\alpha k}|^2 + ||e \vec{k}||^2 |A_{\alpha k}|^2)
= \frac{1}{2} \sum_{\alpha k} \left( X_{\alpha k}^2 + ||e \vec{k}||^2 X_{\alpha k}^2 + Y_{\alpha k}^2 + ||e \vec{k}||^2 Y_{\alpha k}^2 \right). \tag{55.7}
\]
The field momentum is given by a similar calculation, starting with the Poynting vector...
55.3 Quantization

(Your Turn 35D, page 561b):

\[ \vec{P}_{\text{field}} = e_0 \int d^3r \vec{E} \times \vec{B} \]

\[ = e_0 \sum_{\alpha, \vec{k}} \sum_{\beta, \vec{k}'} \int d^3r \frac{1}{2} (-\hat{A}_{\alpha, \vec{k}} \hat{s}(\alpha, \vec{k}) e^{i \vec{k} \cdot \vec{r}} + \text{c.c.}) \times \left( \hat{\vec{v}} \times \frac{1}{2} (\hat{A}_{\beta, \vec{k}} \hat{s}(\beta, \vec{k}) e^{i \vec{k} \cdot \vec{r}} + \text{c.c.}) \right) \]

\[ = -\frac{e_0 L^3}{4} \sum_{\alpha, \vec{k}} \sum_{\beta} (\hat{A}_{\alpha, \vec{k}} \hat{A}^*_{\beta, \vec{k}} \hat{s}(\alpha, \vec{k}) \times (-i \hat{\vec{v}} \times \hat{s}(\beta, \vec{k}) \text{ c.c.}) \]

\[ = \frac{1}{2} \sum_{\alpha, \vec{k}} ' k (i \hat{X}_{\alpha, \vec{k}} - \hat{Y}_{\alpha, \vec{k}}) (\hat{X}_{\alpha, \vec{k}} - i \hat{Y}_{\alpha, \vec{k}} \text{ c.c.}) \]

\[ = \sum_{\alpha, \vec{k}} ' k (\hat{X}_{\alpha, \vec{k}} \hat{Y}_{\alpha, \vec{k}} - \hat{Y}_{\alpha, \vec{k}} \hat{X}_{\alpha, \vec{k}}). \]  

We now have compact formulas for the energy and momentum of the electromagnetic field in terms of the harmonic-oscillator representation (Equation 55.7 and 55.9). The interpretation is that every mode of the field, labeled by \( \alpha \) and \( \vec{k} \), makes an independent contribution to \( \vec{E} \), and also to each component of \( \vec{P}_{\text{field}} \). Note, however, that the momentum gets mixed contributions from the X and Y oscillators. We will soon remove this remaining inconvenience.

55.3 QUANTIZATION REPLACES FIELD VARIABLES BY OPERATORS

Finding the quantum-mechanical version of a harmonic oscillator is a standard problem which will be easy after we make a rather involved change of variables. To motivate the required change, we will break it down into four steps. It is worthwhile to verify each of the steps, which are straightforward if a bit tedious; ultimately the goal is to replace the \( \hat{X} \) and \( \hat{P} \) variables by a set of quantum operators called Q and their hermitian conjugates (Equation 55.21). Note that this chapter uses different typefaces to distinguish quantum operators from their corresponding classical dynamical variables.

Step 1: Quantize

For brevity, at first consider only one pair of modes \( X \) and \( Y \), that is, only a particular \( \alpha, \vec{k} \). We introduce two hermitian operators \( X \) and \( U \), with the property that their commutator is \( [X, U] = i \hbar \). In the energy function, Equation 55.7, we substitute \( X \rightarrow X \) and \( \hat{X} \rightarrow U \) to obtain the hamiltonian operator for \( X \):

\[ H_X = \frac{1}{2} (U^2 + \| \vec{\hat{r}} \|^2 X^2). \]  

\[ ^2 \text{In the analogy to a harmonic oscillator, these represent the position and momentum respectively, but they have no direct connection to physical position \( \vec{r} \) or field momentum \( \vec{P} \).} \]
This operator both represents the energy of a quantum state and also determines its time evolution. For example, the time evolution of \( |\Psi(t)\rangle \) is given by \( \exp(-iH_0 t/\hbar)|\Psi\rangle \). It implies that

\[
\frac{d^2}{dt^2} \langle \Psi_1 | X | \Psi_2 \rangle = \frac{d}{dt} \langle \Psi_1 | i \hbar [H_0, X] | \Psi_2 \rangle = \frac{d}{dt} \langle \Psi_1 | U | \Psi_2 \rangle = \langle \Psi_1 | \frac{i}{\hbar} [H_0, U] | \Psi_2 \rangle = -\|c\vec{k}\|^2 \langle \Psi_1 | X | \Psi_2 \rangle,
\]

which implements the classical equation of motion for the harmonic oscillator in Equation 55.3.

We proceed in the same way with the other oscillator family, introducing hermitian operators \( X \) and \( U \) analogous to \( H_0 \) and \( H_0 \). Then the operator corresponding to \( A_{x,k} \) in Equation 55.4 is

\[
A = (\epsilon_0 L^3 / 2)^{-1/2} (X + iY).
\]

**Step 2: Diagonalize energy**

We could now finish constructing the state space, for example, by writing and solving a set of decoupled Schrödinger equations for each pair of operators \( (X, U) \) and \( (Y, V) \). However, the harmonic oscillator problem has an elegant reformulation that simplifies the math. Change variables once again by defining new operators

\[
S = (2\hbar ||c\vec{k}||)^{-1/2} (||c\vec{k}||X + iU) \quad \text{and} \quad R = (2\hbar ||c\vec{k}||)^{-1/2} (||c\vec{k}||Y + iV).
\]

\( S \) and \( R \) are not hermitian; indeed, it is straightforward to verify that

\[
[S, S^\dagger] = 1, \quad [R, R^\dagger] = 1, \quad [S, R] = [S, R^\dagger] = 0,
\]

\[
H = H_0 + H_Y = \hbar ||c\vec{k}|| (S^\dagger S + R^\dagger R + 1), \quad \text{and}
\]

\[
\vec{P} = i\hbar \vec{k} (S^\dagger R - \text{h.c.}).
\]

In the last formula, “h.c.” denotes the hermitian conjugate, that is, \( R^\dagger S \).

**Step 3: Diagonalize momentum**

The hamiltonian operator has the nice property that \( S \) and \( R \) make independent, additive contributions to it (Equation 55.15). The momentum operator still mixes \( S \) and \( R \), but we can diagonalize it, without spoiling \( H \), by a unitary transformation. Define two **destruction operators** by

\[
Q = (S + iR)/\sqrt{2}, \quad \bar{Q} = (S - iR)/\sqrt{2}.
\]
Your Turn 55B

Show that

\[ [Q, Q^\dagger] = 1, \quad [\bar{Q}, \bar{Q}^\dagger] = 1, \quad [Q, \bar{Q}] = [Q, \bar{Q}^\dagger] = 0, \quad (55.18) \]

\[ H = \hbar|c\vec{k}|(|Q^\dagger Q + \bar{Q}^\dagger \bar{Q} + 1), \text{ and} \]

\[ \vec{P} = \hbar \vec{k}(Q^\dagger Q - \bar{Q}^\dagger \bar{Q}). \quad (55.19) \]

We now have new field operators \( Q \) and \( \bar{Q} \) that, unlike \( S \) and \( R \), diagonalize both the field energy and momentum.

Step 4: Relabel

We now reinstate the mode indices \( \alpha \) and \( \vec{k} \). Until now, all mode sums were over a half-space of discrete \( \vec{k} \) values, but now we can simplify the notation: Define operators for all nonzero \( \vec{k} \) by renaming \( \bar{Q}_{\alpha, \vec{k}} \) as \( Q_{\alpha, -\vec{k}} \). Then

\[ [Q_{\alpha, \vec{k}_1}, Q^\dagger_{\beta, \vec{k}_2}] = \delta_{\alpha \beta} \delta_{\vec{k}_1, \vec{k}_2}, \quad [Q_{\alpha, \vec{k}_1}, \bar{Q}_{\beta, \vec{k}_2}] = 0, \quad \text{for all nonzero } \vec{k}_1 \text{ and } \vec{k}_2. \quad (55.21) \]

Our final formulas then become unrestricted sums:

\[ H = \sum_{\alpha, \vec{k}} \hbar |c\vec{k}| \left( Q^\dagger_{\alpha, \vec{k}} Q_{\alpha, \vec{k}} + \frac{1}{2} \right), \text{ and} \]

\[ \vec{P} = \sum_{\alpha, \vec{k}} \hbar \vec{k} \left( Q^\dagger_{\alpha, \vec{k}} Q_{\alpha, \vec{k}} \right). \quad (55.22) \]

We now have a set of operators in terms of which the energy and momentum of light will have simple interpretations.

55.4 PHOTON STATES

55.4.1 Basis states can be formed by applying creation operators to the vacuum state

We have found a set of field-like operators that obey Maxwell-like equations, and recast them in terms of the \( Q \) and \( Q^\dagger \) operators. Besides giving an elegant approach to quantization, this formulation gives a basis of states that is readily interpretable in light of the Einstein and de Broglie relations: \(^3\)

Your Turn 55C

Show that

\[ [H, Q_{\alpha, \vec{k}}] = -\hbar |c\vec{k}| Q_{\alpha, \vec{k}} \quad \text{and} \quad [\vec{P}, Q_{\alpha, \vec{k}}] = -\hbar \vec{k} Q_{\alpha, \vec{k}}. \quad (55.24) \]

\(^3\)See Section 32.7.3 (page 504).
Equations 55.24 justify the term “destruction operator”:

Applying the destruction operator \( Q_{\alpha,k} \) to a state lowers its energy by \( \hbar ||c_k^\perp|| \), and changes its momentum by \( -\hbar \vec{k} \). Conversely, applying the creation operator \( Q^\dagger_{\alpha,k} \) has the opposite effects.

Next, note that both of the terms in the classical electromagnetic energy function (Equation 55.6) are nonnegative. So it must not be possible to lower that energy indefinitely; there must be a state for which any destruction operator yields zero. We’ll denote that photon ground state by the symbol \( |0\rangle \). Any other state is obtained from this one by the actions of the various creation operators, each of which may be applied any number of times, always raising the energy by \( \hbar ||c_k^\perp|| \) and changing the momentum by \( \hbar \vec{k} \). The spectrum of allowed energy and momentum values suggests a description: It is the same as that of a gas of noninteracting particles, each carrying energy \( \hbar ||c_k^\perp|| \) and momentum \( \hbar \vec{k} \).

**Your Turn 55D**

Show that when a creation operator acts \( n \) times, we can obtain a normalized state as follows:

\[
|n_{\alpha,k}\rangle = \sqrt{\frac{1}{n!}} (Q^\dagger_{\alpha,k})^n |0\rangle.
\]  

(55.26)

More generally, we can define \( |n_{\alpha,k_1}; n_{\alpha,k_2}; \ldots\rangle \) as a state obtained by applying several different creation operators to the ground state, each multiple times, and then normalizing. States of this form with different sets of occupation numbers are all linearly independent and mutually orthogonal. In fact,

The quantum states of light form a linear space spanned by basis vectors of this form, which act like states of noninteracting particles (“photons”).

(55.27)

That is, each one-photon basis state is labeled by a wavevector and a polarization, and carries energy and momentum related by Equation 55.24:

\[
E_{\alpha,k} = \hbar ||c_k^\perp||; \quad \vec{P}_{\alpha,k} = \hbar \vec{k}; \quad \text{so} \quad E_{\alpha,k} = ||\vec{P}_{\alpha,k}||,
\]

(55.28)

implying that photons are massless. For multiphoton states, we add the corresponding quantities, just as we would do with any noninteracting particles.

The interpretation of the quantum basis states as containing particles motivates another commonly used set of terms for the creation and destruction operators: Because they can be interpreted as raising and lowering the number of photons in a state, they are also called raising and lowering operators; \( |0\rangle \) is also called the vacuum state. We may guess that these concepts will be key to understanding how a fluorescent molecule in its excited state can create photons from “nothing” (and how other processes can make photons disappear, Section 32.7.4, page 505).
55.4.2 On being out on a limb

The fact that the collection of occupation numbers, \( \{ n_{\alpha, \vec{k}} \} \), fully determines a basis state is the key insight that leads to the famous spectrum of thermal ("black body") radiation. This aspect of light can alternatively be expressed by saying that the particles of light with given \( \alpha, \vec{k} \) are indistinguishable: They have no further attributes, so all we need to state is how many are present. For example, it doesn’t matter in what order we build a photon state by applying creation operators, because those operators all commute with one another.\(^6\)

It is worth remembering, however, that the tidy picture just described is very far from the historical development. In 1905, Einstein’s light-quantum idea was wildly incompatible with experimental data on the low-frequency end of blackbody spectrum, which was then the hot news in the world of experimental physics. Therefore it is also incompatible with the celebrated Planck formula. No one on the planet, including Einstein, had any idea that a gas of noninteracting particles could have any sort of statistics other than Boltzmann, and that this would ultimately resolve this glaring discrepancy. Somehow, Einstein had the courage to realize that this discrepancy was less important than the other things his proposal got right, and to believe that eventually it would be sorted out.

Our current understanding of the thermal radiation spectrum had to wait another 21 years after Einstein’s original publication, when Dirac undertook an analysis roughly like the one in the preceding sections.

55.4.3 Coherent states mimic classical states in the limit of large occupation numbers

The states we have called “one-photon” are far from being classical. Indeed, no state with a definite number of photons can be an eigenvector of the field operators corresponding to the classical electric and magnetic field, because \( \tilde{A}(\vec{r}) \) involves both creation and destruction operators:

**Your Turn 55E**

Use Equations 55.2, 55.12, 55.13, and 55.17 to show that

\[
\tilde{A}(\vec{r}) = \sum_{\alpha, \vec{k}} \sqrt{\frac{\hbar}{2L^3 \epsilon_{\alpha} ||c\vec{k}||}} \hat{\phi}_{(\alpha, \vec{k})} (Q_{\alpha, \vec{k}} e^{i\vec{k} \cdot \vec{r}} + \text{h.c.}).
\]  

(55.29)

However, we can find eigenvectors of \( Q_{\alpha, \vec{k}} \), called **coherent states**: For any complex number \( u \), define

\[
|u, \alpha, \vec{k}\rangle = \exp\left(-\frac{1}{2}|u|^2\right) \sum_{n=0}^{\infty} (n!)^{-1/2} (u)^n |n_{\alpha, \vec{k}}\rangle.
\]  

(55.30)

---

\(^6\)More precisely, a class of particles that are indistinguishable in this way is called “bosonic.” Another possibility, called “fermionic” particles, has creation operators that mutually **anticommute**.
Chapter 55 Vista: Field Quantization

Your Turn 55F

a. Show that the states \( |u, \alpha, \vec{k}\rangle \) just defined are all properly normalized for any complex number \( u \).

b. Show that \( Q_{\alpha, \vec{k}} |u, \alpha, \vec{k}\rangle = u |u, \alpha, \vec{k}\rangle \), and hence also \( \langle u, \alpha, \vec{k}| Q_{\alpha, \vec{k}}^\dagger = u^* \langle u, \alpha, \vec{k}| \).

c. Then show that Equation 55.29 implies

\[
\langle u, \alpha, \vec{k}| \hat{A}(\vec{r}) |u, \alpha, \vec{k}\rangle = (2L^3 \epsilon_0 |c\vec{k}| / \hbar)^{-1/2} \frac{\vec{\xi}_{(\alpha,\vec{k})}}{} ue^{i\vec{k}.\vec{r}} + \text{c.c.}
\]

Your Turn 55G

The coherent states are superpositions of states with different numbers of photons. Find the probabilities of getting exactly \( \ell \) photons in a measurement on a coherent state by computing the modulus squared of each individual term of Equation 55.30. Is this a distribution you have seen previously?

Your result in Your Turn 55F shows that the coherent state based on a particular wavevector and polarization is the quantum analog of a classical single-mode state (Equation 55.2, page 769). Moreover, Your Turn 55G implies that as \( |u| \) becomes large (and hence also the expectation of the photon number), the relative standard deviation of the photon number in this state goes to zero, leading to classical behavior. In this limit, the coherent states correspond to classical states of the electromagnetic field, for example, the radiation emitted by a radio broadcast antenna.\(^7\)

This section has established contact between the field quantization procedure in this chapter, the particle picture from earlier chapters, and Maxwell’s original classical fields. The distinction between particle-like (occupation number representation) and wavelike (coherent state representation) has melted away into just two different choices of basis for the state space. Perhaps Michael Faraday had some tenuous intimation of this idea when he wrote, “The difference between a supposed little hard particle and the powers around it, I cannot imagine.”

55.5 INTERACTION WITH ELECTRONS

55.5.1 Classical interactions involve adding source terms to the field equations

If we wish to study the creation of light by a molecule, then we must acknowledge that the light field interacts with that molecule’s electrons. In the presence of charged matter, we can no longer find a gauge transformation that eliminates the scalar potential \( \psi \), though

\(^7\)Books on quantum optics show that the light created by a single-mode laser, operated well above threshold, is also a coherent state (Loudon, 2000, chap. 7).
we can still impose $\mathbf{\nabla} \cdot \mathbf{A} = 0$. The electric Gauss law then says

$$\mathbf{\nabla} \cdot \mathbf{E} = -\nabla^2 \psi = \rho_q / \varepsilon_0,$$

where $\rho_q$ is the charge density. This formula looks just like the corresponding equation in electrostatics, and it leads to the usual potential that binds the molecule’s electrons to its nuclei.

Ampère’s law also involves charges, via the electric charge flux $\mathbf{j}(t, \mathbf{r})$:

$$\mathbf{\nabla} \times \mathbf{B} = \mu_0 \mathbf{j} + \mu_0 \varepsilon_0 \frac{\partial}{\partial t} \mathbf{E}.$$

Casting everything into plane wave mode expansions as before gives the full Maxwell equations as

$$k^2 \psi_\mathbf{k} = \frac{1}{\varepsilon_0} \rho_{q,\mathbf{k}} \quad \text{and}$$

$$\frac{d^2}{dt^2} \mathbf{A}_{\mathbf{k}} + c k \mathbf{A}_{\mathbf{k}} = -i k \frac{d \psi_\mathbf{k}}{dt} + \frac{1}{\varepsilon_0} \mathbf{j}_{\mathbf{k}},$$

where $c = (\mu_0 \varepsilon_0)^{-1/2}$ and $\psi_\mathbf{k}, \rho_{q,\mathbf{k}},$ and $\mathbf{j}_{\mathbf{k}}$ are the plane-wave components of $\psi$, $\rho_q$, and $\mathbf{j}$, respectively. We now take the dot product of both sides of Equation 55.32 with the two transverse basis vectors $\hat{\mathbf{z}}_{(\alpha, \mathbf{k})}$ to find the desired generalization of Equation 55.3:

$$\frac{d^2}{dt^2} A_{\alpha, \mathbf{k}} = -c k \mathbf{A}_{\alpha, \mathbf{k}} + \frac{1}{\varepsilon_0} \mathbf{j}_{\mathbf{k}} \cdot \hat{\mathbf{z}}_{(\alpha, \mathbf{k})} \quad \text{for each } \mathbf{k} \cdot \alpha.$$

The scalar potential $\psi$ has dropped out of this equation of motion.

### 55.5.2 Electromagnetic interactions can be treated perturbatively

There is no need to quantize the scalar potential $\psi$, because Equation 2.4 shows that, in Coulomb gauge, it is not an independent dynamical variable: It just tracks whatever the charge density is doing.

The last term of Equation 55.33 describes the interaction of the vector potential with charge flux. To discuss the radiation of a molecule, we treat this term as a perturbation. That is, we set up an “unperturbed” hamiltonian operator describing the quantum mechanics of the electrons making up the molecule, with their Coulomb attraction to the nuclei mediated by the scalar potential $\psi$ as usual. There is another term describing the free electromagnetic field (Equation 55.22). To these terms we then add the perturbation

$$- \int d^3 r \mathbf{j}(\mathbf{r}) \cdot \mathbf{A}(\mathbf{r}),$$

where $\mathbf{j}(\mathbf{r})$ is the operator version of the charge flux and $\mathbf{A}(\mathbf{r})$ is given by Equation 55.29. This term modifies the quantum equations of motion, introducing the last part of Equation 55.33.
Each electron in the atom or molecule of interest contributes a delta function to
that is localized at the electron’s position $\vec{r}_e$, with strength equal to its charge, $-e$, times its velocity,$^8$ $\vec{p}_e/m_e$. Thus, each electron makes a contribution to the integral in Equation 55.34 equal to

$$-\sum_{\alpha,\vec{k}} \frac{\hbar}{2L^3 \epsilon_0 ||\vec{c}||} \hat{\xi}_{(\alpha,\vec{k})} \cdot (-e)(\vec{p}_e/m_e)(Q_{\alpha,\vec{k}}e^{i\vec{k} \cdot \vec{r}_e} + \text{h.c.}). \quad (55.35)$$

The effect of this perturbation is to allow transitions between eigenstates of the unperturbed hamiltonian operator, that is, between states that would be stationary were it not for the perturbation term. For example, the transitions that interest us are those from a molecule with initially excited electron state and no photons present, to a deexcited electron state and one photon present. To find the probability per unit time that this transition will occur, we need to compute the modulus squared of Equation 55.35 sandwiched between the initial and final states.$^9$ The hermitian conjugate term, involving $Q^\dagger$, can create the photon, so we want the matrix element of the remaining factors of this term sandwiched between the molecular states.

To make progress, notice that for transitions in the visible spectrum, $k \approx 10^{-2}$ nm$^{-1}$. But $r_e$ cannot exceed the size of the atom or molecule, typically $\approx 1$ nm, so $k \cdot r_e$ is a small dimensionless number. Accordingly, we will approximate $\exp(i\vec{k} \cdot \vec{r}_e)$ by its leading-order Taylor series term, which is $1$—the electric dipole approximation.$^{10}$

55.5.3 The dipole emission pattern

We now ask for the probability that the emitted photon will be observed to be traveling in a particular direction with a particular energy and polarization. The preceding section argued that dropping overall constant factors, the answer is proportional to

$$\left| \left\langle \text{ground}; \alpha, \vec{k} \right| \sum_{\beta,\vec{k}} Q^\dagger_{\beta,\vec{k}} \hat{\xi}_{(\beta,\vec{k})} \cdot \hat{\vec{p}}_e |\text{excited} \rangle \right|^2$$

$$= \left| \left\langle \text{ground} | \hat{\vec{p}}_e |\text{excited} \rangle \cdot \hat{\xi}_{(\alpha,\vec{k})} \right|^2. \quad (55.36)$$

One further transformation helps to clarify the meaning of this quantity. The electron momentum operator, whose matrix element we need, can be rephrased in terms of the electron position operator, as the commutator

$$[H_e, \vec{r}_e] = -\frac{i\hbar}{m} \hat{\vec{p}}_e.$$

Sandwich this relation between the ground and excited states to find

$$\langle \text{ground} | (E_0 \vec{r}_e - \vec{r}_e E_0) |\text{excited} \rangle = \frac{-i\hbar}{m} \langle \text{ground} | \hat{\vec{p}}_e |\text{excited} \rangle.$$

The right-hand side of this formula is a constant times the quantity needed in Equation 55.36. The left-hand side can be written in terms of the electric dipole moment op-

---

$^8$See Section 34.9.2 (page 538).

$^9$Quantum mechanics textbooks call this scheme the “Golden Rule” of time-dependent perturbation theory.

$^{10}$See Chapter 43.
operator, $\hat{D}_E = -e\vec{r}_e$, so we find that the probability of photon emission involves the matrix element of the dipole moment, a vector called the molecule's transition dipole. This is encouraging news: Chapter 43 showed that in the classical theory, the rate of energy radiation is also proportional to the amplitude squared of the electric dipole moment.

If the molecular states are such that the transition dipole is nonzero, then we can choose a coordinate system in which it points along the $z$ axis:

$$\langle \text{ground}\hat{D}_E | \text{excited} \rangle = D_E \hat{z}. \quad (55.37)$$

Suppose that, as is the case in many experiments, we record every photon received regardless of its polarization. The sum of Equation 55.36 over $\alpha$ includes the factor

$$\sum_\alpha \hat{z} \cdot \hat{\xi}_{(\alpha \vec{k})} \hat{\xi}_{(\alpha \vec{k})} \cdot \hat{z}. \quad (55.38)$$

We can simplify this expression by realizing that it involves the projection of $\hat{z}$ onto the plane perpendicular to $\vec{k}$. Another expression for that projection operator is $\hat{\mathbf{1}} - \hat{k} \hat{\mathbf{k}}$, so we get

$$\hat{z} \cdot (\mathbf{1} - \hat{k} \hat{\mathbf{k}}) \cdot \hat{z} = \hat{z} \cdot \hat{z} - (\hat{z} \cdot \hat{k})^2 = 1 - \cos^2 \theta = \sin^2 \theta, \quad (55.39)$$

where $\theta$ is the polar angle between the direction of observation, $\hat{k}$, and the transition dipole.

Equations 55.39 and 55.36 show that the probability density function for the angles at which photons are emitted has a “dipole doughnut” pattern. No photons are emitted along $\pm \hat{z}$; instead, they are preferentially emitted in the equatorial belt $\theta \approx \pi / 2$. A similar argument shows that the probability to absorb light also follows a dipole pattern.

The mean rate at which photons are emitted is determined by the transition dipole $D_E$ defined by Equation 55.37, which itself is essentially the matrix element of the molecule's electric dipole moment operator.

If the matrix element of the dipole moment operator is nonzero, then the dominant mechanism of energy loss by a molecule is the one just described, with its characteristic angular distribution $\varphi(\theta, \varphi) \propto \sin^2 \theta. \quad (55.40)$

This section has resolved the puzzle posed at the start of this chapter: The pattern of photon emission observed in defocused orientation imaging (Figure 55.1) agrees with the dipole radiation pattern in classical electrodynamics because the same angular factors enter each calculation.

### 55.6 VISTAS

#### 55.6.1 Many invertebrates can detect the polarization of light

Many species of insects, including honeybees, ants, crickets, flies, and beetles, have the ability to detect and act on the polarization of light. The first firm evidence for this...
sense was obtained by K. von Frisch in his studies of honeybees in 1948. Von Frisch knew that when returning from a successful foraging trip, a worker bee uses the location of the Sun in the sky to determine its own orientation. With this information, the bee can effectively integrate its instantaneous velocity to get an overall vector indicating the displacement to the source of food. Upon its return to the hive, the bee must communicate this information to others, via a “dance.”

The dance includes a segment of straight walking with a tail-wagging motion, indicating direction to the food by the direction of this straight segment. Because the bee only knows the direction relative to that of the Sun, it must again determine the Sun’s location before it can know which way to walk, and the others watching it must in turn remember that direction relative to the Sun if they are to follow that course. But remarkably, von Frisch found that the returning bee could successfully communicate even when he blocked the view of the Sun at the hive: As long as a small patch of blue sky was visible, communication was accurate. Von Frisch discussed the phenomenon with physicist H. Benndorf, who pointed out that the polarization pattern of the blue sky is related to the location of the Sun.

Von Frisch therefore hypothesized that bees could discern and act on the polarization of light. To test his hypothesis, he filtered the sky light visible to the bees through polarizers, modifying this one aspect of the bee’s environment while holding everything else unchanged. When the polarizer’s axis aligned with that of the sky’s polarization, then it had no effect other than to enhance the degree of polarization, and the bee’s dance was unaltered. When the polarizer was rotated, however, altering the apparent direction of polarization of the blue sky, then the bee’s dance changed, inaccurately reporting the location of the food.

55.6.2 Invertebrate photoreceptors have a different morphology from vertebrates’

Before describing how a polarization sense is possible, let’s first see why most vertebrates don’t have it. Photoreceptor cells in the vertebrate eye contain stacks of membrane layers oriented perpendicular to the incoming light. Embedded receptor proteins (rhodopsin) hold their light-sensitive cofactors (chromophores) with transition dipole parallel to the membrane, but within that plane the orientation is random. Thus, each chromophore’s transition dipole also points randomly in the plane perpendicular to the incoming photons’ direction of motion. Regardless of whether the incoming photons are polarized, their polarization vectors make random, uniformly distributed angles with the transition dipoles they encounter. Thus, although the photon’s probability to be absorbed by any particular chromophore depends on polarization, the overall absorption probability does

---

14 Polarization vision can also be used for communication: Some insects create polarized light, presumably to identify to others of their species (Figure 52.4, page 728).
15 Distance to the food is encoded in other characteristics of the dance such as tempo. To repeat the dance, the bee must loop around to the start of the straight segment; it does this loop without the tail-wagging motion, which indicates to others that this segment of the dance is to be ignored.
16 If no blue sky was visible, either by design or on an overcast day, then communication failed.
17 Section 47.7.2 (page 682).
Invertebrate photoreceptors. (a) Cross-section through one facet (ommatidium) of a generic insect eye. Light passes through the lens (left) to the rhabdom (right). (b) Enlargement of the orange box in (a). Several long photoreceptor cells \((R1–R7)\) run lengthwise along the ommatidium (in this example, from a crab). Each has many parallel, hairlike projections (microvilli, \(m\)) that together form the rhabdom along the central axis of the ommatidium. Each photoreceptor cell has many microvilli all oriented parallel to each other, but differently from those of neighboring cells. (c) Enlargement of the green box in (b). Electron micrograph of the microvilli confirming this arrangement in the mantis shrimp \(Gonodactylus oerstedii\). Alternating layers of microvilli are seen either end-on or sideways. [(b) From Stowe, 1980. (c) Courtesy Christina A. King-Smith.]

Insect and crustacean eyes have a different morphology (Figure 55.2). For example, bees have compound eyes each consisting of about 5000 individual units called ommatidia, about 150 of which are specialized for polarization vision (those in the “dorsal rim area”). Each ommatidium has its own rudimentary lens serving several photoreceptor cells. Each photoreceptor is long, in order to present many light-sensitive molecules to incoming photons that traverse its length; each guides light down its length much like an

---

18Actually some fish, for example the northern anchovy \(Engraulis mordax\), have cone cells with layers oriented parallel to the incoming light, enabling polarization vision (Horváth, 2014, chap. 9).

19[There are two major groups of animal photoreceptors, known as ciliary and rhabdomeric. Both use opsins, but they function in very different ways. Scientists used to think that ciliary receptors were only found in vertebrates, and rhabdomeric ones were only in invertebrates. But that’s not true: Both kinds of receptors are found in both groups. And both are found in the scallop, which has one retina full of ciliary photoreceptors and one full of rhabdomeric ones. Why? It’s unclear, although one retina appears to be used to detect moving objects and the other is used for selecting habitats. – Byong22a page 65]
optical fiber. Unlike the stack of disks bearing rhodopsin in the vertebrate photoreceptor, however, these rhabdomeric receptor cells embed their chromophores in an array of parallel, tubular projections called the rhabdomere. Each of these tubes (microvilli) is oriented with its long axis perpendicular to the incoming light; each microvillus in turn carries chromophores with their transition dipoles predominantly parallel to its axis.

The discussion in Section 55.5.2 averaged light emission probability over all polarizations. Had we not taken this step, we would have found that the probability to emit (or absorb) light depends on the light’s polarization relative to the molecule’s transition dipole. Thus, the effect of the insect’s ommatidial arrangement is that each photoreceptor cell has an overall preference for catching photons of a particular polarization. By comparing the outputs of different photoreceptors viewing the same patch of sky through the same lens, the bee can determine the orientation of polarization relative to that of its head, and from that information deduce its orientation relative to the Sun.

"Most insect, crustacean, and cephalopod [eyes] typically have two classes of photoreceptors that are stimulated by horizontally or vertically polarized light. By comparing their two receptors, they can distinguish between light that’s polarized to different extents, or at different angles. Cephalopods are more sensitive to polarization than any other animals. (FN: Temple S et al 2012. Current Biol 22:R121) Shelby Temple and his colleagues found that the mourning cuttlefish can spot the difference between two kinds of polarized light whose planes of vibration differ by just one degree. These animals are colorblind, but they might use polarization as a replacement, to add rich detail to their visual world. Mantis shrimps have this arrangement in the top hemisphere of their eyes. But in the bottom hemisphere, their polarization receptors are rotated by 45 degrees. And in rows 5 and 6 of the midband, they have something unique. Polarized light usually oscillates in a single fixed plane, but that plane can sometimes rotate, so the light travels along a twisting helix. This is called circular polarization. And as Marshall’s postdoc Tsyr-Huei Chiou found in 2008 (FN: Chiou et al 2008. Current Biol 18:429), mantis shrimps are the only animals that can see it. The bottom rows of their midbands have photoreceptors that are tuned to circularly polarized light, spiraling either clockwise or anticlockwise. So mantis shrimps have six classes of polarization receptors-vertical and horizontal, two diagonals, clockwise and anticlockwise. They can also rotate their eyes to enhance the polarization contrast between an object and its background, making them the first known animals with dynamic polarization vision (FN Daly et al 2016. Nat Communs 7:12140). In the bottom two rows of the midband, photoreceptors are arranged in an way that untwists circularly polarized light so that it becomes linearly polarized instead. That’s how mantis shrimp can sense it. (FN Cronin T Current Zool 64:513)” – Byong22 pp112-114

55.6.3 Some transitions are far more probable than others

Section 55.5 focused on the relative mean rates to emit photons in different directions. To find the absolute rates, we need various other factors provided by the “Golden Rule”
of time-dependent perturbation theory. The derivation of the rule also shows why energy must be conserved in photon emission and absorption, or more precisely, it must be conserved to within a tolerance set by the uncertainty relation.

For simplicity, Section 55.2 chose to expand the vector potential $\vec{A}$ in a basis of linearly polarized, plane wave states. Other bases may be better adapted to the problem at hand, for example, a basis of circularly polarized plane waves. Also, a basis of outgoing spherical waves, centered on the emitting object, is better suited to study light emitted by a very small object and traveling out to infinity. That basis can be chosen such that each element carries definite angular momentum away from the emitter. When we do this, we find that certain kinds of photons cannot be emitted at all by certain kinds of transitions, because doing so would violate the conservation of angular momentum. Other transitions appear impossible when we make the approximation $\exp(i\vec{k} \cdot \vec{r}_e) \approx 1$, as was done in Section 55.5.2, but not when we retain higher terms in the Taylor series. Such transitions are called “forbidden,” but more precisely their rates are just suppressed by powers of the small factor $(kr_e)^2$.

The statement that some transitions are “forbidden” is an example of a selection rule. Another class of selection rules arises from considerations of electron spin in multi-electron atoms or molecules. It is possible for a molecule to get trapped in an excited state, from which transitions to the ground state are suppressed by a spin selection rule. Such an excited state can eventually make its transition, but with mean rate far slower than most fluorescence transitions, leading to the phenomenon of phosphorescence (ultra-slow fluorescence). Spin selection rules also ensure very slow exit from the dark states of some fluorophores, which is useful for localization microscopy.

### 55.6.4 Lasers exploit a preference for emission into an already occupied state

Sections 55.5.2–55.5.3 restricted attention to the case in which a photon is emitted into a world originally containing no photons. Although photons do not interact in the usual sense of colliding, nevertheless a very important new phenomenon arises when we consider adding a photon to a state that is already occupied. If a mode initially contains $n$ photons, Equation 55.26 (page 774) implies

$$\langle n + 1|Q^\dagger|n\rangle = \langle 0|\frac{1}{\sqrt{n+1+1}}Q^{n+1}(Q^\dagger)^{n+1}\frac{1}{\sqrt{n+1}}|0\rangle = \langle 0|\sqrt{\frac{(n+1)!}{n!}}|0\rangle = \sqrt{n+1}.$$  

This factor gets squared when it enters into the rate for photon emission into this mode. Because this matrix element depends on $n$, we conclude that

When an atom or molecule emits a photon, it preferentially chooses a mode that is already occupied. (55.41)

If we have a population of many excited atoms or molecules, then this result implies that there can be an avalanche-type effect, in which one particular mode gets the vast majority of all emitted photons. This mechanism for obtaining nearly single-mode light is called light amplification by stimulated emission of radiation—the **laser**.
FURTHER READING

*Semipopular:*
Walmsley, 2015.
Animals with polarization vision: Yong, 2022; Chitka, 2022.

*Intermediate:*
Quantum theory of light: Grynberg et al., 2010; Lipson et al., 2011; Leonhardt, 2010; Loudon, 2000; Nelson, 2017.
Quantum versions of momentum and angular momentum of light: Grynberg et al., 2010; Andrews & Bradshaw, 2022.

*Technical:*
Defocused orientation imaging: Toprak et al., 2006; Böhmer & Enderlein, 2003; Bartko & Dickson, 1999a; Bartko & Dickson, 1999b.
APPENDIX A

Units and Dimensional Analysis

I think that the progress of science, both in the way of discovery, and in the way of diffusion, would be greatly aided if more attention were paid in a direct way to the classification of quantities.

—James Clerk Maxwell

This appendix recalls some general ideas about units and dimensions in physics. Chapter 16 carries the discussion onward to electrodynamics.

Some physical quantities are naturally integers, like the number of discrete clicks made by a Geiger counter. But others are continuous, and most continuous quantities must be expressed in terms of conventional units. This book uses the Système Internationale, or SI units, but you’ll need to be able to convert units when reading other works. Units and their conversions in turn form part of a larger framework called dimensional analysis.

Dimensional analysis gives a powerful method for catching algebraic errors, as well as a way to organize and classify numbers and situations, and even to guess new physical laws, as we’ll see in Section A.4.

To handle units systematically, remember that

A “unit” acts like a symbol representing an unknown quantity. Most continuous physical quantities should be regarded as the product of a pure number times one or more units.

(A few physical quantities, for example, those that are intrinsically integers, have no units and are called dimensionless.) We carry the unit symbols along throughout our calculations. They behave just like any other multiplicative factor; for example, a unit can cancel if it appears in the numerator and denominator of an expression. We know relations among certain units; for example, we know that 1 inch $\approx 2.54$ cm. Dividing both sides of this formula by the numeric part, we find 0.39 inch $\approx 1$ cm, and so on.

A.1 BASE UNITS

The SI chooses “base” units for length, time, mass, and electric charge: Lengths are measured in meters (abbreviated m), masses in kilograms (kg), time in seconds (s), and electric charge in coulombs (which this book abbreviates as coul). The system also creates related

---

1. One exception involves temperatures expressed using the Celsius and Fahrenheit scales, each of which differ from the absolute (Kelvin) scale by an offset.

2. The standard abbreviation is C, but this risks confusion with the speed of light, a concentration or capacitance variable, or a generic constant.
units via the prefixes giga ($= 10^9$), mega ($= 10^6$), kilo ($= 10^3$), centi ($= 10^{-2}$), milli ($= 10^{-3}$), micro ($= 10^{-6}$), nano ($= 10^{-9}$), pico ($= 10^{-12}$), or femto ($= 10^{-15}$), abbreviated as G, M, k, d, c, m, μ, n, p, and f respectively. Thus, 1 nm is a nanometer (or $10^{-9}$ m), 1 μg is a microgram, and so on. A symbol like $\mu m^2$ means $(\mu m)^2 = 10^{-12} m^2$, not “$\mu(m^2)$” = $10^{-6} m^2$.

A.2 DIMENSIONS VERSUS UNITS

Other quantities, such as electric current, derive their standard units from the base units. But it is useful to think about current in a way that is less strictly tied to a particular unit system. Thus, we define abstract dimensions, which tell us what kind of quantity a variable represents. For example,

- The symbol $L$ denotes the dimension of length. The SI assigns it a base unit called “meters,” but other units exist with the same dimension (for example, miles or centimeters). Once we have chosen a unit of length, we then also get derived units for area ($m^2$) and volume ($m^3$), which have dimensions $L^2$ and $L^3$, respectively.
- The symbol $M$ denotes the dimension of mass. Its SI base unit is the kilogram.
- The symbol $T$ denotes the dimension of time. Its SI base unit is the second.
- The symbol $Q$ denotes the dimension of electric charge. Its SI base unit is the coulomb.
- Electric current has dimensions $QT^{-1}$. The SI assigns it a standard unit coul/s, also called “ampere” and abbreviated A.
- Energy has dimensions $ML^2T^{-2}$. The SI assigns it a standard unit kg m$^2$/s$^2$, also called “joule” and abbreviated J.
- Power (energy per unit time) has dimensions $ML^2T^{-3}$. The SI assigns it a standard unit kg m$^2$/s$^3$, also called “watt” and abbreviated W.

Suppose that you are asked on an exam to compute an electric current. You work hard and write down a formula made out of various given quantities. To check your work, write down the dimensions of each of the quantities in your answer, cancel whatever cancels, and make sure the result is $QT^{-1}$. If it’s not, you may have forgotten to copy something from one step to the next. It’s easy, and it’s amazing how quickly you can spot and fix errors in this way.

When you multiply or divide two quantities, the dimensions combine like numerical factors: Particle flux ($T^{-1}L^{-2}$) times area ($L^2$) has dimensions appropriate for a rate ($T^{-1}$). On the other hand, you cannot add or subtract terms with different dimensions in a valid equation, any more than you can add rupees to centimeters. Equivalently, an equation of the form $X = Y$ cannot be valid if $X$ and $Y$ have different dimensions. (If either $X$ or $Y$ equals zero, however, then we may omit its units without ambiguity.)

You can add dollars to yuan, with the appropriate conversion factor, and similarly cubic centimeters to fluid ounces. Cubic centimeters and fluid ounces are different units.

---

Some books use $I = Q/T$, a “current” dimension, instead of $Q$. 

---
that both have the same dimensions \((L^3)\). We can automate unit conversions, and reduce errors, if we restate the conversion 1 US fluid ounce \(\approx 29.6\, \text{cm}^3\) in the form

\[
1 \approx \frac{\text{US fluid ounce}}{29.6\, \text{cm}^3}.
\]

Because we can freely insert a factor of 1 into any formula, we may introduce as many factors of the above expression as we need to cancel all the ounce units in that expression. This simple prescription (“multiply or divide by 1 as needed to cancel unwanted units”) eliminates confusion about whether to place the numeric factor 29.6 in the numerator or denominator.

### A.2.1 Densities

The words “density,” “areal density,” and “linear density” refer to the amount of something per volume, area, or length respectively. Thus, charge flux could instead be called “areal density of current,” but not “current density.” This book will use the two-syllable form, not the nine-syllable synonym, for that quantity.

### A.2.2 Angles

Angles are dimensionless: We get the angle between two intersecting rays, in the dimensionless unit radians (abbreviated rad), by drawing a circular arc of any radius \(r\) between them and centered on the intersection, then dividing the length of that arc (with dimensions \(L\)) by \(r\) (with dimensions \(L\)). Another clue is that if \(\theta\) carried dimensions, then trigonometric functions like sine and cosine wouldn’t be defined (see Section A.2). The angle corresponding to a complete circle is \(2\pi\) rad. An alternative expression for this quantity is 360 deg.

What about degrees versus radians? We can think of deg as a convenient or traditional unit\(^4\) with no dimensions: It’s just an abbreviation for the pure number \(\pi / 180\). The radian represents the pure number 1; we can omit it. Stating it explicitly as rad is just a helpful reminder that we’re not using degrees. Similarly, when phrases like “cycles per second” or “revolutions per minute” are regarded as angular frequencies, we can think of the words “cycles” and “revolutions” as dimensionless units (pure numbers), both equal to \(2\pi\).

Other traditional units of angle include arcmin = \(\text{deg} / 60\), and arcsec = \(\text{deg} / 3600\). The former is sometimes abbreviated by a prime, and the latter by a double prime, as in GPS coordinates: \(42^\circ\, 22'\, 42.29''\), but this book uses primes for other purposes. We will use the less confusing abbreviations arcmin and arcsec in place of ′ and ″.

Angular area (also called solid angle) is also dimensionless. Given a patch on the surface of a sphere, we get its angular area, in the dimensionless unit steradians (abbreviated sr), by finding the area of that patch and dividing by the sphere’s radius squared.

\[\text{We will use this abbreviation instead of the shorter } ^\circ, \text{ to avoid potential conflict with the temperature unit.}\]
A.2.3 Functions applied to dimensional quantities

If \( x = 1 \text{ m} \), then we understand expressions like \( 2\pi x \) (with dimensions \( \text{L} \)), and even \( x^3 \) (with dimensions \( \text{L}^3 \)). But what about \( \sin(x) \) or \( \log_{10} x \)? These expressions are meaningless; more precisely, they don’t transform in any simple multiplicative way when we change units, unlike say \( x/26 \) or \( x^2 \).

A.2.4 Additional units

Additional SI units

circular frequency: One hertz (\( \text{Hz} = \text{s}^{-1} \)) denotes one complete cycle per second. Use this unit name to emphasize that you are are not stating an angular frequency. Circular frequency is often abbreviated simply “frequency” and denoted by the symbol \( v \).

angular frequency: Change in angle per time. When stating an angular frequency, use an angular unit like \( \text{rad/s} \) to emphasize that you are not stating a circular frequency. Angular frequency is often denoted by the symbol \( \omega \) or \( \Omega \). The circular frequency \( 1 \text{ Hz} \) corresponds to angular frequency \( 2\pi \text{ rad/s} \).

temperature: One kelvin (\( K \)) can be defined by saying that the atoms of an ideal monoatomic gas have mean kinetic energy \( (3/2)k_B T \), where \( k_B = 1.38 \cdot 10^{-23} \text{ J K}^{-1} \).

resistance and conductance: One ohm (\( \Omega \)) equals one volt per ampere. One siemens is an inverse ohm: \( 1 \text{ Si} = 1 \text{ \Omega}^{-1} \).

electric potential: One volt (\( \text{V} \)) equals 1 J/coul.

Traditional but non-SI units

mass: One dalton (also called “unified atomic mass unit,” and abbreviated \( u \)) is 1 Da \( \approx 931.5 \text{ MeV/c}^2 \).

time: One minute is 60 s, and so on.

length: One Ångstrom unit (\( \text{Å} \)) equals 0.1 nm.

volume: One liter (\( \text{L} \)) equals \( 10^{-3} \text{ m}^3 \). Thus, \( 1 \text{ mL} = 1 \text{ cm}^3 \).

number density: A 1 M solution has a number density of \( 1 \text{ mole/L} = 1000 \text{ mole m}^{-3} \). In this book, the word “mole” represents the exact integer\(^6\) \( 6.02214076 \cdot 10^{23} \).

energy: An electron volt (eV) equals \( e \times (1 \text{ V}) = 1.60 \cdot 10^{-19} \text{ J} = 96 \text{ kJ/mole}. \) Here \( e \) is the electric charge on a proton. An erg (\( \text{erg} \)) equals \( 10^{-7} \text{ J} \). Thus, \( 1 \text{ kcal mole}^{-1} = 0.043 \text{ eV} = 6.9 \cdot 10^{-21} \text{ J} = 6.9 \cdot 10^{-14} \text{ erg} = 4.2 \text{ kJ mole}^{-1} \).

---

\(^5\)One way to see why such expressions are meaningless is to use the Taylor series expansion of \( \sin(x) \), and notice that it involves adding terms with incompatible units.

\(^6\)Thus we do not introduce a dimension for “amount of substance.”
A.3 ABOUT GRAPHS

When you make a graph involving a continuous quantity, state the units of that quantity in the axis label. For example, if the axis label says waiting time \([\text{s}]\), then we understand that a point aligned with the tick mark labeled 2 represents a measured waiting time that, when divided by 1 s, yields the pure number 2.

The same interpretation applies to logarithmic axes. If the axis label says flash photon density \([\text{photons/}\mu\text{m}^2]\), and the tick marks are unequal, then we understand that a point aligned with the first minor tick after the one labeled 10 represents a quantity that, when divided by the stated unit, yields the pure number 20 (in this case, 20 photons/\mu m^2). Alternatively, we can make an ordinary graph of the logarithm of a quantity \(x\), indicating this in the axis label, which says \(\log_{10} x\) or \(\ln x\) instead of \(x\). The disadvantage of the second system is that, if \(x\) carries units, then strictly speaking we must instead write something like \(\log_{10}(x/(1 \mu m^2))\) or \(\log_{10}(x [\text{a.u.}])\), because the logarithm of a quantity with dimensions has no meaning.

Sometimes a quantity is given in some unknown or unstated unit. It may not be necessary to be more specific, but you should alert your reader by saying something like emission spectrum [arbitrary units]. Many authors abbreviate this as “[a.u.]”

A.4 PAYOFF

Suppose that we wanted a relation between the period \(T\) and radius \(R\) of planetary orbits, but we couldn’t solve the equations of motion. We know from Galileo that the mass of the planet is immaterial, but the mass of the Sun may not be. We know that Newton’s constant must be relevant. What combinations of \(R\), \(M_{\text{sun}}\), and \(G_N\) have dimensions of time?

Consider the combination \(G_N^{\alpha} M_{\text{sun}}^{\beta} R^\gamma\) and adjust the exponents to give the whole thing the desired dimensions: \(\alpha = \beta, 3\alpha = -\gamma\), and \(\alpha = -1/2\). In this way we find, without solving for elliptical orbits, Kepler’s relation \(T \propto R^{3/2}\).!

Another useful application of dimensional analysis is in estimating an integral. For example, suppose that we wish to compute the total energy of black-body radiation:

\[
\int_0^\infty d\omega \ I(\omega) = \int_0^\infty d\omega \ \frac{\hbar \omega^3}{\pi^2 c^2 (e^{\hbar \omega/k_B T} - 1)}.
\]

To see what’s going on, find a dimensionless integration variable in terms of which the denominator is simple: \(u = \hbar \omega/k_B T\). Then changing variables shows that the total energy is an interesting part, \((k_B T)^4/(\pi^2 c^2 \hbar^3)\), which displays the parameter dependences, times a constant. The remaining integral itself, here \(\int_0^\infty du \ u^3/(e^u - 1)\), is not so interesting; it’s just a single universal number of order 1 that has been purged of any dependence on parameters.
Global List of Symbols

Good notation should serve you—not the other way round.
— Howard Georgi

Throughout this book the word “vector” is used specifically to mean a set of three numbers that points in space (or four numbers that point in spacetime). More abstract notions of vector, like the state vector of quantum mechanics, exist but don’t follow the particular transformation rules we use here.¹

B.1 MATHEMATICAL NOTATION

We need a notational system that is precise enough to express intricate ideas unambiguously, yet flexible enough to not be a burden when we know what we’re doing. If possible, we also want a system that nudges us to write down correct formulas rather than wrong ones.

Abbreviated words

c.c. Complex conjugate of the preceding term(s).

|_ret Evaluated at “retarded time” (observation time minus R/c); see Section 25.4, page 393.

Operations

z* Complex conjugate of a complex number z.

|z| Absolute value of a complex number, = \sqrt{z^*z}.

\nabla Laplace operator.

□ Wave operator, also called dalembertian. (Some books write □² instead of □, to parallel the symbol \nabla².)

★ Hodge dual operation (Section 15.9'b, page 235 and Section 34.9', page 544).

\vec{a} \cdot \vec{b} Dot product of two 3-vectors, itself a scalar. \vec{X} \cdot \vec{Y}, invariant scalar product of two 4-vectors.

\|\vec{b}\| Length of a real 3-vector, = \sqrt{\vec{b} \cdot \vec{b}}. For a complex vector, we instead write explicitly \sqrt{\vec{b}^* \cdot \vec{b}} when that’s what is meant. Similarly, \|\vec{X}\|^2 = \vec{X} \cdot \vec{X} is the invariant scalar product of a real 4-vector with itself (possibly a negative quantity). For a complex 4-vector, we instead write \vec{X}^* \cdot \vec{X} if that’s what is meant.

¹ A mathematician might therefore say “rank-1 tensor” wherever this book says “vector.”
\( \vec{a} \times \vec{b} \) cross (vector) product of two 3-vectors.

\( \vec{a} \otimes \vec{b} \) dyad (tensor) product of two vectors, itself a rank-2 tensor. (Other authors call it “outer product.”) The generalization to tensors of any rank is called “tensor product.” (Some books omit the symbol \( \otimes \) and just write \( \vec{a} \vec{b} \).)

\( \mathbf{T}^{[S]} \) and \( \mathbf{T}^{[A]} \) symmetric and antisymmetric parts of a rank-2 tensor (Equation 32.9, page 493).

**Complex notation**

The real part of a complex expression will always be written out in full, usually as \( \frac{1}{2}X + \text{c.c.} \). (Beware that many authors abbreviate by dropping the \( \frac{1}{2} \) and the +c.c.; you are supposed to understand that in any complex expression, the real part is meant.)

Sometimes when we wish to discuss the real and imaginary parts separately, they will be called \( X^{(R)} \) and \( X^{(I)} \). Some books instead write \( X' \) and \( X'' \), but we use primes for other purposes; see below.

**Other modifiers**

An overbar on a symbol can denote peak value (amplitude) of a sinusoidally varying quantity with the same letter name, for example, \( f(t) = \bar{f} \cos(\omega t) \). More generally, such quantities may be complex; then \( f(t) = \frac{1}{2} \bar{f} e^{-\text{ii}at} + \text{c.c.} \).

Sometimes an overbar can instead be used to indicate the nondimensionalized version of some quantity.

A dot over a function name can mean a derivative with respect to time. A prime following a function name can mean a derivative with respect to a spatial coordinate. Primes have other uses, however; see below.

**3-vectors and -tensors**

Many books use boldface type to denote 3-vectors and 3-tensors. That’s hard to draw on a piece of paper or chalkboard, so this book uses an arrow above the variable’s name to denote a 3-vector and a double arrow to denote a 3-tensor of rank two. Tensors of higher rank will always appear with explicit indices indicating their components (and no arrow), for example, the Levi-Civita tensor \( \varepsilon_{ijk} \).

The components of a vector or tensor in some coordinate system are always denoted with subscripts. Most books drop the boldface or arrow when referring to the components of a vector or tensor, but we will retain it, to emphasize that those quantities are part of a particular class of geometrical objects.

In cartesian coordinates, specific index values can be labeled \( x, y, \) or \( z \), or equivalently 1, 2, or 3 respectively. In other coordinates, explicit names are used, such as \( r, \varphi, \) and \( z \) for cylindrical coordinates. Generic index values are represented by Latin-alphabet letters. When the same index of this sort appears twice in an expression, summation is implied (unless otherwise stated). When the same index appears on either side of an equality,
then several equalities are being asserted, one for each value (or each combined value if multiple such pairs of “loose” indices appear).

When transforming, we sometimes use $i, j, k \ldots$ indices for components with respect to the original coordinate system and $a, b, c \ldots$ for the transformed system. Sometimes $\mathbf{\tilde{v}}_x$ is used as a synonym for $\mathbf{\tilde{v}}_1$, and so on for $\mathbf{\tilde{v}}_y$ and $\mathbf{\tilde{v}}_z$.

$\nabla$ Spatial gradient operator. Its cartesian components $\nabla_i$ are the partial derivatives $\partial / \partial \tilde{r}_i$.

$\mathbb{1}$ Identity tensor (identity matrix regarded as a 3-tensor), also called “unit tensor.” Its cartesian components $\mathbb{1}_{ij}$ are given by the “Kronecker delta” symbol: $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.

When a letter that is normally used for a vector appears without an over-arrow or index, that notation usually refers to the length of the corresponding vector; for example, $\mathbf{d}r$ indicates the length of $\mathbf{\tilde{r}}$. However, $\mathbf{d}^3r$ denotes $\mathbf{d}x\mathbf{d}y\mathbf{d}z$ (which is not a vector).

A differential element of surface has area denoted generically by $d^2\Sigma$, or $d^2\mathbf{r} = \mathbf{d}x\mathbf{d}y$ if specific 2D cartesian coordinates are used. When multiplied by a perpendicular unit vector, it becomes the vector $d^2\mathbf{\tilde{r}}$. We must then specify which of two perpendiculars is meant, for example, the outward-pointing direction if the overall surface is closed, or the one associated by a right-hand rule to a particular choice of direction around the boundary of the overall surface.

If a 3-vector is normalized to unit length, it gets a hat (circumflex) instead of an arrow, for example, the coordinate basis vectors $\mathbf{\hat{x}}, \mathbf{\hat{y}},$ and $\mathbf{\hat{z}}$. These are constant unit vectors, but the radial unit vector $\mathbf{\hat{r}} = \mathbf{\tilde{r}}/r$ is a vector field.

When we have a collection of related vectors, for example, the positions of many particles, they may be distinguished by a subscript in parentheses, to avoid confusion with a vector component index. Thus, $\mathbf{\tilde{r}}_{(\ell)}$ is the position of particle $\ell$; its $x$ component is then $\tilde{r}_{(\ell)}^x$ and so on. By extension, $k_{(\pm)}$ does not refer to the components of a wavevector in the helicity basis (it has no overarrow); rather, it refers to the length of the wavevector in each of two cases (positive and negative helicity, Section 51.4, page 716).4

A few “alternate” versions of vector quantities will even get an upside-down hat (háček) instead of an arrow.

When a letter that is normally used for a rank-2 tensor appears without an over-arrow, that may indicate that in this instance, the tensor is assumed to be an overall scalar times the identity tensor. For example, an isotropic polarizability may be written as $\alpha$, shorthand for $\alpha \mathbb{1}$.

Tilde versus prime

Sometimes each member of a collection of vectors will be related to a corresponding member of another collection by a common operation, for example, a physical, or “active,”

---

1 Some books use the symbols $\mathbf{i}, \mathbf{j}, \mathbf{k}$, or simply $\mathbf{i}, \mathbf{j}, \mathbf{k}$, to represent the unit vectors that this book calls $\mathbf{\hat{x}}, \mathbf{\hat{y}}, \mathbf{\hat{z}}$.

4 This elaborate notation is not needed for a scalar quantity like refractive index: $n_\pm$ unambiguously indicates two cases.
rotation. We may use the same symbol for each set to emphasize the correspondence, but distinguish the modified ones with a tilde: $\vec{V}$ in place of $\vec{V}$, or even $\vec{r}(\xi)$ in place of $\vec{r}(\xi)$.

Primes will usually indicate a completely different concept. Sometimes we will express a single vector in terms of more than one coordinate system. Then the components (ordinary numbers) used to represent that vector will have two different forms, which we will write as $\vec{V}_i, i = 1, 2, 3$ and $\vec{V}^\alpha_a, a = 1, 2, 3$ respectively. In each case, we are referring to the same vector $\vec{V}$. What’s being rotated is the coordinate system, not $\vec{V}$, but this introduces a “passive” transformation on the components.

Similar remarks apply to higher-rank 3-tensors.

Occasionally, prime will instead be used to mean a derivative with respect to a spatial coordinate. Also, tilde will sometimes indicate a gauge transformation.

4-vectors and -tensors

Many books use no typographical signal to indicate 4-vectors and 4-tensors, but we use an underscore, regardless of rank. As with 3-quantities, we’ll retain the underscore even when referring to specific components, to emphasize that they have particular transformation rules under change of coordinate system, for example $p^\mu$. However, $d^4X$ denotes $dctdxdydz$ (which is not a 4-vector).

The components of a 4-vector or 4-tensor in some coordinate system are denoted with sub- and superscripts. Subscript indices are distinct from superscript indices, as explained in Chapters 32–33. They start from 0 (time), so that in an inertial coordinate system index values 1,2,3 still correspond to $x, y, z$ respectively.

Generic index values are represented by Greek-alphabet letters. When the same index of this sort appears twice in an expression, summation is implied (unless otherwise stated). When the same index appears on either side of an equality, then several equalities are being asserted, one for each value (or each combined value if multiple such pairs of “loose” indices appear).

When transforming, we sometimes use $\mu, \nu, \lambda, \ldots$ indices for components with respect to the original coordinate system and $\alpha, \beta, \gamma, \ldots$ for the transformed system.

Often, a 4D quantity has a name similar to that of the 3D quantity related to its spatial components.

When the same letter of the alphabet is used for both a 3-vector and a 4-vector, it is understood that the spatial part of the 4-vector is the same as the corresponding 3-vector in some inertial coordinate system. Thus, for example, the $x$-component of relativistic momentum can be called either $p_x^1$ or $\vec{p}_x$.

The usage of tilde (active) and prime (passive) is the same as for three-dimensional objects.

- $\partial$ Spacetime gradient operator [dimension $L^{-1}$].
- $\left(\begin{array}{c}p \\ q \end{array}\right)$ Denotes the rank of a tensor with $p$ upper and $q$ lower indices.
- $g$ Denotes the tensor whose components in any inertial coordinate system form the matrix diag($-1, +1, +1, +1$).
Spinors

See Section 34.11’ (page 545).

Matrices and quantum mechanical operators

These objects are set in sans-serif type, $\mathbf{M}$. They are arrays of numbers that do not necessarily transform in the specific manner of tensors upon coordinate change. See Appendix D for more.

When matrices appear next to each other, the usual rules of matrix multiplication are implied. In addition, sometimes we will use square brackets such as $\left[ \vec{r} \right]$ to denote the three cartesian components of $\vec{r}$ in a particular coordinate system, regarded as a column vector and subject to matrix multiplication rules; similarly $\left[ \vec{K} \right]$ refers to a $3 \times 3$ matrix and $\left[ \hat{F} \right]$ to a $4 \times 4$ matrix. This notation can be very compact, though it is only useful for objects of rank less than three, and only when all contractions involve the column index of one object in a string, and the row index of the next one. More complex formulas require that we use explicit index notation.

Relations

$\sim$ Has the same dimensions as.
$\approx$ Is approximately equal to.

Miscellaneous

When a statement involves a single $\pm$ sign, it’s an abbreviation for making two statements. When a statement involves two $\pm$ signs, it’s still an abbreviation for making two statements: One where both of them represent $+$, and another where both represent $\lambda$. When a statement uses a $\pm$ sign and also a $\mp$ sign, it’s still an abbreviation for making two statements, but with opposite signs.

The usual square root of minus one is indicated in roman type ($i$) to distinguish it from say, an index. Some engineering texts instead use the letter $j$ to represent this quantity. Some computer math systems instead refer to this quantity as $I$ or as $j$. (The other square root of minus one is then $-i$.)

The base of natural logarithms is indicated in roman type ($e$) to distinguish it from the charge on a proton ($e$), a constant of Nature.

The differential symbol is indicated in roman type ($d$) to distinguish it from any variable called $d$, which might denote a distance.

B.2 DIMENSIONS AND UNITS

See Appendix A and Chapter 16.
Appendix B Global List of Symbols

B.3 NAMED QUANTITIES

We have a lot of quantities, and only a limited number of letters of the alphabet, so inevitably some symbols will be overloaded with more than one meaning. Sometimes the meanings will be disambiguated by upper/lower case, or by tensor rank. In other cases, you just have to determine the desired meaning by context.

Latin alphabet

- \( a \) Size of a finite distribution of charge and/or current, often taken to be much smaller than some other scale.
- \( A \) Three-dimensional magnetic vector potential.
- \( A \) Four-vector potential.
- \( b \) Generic name for a constant. \( B \), generic name for the amplitude of a sinusoidally-varying quantity.
- \( B_{ij} \) Shape operator for a 2D surface in 3-space.
- \( \bar{B} \) Magnetic induction (often called “magnetic field”); \( \bar{B} \), modified form, = \( c\bar{B} \) [same dimensions as electric field, \( \text{ML}^{-2}\text{T} \)].
- \( c \) Speed of light in vacuum. \( c_s \), speed of vibrations in a medium, e.g. a spring.
- \( c_{Na}^+ \), etc. Number density (concentration) of an ion species or other discrete, countable class of objects, usually in water solution [dimensions \( \text{L}^{-3} \)].
- \( C \) Capacitance.
- \( \mathcal{C} \) Areal density of capacitance.
- \( D_{ion} \) Diffusion constant for some species of ions in solution.
- \( D_r \) Retarded green function for the wave operator (dalemberttian).
- \( \mathcal{D} \) Electric displacement (analog of \( \mathcal{E}_0 \mathcal{E} \) in a medium).
- \( \mathcal{D}_r \) Electric dipole moment. \( \mathcal{D}_r \), its quantum version (Section 55.5.2, page 777) [dimensions \( \text{Q}\text{L} \)]. \( D_e \), the length of its matrix element (the “transition dipole”).
- \( \mathcal{D}_m \) Magnetic dipole moment; \( \mathcal{D}_m = \mathcal{D}_m/c \), modified form (with same units as electric dipole moment).
- \( e \) Charge on a proton.
- \( \hat{e}_{(i)} \) Basis of three mutually perpendicular 3-vectors, possibly depending on a chosen basepoint, not necessarily of unit length. Each has three components: \( \hat{e}_{(i)} \).
- \( \hat{e}_{(i)} \) Normalized basis of three mutually perpendicular 3-vectors. If they are associated to the main cartesian coordinate system being used, then they have abbreviated names \( \hat{x} = \hat{e}_{(1)} \), \( \hat{y} = \hat{e}_{(2)} \), and \( \hat{z} = \hat{e}_{(3)} \).
- \( \mathcal{E} \) Electric field [dimensions \( \text{ML}^{-2}\text{T} \)].
- \( \mathcal{E} \) Energy, usually the relativistic (correct) form. \( \mathcal{E} \), specifically the relativistic energy when it is necessary to distinguish it from the newtonian quantity \( \mathcal{E}_N \).
- \( \mathcal{E}_{\text{FRET}} \) Fluorescence resonance energy transfer efficiency [dimensionless] (Chapter 4).
- \( f \) Focal length of a lens [dimensions \( \text{L} \)] (Section 41.4.2, page 623).
B.3 Named Quantities

807

Linear tension, for example in a spring or along a 1D interface [dimensions of force] (Chapters 7, 27).

\( \mathcal{F} \) Helmholtz free energy.

\( F \) Faraday 4-tensor.

\( g \) Conductance per area.

\( \gamma \) Metric in two or three dimensions.

\( g \) Metric 4-tensor. In special relativity, this is a rank-(\( \frac{0}{2} \)) tensor whose 16 components in any E-inertial coordinate system, \( g_{\mu\nu} \), are always the same numerical constants. The same letter \( g \) can also be used to refer to the dual metric tensor, a rank-(\( \frac{2}{0} \)) tensor whose 16 components in any E-inertial coordinate system, \( g^{\mu\nu} \), are the same numerical constants as those of \( g_{\mu\nu} \). The notation is unambiguous because applying the index-raising operation to the first version does yield the second one.

\( G \) Conductance.

\( G \) Gauss curvature of a surface in space (Chapter 7).

\( G_N \) Newton gravitation constant.

\( \hat{h} \) Displacement (position) of an object relative to the origin of coordinates or other reference point. \( h \), generic symbol for a distance.

\( h \) reduced Planck constant [dimensions \( mL^2T^{-1} \)].

\( H \) Mean curvature of a surface in space (Chapter 7).

\( \bar{H} \) Magnetic intensity (analogous to \( B/\mu_0 \) but includes a medium).

\( H \) Quantum hamiltonian operator (Chapter 55) [dimensions \( LT^{-2} \)].

\( I \) Electric current [dimensions \( QT^{-1} \)]. \( I_x \) axial current in a cable. \( I_r \), radial (“leak”) current in a cable.

\( \tilde{I} \) Matrix form of the Stokes parameters (Chapter 24) [dimensions \( (V/m)^2 \)].

\( J \) Linear density of a line current source (Section 8.7.2, page 122) [dimensions \( QT^{-1}L^{-1} \)].

\( \bar{J} \) Moment of inertia tensor of a rigid body.

\( j \) Electric charge flux [dimensions \( QL^{-2}T^{-1} \)]; \( j_f \), free part; \( j_{b,p} \), bound part from electric polarization; \( j_{b,M} \), bound part from magnetic polarization; \( j^{[1D]} \), one-dimensional version; \( j^{[2D]} \), 2D charge flux in a surface (sometimes called “surface current density”); \( j_f^{[2D]} \), free part.

\( \tilde{J} \) Quantum charge flux operator.

\( \bar{J} \) Flux of energy.

\( J_{\text{ion}} \) Number flux of ions of some species [dimensions \( L^{-2}T^{-1} \)].

\( J \) Electric charge 4-flux (sometimes called “4-current”); \( J \), scalar analog sometimes used in simplified formulas.

\( \tilde{J} \) Jones tensor (on the 2D space of directions transverse to plane wave wavenumber) (Section 46.3, page 676).

\( J \) Source term for scalar wave equation (Equation 25.6, page 393).
\[ J \] Generic conserved 4-flux arising from a continuous symmetry (Equation 40.16, page 612).

\[ k \] Generic name for a Hooke-law spring constant.

\[ k_B \] Boltzmann constant; \( k_B T \), thermal energy; \( k_B T_r \), at room temperature.

\[ K \] Temporary name for relativistic energy/\( c \), later named \( p^0 \).

\[ \tilde{K} \] Hooke-law spring tensor.

\[ K^{\mu \nu}_{\lambda \sigma} \] 4D susceptibility operator (Section 51.3b, page 721).

\[ \ell \] Generic index for enumeration, for example, a set of particles or elements of a continuous source. Can also indicate which of several ion species is under consideration.

\[ \ell_B \] Bjerrum length (Equation 10.29).

\[ \tilde{\ell} \] Parametric representation of a generic curve in space; \( d \tilde{\ell} \), small element.

\[ L \] Inductance.

\[ \tilde{L} \] Angular momentum.

\[ \mathcal{L} \] Lagrangian density (Section 40.2, page 606–40.3).

\[ m \] Mass [dimensions \( \text{M} \)]; \( m_e \), mass of electron.

\[ m \] Generic 3-space index.

\[ \tilde{M} \] Volume density of magnetic dipole moment; \( \tilde{M} = \tilde{M}/c \), modified form (same units as \( \tilde{P} \)).

\[ n \] refractive index.

\[ n_{\kappa, \alpha} \] Photon occupation number (Equation 55.26, page 774) [dimensionless].

\[ n_e \] Number density of unbound electrons (usually in plasma) or of electrons that could potentially unbind [dimensions \( \text{L}^{-3} \)].

\[ M^{\mu \nu \lambda \sigma} \] Angular momentum flux tensor (Section 35.5, page 559).

\[ p \] Order of a multipole (called a “2\( p \)-pole”), equal to the rank of the 3-tensor that specifies it. Rank of a generic 3-tensor.

\[ p \] Pressure [dimensions \( \text{ML}^{-1} \text{T}^{-2} \)].

\[ \tilde{p} \] A particle's 3-momentum, usually the relativistic (correct) form. \( \tilde{p} \), specifically the relativistic momentum when it is necessary to distinguish it from the newtonian quantity \( p^N \).

\[ \tilde{P} \] A particle’s 4-momentum.

\[ p_e \] Quantum electron momentum operator (Section 55.5.2, page 777) [dimensions \( \text{ML}^2 \text{T}^{-1} \)].

\[ \mathcal{P} \] Volume density of electric dipole moment (“polarization density”).

\[ \tilde{P}_{\text{field}} \] momentum of electromagnetic field (Equation 55.8, page 771) [dimensions \( \text{ML}^2 \text{T}^{-1} \)]. \( \tilde{P} \), corresponding quantum operator.

\[ \mathcal{P} \] Power [dimensions \( \text{ML}^2 \text{T}^{-3} \)].

\[ \text{Prob} \] Probability (a real, dimensionless quantity between 0 and 1). \( \phi(x) \), Probability density function for a continuous random variable \( x \) [dimensions match those of \( 1/x \)].
q Electric charge.
Q, \( Q^+ \) lowering (destruction) and raising (creation) operators, respectively, for electromagnetic field (Equation 55.17, page 772) [dimensionless].
\( \tilde{Q}_e \) Electric quadrupole 3-tensor. \( \tilde{Q}_m \) Magnetic quadrupole 3-tensor.
r Radial coordinate in spherical polar system (distance from origin).
\( r_c \) Classical electron radius.
\( \tilde{r} \) Three-dimensional position vector, with cartesian components \( \tilde{r}_i = (x, y, z)^t \). Sometimes specifically the field point (observer location); then \( \tilde{r}_e \) denotes source point.
\( \tilde{r}_e \) Electron position [dimensions L]; \( \tilde{r}_e \), corresponding quantum operator (Equation 55.35, page 778).
\( r_F \) Förster radius (Section 4.3.2, page 62) [dimensions L].
\( \tilde{R} \) Displacement between source point and field point; \( R_{\text{traj}} \), for field point evaluated somewhere on a particle trajectory.
\( R \) Electrical resistance. \( R_x \), axial resistance along a cable. \( R_r \), radial (“leak”) resistance out of a cable.
\( \mathcal{R} \), reflection factor [dimensionless].
s Arc length parameter along a curve in 3-space.
\( s_i \) Individual Stokes parameters (Chapter 24) [dimensions (V/m)^2].
\( S \) Action functional (Sections 40.2–40.3, page 606) [dimensions ML/T].
\( S \) Eikonal function describing a collection of rays [dimensions L].
\( S \) A 3D rotation, or the \( 3 \times 3 \) matrix representing it; \( S_{ij} \), its explicit components.
t Time, as measured in an inertial coordinate system (either G-inertial in newtonian physics or E-inertial in relativistic physics). Sometimes specifically the time of an observation; then \( t_e \) denotes source time. \( t_r \), retarded time (intersection of a particle trajectory with the past light cone of the observation event). \( t_c \), observer’s time minus (distance to center of a source)/c.
\( T \) Interfacial surface tension (Chapter 7) [dimensions \( \text{ML}^2 \text{T}^{-2} \)].
\( T \) Absolute temperature.
\( T \) Transmission factor [dimensionless].
\( \tilde{T} \) Momentum flux 3-tensor (called “stress tensor” in some books).
\( T \) energy–momentum flux tensor (called “stress-energy tensor” in some books).
\( T \) generator of a rotation (Equation 3.12, page 46).
u, v Light-cone coordinates.
u Displacement of a continuous spring.
\( U \) Potential energy of a particle.
\( U \) Four-velocity. Its three spatial components are not equal to the components of ordinary velocity \( \dot{u} \).
\( u_m \), velocity of a material medium that supports waves (spring, water, æther, ...).
v Depolarization (\( \Delta \psi \) shifted by \( \psi^0 \)); \( u_1 \) and \( u_2 \), special fixed-point values (Figure 12.4); \( \tilde{\psi}(t) \), depolarization waveform of a traveling wave; \( \tilde{\psi} \), dimensionless rescaled form.
Appendix B Global List of Symbols

\( \vec{u} \) Velocity; that is, the time derivative of the position of a particle in an inertial coordinate system (either G-inertial in newtonian physics or E-inertial in relativistic physics). \( \vec{u'} \), velocity of a Galilean or Lorentz boost.

\( V \) A region in 3-space, or its volume; \( \partial V \), the boundary of \( V \), that is a closed surface. An area element \( d^2\vec{S} \) of that surface is conventionally taken to point outward.

\( w \) Generic length variable, for example, thickness of a layer.

\( \mathcal{W} \) Hamilton’s principal function (Section 21.7, page 334) \[ \text{dimensions} \ M L T^2 \].

\( x, y, z \) Right-handed cartesian coordinates of 3-space, or spatial components of a right-handed E-inertial coordinate system on spacetime. \( \hat{x}, \hat{y}, \hat{z} \), corresponding unit vectors, also called \( \hat{e}_1, \hat{e}_2, \) and \( \hat{e}_3 \).

\( \mathcal{X} \) Four-vector coordinates of an event. Sometimes specifically the field (observation) point; then \( \mathcal{X}_s \) is the source event.

Greek alphabet

\( \alpha \) Electric polarizability of a molecule or other small object; \( \alpha_m \), magnetic polarizability. \( \vec{\alpha} \), polarizability tensor of an anisotropic object.

\( \beta \) Cross-polarizability of a single chiral molecule.

\( \vec{\beta} \) Velocity of a particle divided by \( c \).

\( \gamma \) Abbreviation for \( 1/\sqrt{1 - \vec{\beta}^2} \) (Section 30.3.1, page 454).

\( \gamma \) Another generic quantity name.

\( \Gamma(\xi) \) Parametric representation of a 4D trajectory (curve in spacetime). \( \vec{\Gamma} \), the spatial part of such a trajectory, for example, \( \vec{\Gamma}(t) \), a trajectory specifically parameterized by lab time.

\( \vec{\Gamma} \) Dual representation of the magnetic dipole moment as an antisymmetric 3-tensor of rank 2.

\( \delta^{(n)} \) Product of \( n \) Dirac delta functions \[ \text{dimensions} \ M L T^2 \].

\( \Delta \) Amount by which some quantity changes. Usually used as a prefix: \( \Delta x \) denotes a small change in \( x \).

\( \varepsilon \) Dielectric permittivity of a medium \[ \text{dimensions} \ Q^2 M^{-2} L^{-3} \]; \( \varepsilon_0 \), permittivity of vacuum. The dimensionless ratio \( \varepsilon/\varepsilon_0 \) is called the “dielectric constant,” but we don’t assign any symbol to it.

\( \varepsilon_{ijk} \) Components of the 3D Levi-Civita tensor in a particular cartesian coordinate system. In a right-handed system, \( \varepsilon_{123} = +1 \). \( \varepsilon_{\mu\nu\lambda} \), components of the 4D Levi-Civita tensor in a particular E-inertial coordinate system. In a right-handed system, \( \varepsilon_{0123} = +1 \).

\( \varepsilon_{\text{multi}} \) Multipole parameter (Equation 43.8) \[ \text{dimensionless} \].

\( \varepsilon_{a\beta} \) and \( \varepsilon_{\dot{a}\dot{\beta}} \), spinor metrics (Equations 34.31 and 34.36, page 548).

\( \vec{\xi} \) Polarization 3-vector (complex amplitude for vector potential) for a plane EM wave; \( \xi_{(1)}, \xi_{(2)} \), linear polarization basis (real); \( \xi_{(+)}, \xi_{(-)} \), circular polarization basis (complex); \( \xi_{(\alpha\beta)} \), basis of unit polarization vectors (\( \alpha = 1, 2 \)) for plane waves traveling...
B.3 Named Quantities

along $\vec{k}$ (Equation 55.2, page 769) [dimensionless].

$\xi_{(x;k)}$ Basis of unit polarization vectors ($x = 1, 2$) for plane waves traveling along $\vec{k}$ (Equation 55.2, page 769) [dimensionless].

$\xi$ Polarization 4-vector (complex amplitude for 4-vector potential of a plane wave).

$\eta$ Viscous drag coefficient for a particle in fluid; $\vec{\eta}$, same for anisotropic particle [dimensions $\text{ML}^{-1}$].

$\eta$ Bulk cross-polarizability of a chiral material.

$\eta_l$ Integers specifying a mode in a cavity (Section 55.2, page 769) [dimensionless].

$\vartheta$ Polar angle in spherical polar coordinates [dimensionless].

$\vartheta$ Angle between an incoming wave’s linear polarization and the line of sight to an observer.

$\vartheta$ Velocity of neural action potential.

$\Theta$ Step function [dimensionless].

$\kappa$ Electric conductivity of a medium; elastic stretch modulus of a continuous spring; curvature of a curve in a plane (Chapter 7).

$\lambda$ Wavelength of a plane or spherical wave.

$\lambda_0$ Debye length.

$\lambda_c$ Space constant of a nerve axon or other cable.

$\Lambda$ A Lorentz transformation linking two E-inertial coordinate systems, or the $4 \times 4$ matrix representing it; $\Lambda_{\nu\mu}$, its explicit components.

$\mu$ Magnetic permeability of a medium [dimensions $\text{ML}^{-1}Q^{-2}$]; $\mu_0$, permeability of vacuum.

$\nu$ Circular frequency (often abbreviated “frequency”) of a sinusoidally varying quantity (cycles per unit time) [dimensions $T^{-1}$].

$\xi$ Generic parameter for a curve in space (not necessarily arc length) or spacetime (not necessarily proper time); $\vec{\xi}$, constant 3-vector used when constructing a dipole spherical wave.

$\Xi$ Gauge-transformation parameter.

$\rho$ Radial coordinate in cylindrical coordinate system.

$\rho$ Generic symbol for volume density of a quantity that has dimensions; $\rho_q$, electric charge density [dimensions $\text{QML}^{-3}$]; $\rho_f$ and $\rho_b$, free and bound electric charge densities; $\rho_\xi$, energy density; $\rho_m$, mass density.

$\rho_q^{[1D]}$, linear electric charge density (coul/m); $\rho_\xi^{[1D]}$, linear energy density; $\rho_m^{[1D]}$, linear mass density.

$\sigma$ Generic symbol for areal density of a scalar quantity; $\sigma_q$, areal charge density; $\sigma_f$, free areal charge density; $\sigma_b$, bound areal charge density.

$\sigma$ Scattering cross section.

$[\vec{\sigma}]$ Three Pauli matrices (Equation 34.34, page 547); $[\sigma]$, Pauli matrices augmented with identity matrix (Equation 34.38, page 549).
\( \Sigma \) A 2D surface, or its area; \( d\hat{\Sigma} \), infinitesimal surface element, including a choice of perpendicular vector, that is, differential of area times the chosen unit vector. \( \partial\Sigma \), boundary of a surface \( \Sigma \), that is, a closed curve with a direction chosen by applying the right-hand rule to the chosen perpendicular.

\( \tau \) A particle's proper time; equivalently, proper time parameter along a trajectory in spacetime; equivalently, the time recorded by an imagined clock carried along with the particle. If the particle's trajectory is accelerating, then proper time will not agree with time \( t \) in any fixed E-inertial coordinate system. The proper time difference squared between two time-like separated events is also the invariant interval between them.

\( \tau_{\text{cable}} \) Time constant of a nerve axon or other cable.

\( T \) Rapidity parameter of a Lorentz boost.

\( \phi \) Azimuthal angle in either cylindrical or spherical polar coordinate system [dimensionless].

\( \phi \) Phase shift of one sine function relative to another.

\( \phi_N \) Newtonian gravitational potential.

\( \Phi_{\kappa, \omega} \) The complex function \( e^{i(k \cdot \hat{r} - \omega t)} \) (dimensionless).

\( \Phi_B \) Integral of \( \vec{B} \cdot d\hat{\Sigma} \) over an area. \( \Phi_B = \Phi_B / c \), modified version.

\( |\Phi\rangle \), vector in quantum-mechanical state space.

\( \chiE \) Dielectric susceptibility (bulk electric polarizability of a medium divided by \( \varepsilon_0 \)); \( \chi_m \), magnetic susceptibility (related to bulk polarizability of a medium); \( \tilde{\chi}_m \), modified form. For anisotropic media, these are replaced by tensors.

\( \psi \) Scalar potential field, also called electric potential [dimensions \( \text{ML}^2 \text{T}^{-2} \text{Q}^{-1} \)]. In electrostatics, also called the electrostatic potential. \( \tilde{\psi}_s \), its dimensionless form (in static or quasi-static situations), \( \tilde{\psi} \), amplitude of a potential varying sinusoidally in time. \( \psi_{[p]} \), standard 2\( p \)-pole potentials. \( \psi_{\text{in}} \), potential inside a neuron; \( \psi_{\text{out}} \), potential outside (often taken to be zero). \( \psi_{\text{Nernst}} \), Nernst potential; \( \psi^0 \), quasisteady resting potential; \( \nu \), membrane potential relative to \( \psi^0 \).

\( \omega \) Angular frequency (radians per unit time).

\( \omega_p \) Plasma frequency.

\( \tilde{\omega} \) Angular frequency of rigid body rotation, with direction corresponding to its axis of rotation via the right-hand convention.

\( \tilde{\omega} \) Dual representation of \( \vec{B} \) as an antisymmetric 3-tensor of rank 2 (Equations 15.2–15.3).

\( \Omega \) Solid angle (sometimes called angular area).
Appendix C

Numerical Values

If the model explains all the facts, then there’s something wrong—because always some of the facts are wrong.

—Aharon Katchalsky

C.1 FUNDAMENTAL CONSTANTS

Newtonian gravitation constant, \( G_N \approx 6.7 \cdot 10^{-11} \text{ m}^3\text{kg}^{-1}\text{s}^{-2} \).

Planck constant (reduced), \( \hbar \approx 1.05 \cdot 10^{-34} \text{ J s} \).

Proton charge, \( e \approx 1.6 \cdot 10^{-19} \text{ coul} \). Electron charge is \( -e \). Useful: \( e^2/(4\pi\varepsilon_0) \approx 1.44 \text{ eV nm} = 1.44 \text{ MeV fm} \).

Electron mass, \( m_e \approx 9.1 \cdot 10^{-31} \text{ kg} \).

Speed of light, \( c \approx 3.0 \cdot 10^8 \text{ m/s} \).

Avogadro number, \( N_{\text{mole}} \approx 6.02 \cdot 10^{23} \).

Boltzmann constant, \( k_B \approx 1.38 \cdot 10^{-23} \text{ JK}^{-1} \). Typical thermal energy at room temperature \( k_B T \approx 4.1 \text{ pJ nm} \approx 4.1 \cdot 10^{-21} \text{ J} \approx 2.5 \text{ kJ mole}^{-1} \approx 0.59 \text{ kcal mole}^{-1} \approx 0.025 \text{ eV} \).

Permittivity of vacuum, \( \varepsilon_0 \approx 8.85 \cdot 10^{-12} \text{ coul}^2\text{N}^{-1}\text{m}^{-2} = 8.85 \cdot 10^{-12} \text{ F/m} \). Permeability of vacuum, \( \mu_0 \approx 4\pi \cdot 10^{-7} \text{ m kg coul}^{-2} \).

C.2 MATERIAL PROPERTIES AT LOW FREQUENCY

Resistivity of salt solution at concentration 100 mM: \( \kappa \approx 0.1 \Omega^{-1}\text{m}^{-1} \).

Liquid water: \( \varepsilon \approx 80\varepsilon_0 \) at low frequency.

C.3 PROPERTIES AT OPTICAL FREQUENCIES

C.3.1 Refractive index for visible light

These approximate values neglect dispersion (dependence on wavelength).

Air at standard temperature and pressure: \( n_{\text{air}} \approx 1.0003 \). This book uses the approximate value 1, except when studying the mirage phenomenon; there, we use more precise values for light of wavelength 633 nm. At 30 °C: \( n_{\text{air}} \approx 1.00026 \); at 50 °C: \( n_{\text{air}} \approx 1.00024 \).

Water at 20°: \( n_w \approx 1.33 \) for 650 nm (red) light; \( \approx 1.34 \) appropriate 400 nm (blue) light.
Appendix C Numerical Values


Fish eyeball lens: Typically $n \approx 1.52$ near the center, shading continuously to $\approx 1.38$ at the periphery.

C.4 MISCELLANEOUS

Earth’s magnetic field strength at surface, approx $5 \cdot 10^{-5}$ T.

Earth radius $6.4 \cdot 10^6$ m.

Maximum energy of solar radiation per area at Earth surface: $1.4$ kW/m$^2$.

Mass of Sun $2.0 \cdot 10^{30}$ kg.
Matrix Refresher

(Unit matrix.
S 3D rotation matrix.
Λ 4D Lorentz transformation matrix.

Matrix transpose (exchange rows and columns). A matrix satisfying $A = A^t$ is said to be symmetric. In $d$ dimensions, a symmetric matrix has $d(d + 1)/2$ independent entries, because those below the diagonal duplicate those above it.

A matrix satisfying $A = -A^t$ is said to be antisymmetric. It must therefore equal zero along its diagonal, and it has $d(d - 1)/2$ independent entries.

If $A$ is (anti)symmetric, then the same is true of $B^t A B$.

If instead the transpose of a real matrix $M$ equals the inverse, then $M$ is called orthogonal.

$M^*$ Complex conjugate each entry. If it equals $M$, then the matrix is real.

$M^\dagger$ Hermitian conjugate (also called “hermitian adjoint”), same as $M^t \dagger$. The superscripted symbol is traditionally pronounced “dagger.” $M$ is said to be hermitian if it equals its hermitian conjugate. If instead the hermitian conjugate equals the inverse, then $M$ is called unitary.

$$(AB)^\dagger = A^\dagger B^\dagger$$
$$(AB)^{-1} = A^{-1} B^{-1}$$
$$(A^{-1})^\dagger = (A^\dagger)^{-1}$$
$$\det(AB) = \det A \det B$$

A symmetric real matrix has a set of real eigenvectors (“principal directions”) that span its entire space and that are mutually perpendicular.
Animated graphics

The ability to create scientific animations will be very valuable to you, for example, to create a striking graphic for a presentation. Many problems in this book ask you to exercise this skill. You may be able to do a video screen capture of your animation as it runs in your favorite math software. But image quality may be poor, and it may not run at the desired frame rate.

If you do your computing with Python, then note:

- Step one is to make an animation that plays within Python. You may find the free Celluloid module: github.com/jwkvan/celluloid to be helpful, or the more daunting but more flexible FuncAnimation from the matplotlib.animation module.¹
- Next you must save to a common video format:
  - You can directly generate a file in one of the common video formats (mp4, mpg, mov), which in turn will be viewable in a browser, embeddable into a presentation, uploadable to YouTube, Vimeo, and so on (and from there to your social media pals!) by installing a separate application called ffmpeg in a place where Python can access it, then using the save method of an animation created in Python, which calls ffmpeg.²
  - Without installing any extra software, you can generate gif animation files with PillowWriter.
  - Without installing any extra software, you can generate an HTML5 video file (viewable in a browser) by using the to_jshtml method of an animation created in Python.

An alternative available for any math software is to create a folder containing many still images (individual video frames). You can then try one of these:

- Call ffmpeg, or some other encoder for rendering, from your system’s command-line prompt. You may need the obscure option -pix_fmt yuv420p to generate movies viewable on other platforms, for example:
  
  $ ffmpeg -i frames%d.png -pix_fmt yuv420p movmovie.mp4

- macOS: Open QuickTime Player, hit File/Open Image Sequence, select all of the image files and hit Choose Media.
- Windows or macOS: Other free software such as VLC or ImageJ may be able to turn still images into a video format.

²Users of Google Colab may find that ffmpeg or something equivalent is automatically available.
APPENDIX F

Formulas

[[Not ready]]
Glossary

Here are a number of terms with brief definitions emphasizing the aspects discussed in this book. Beware that various authors use slightly different shades of meaning for some of these terms. In some cases, further details are available via bold entries in the Index. Some terms not listed here may nevertheless have such entries.

A few of the words and phrases listed here are flagged below as *not* used in the main text, but are in common enough use elsewhere to be worth connecting to the equivalents used in this book.

**ABERRATION, CHROMATIC**: Effect of dispersion in the material making up a lens, leading to poor image formation.

**ABERRATION, SPHERICAL**: Failure of a lens with perfectly spherical surfaces to focus light perfectly to a pointlike spot.

**ABERRATION, OF STARLIGHT**: Apparent displacement of distant sources of light relative to each other, caused by motion of observer.

**ACausal**: Leading to a violation of causality and hence forbidden.

**ACCEPTOR FLUOROPHORE**: One whose excitation peak is matched to another (donor) fluorophore's emission peak.

**ACHIRAL**: Synonym for **NONCHIRAL**.

**ACTION FUNCTIONAL**: Assigns a single number (“the action”) to a complete history of a physical system. Evaluated by substituting the trajectory’s position and velocity functions into a lagrangian density function, then integrating over time or spacetime.

**ACTION POTENTIAL**: Spacetime course of an electric disturbance along a nerve axon, assuming voltage gating. Mathematically modeled in this book as a traveling wave solution to the nonlinear cable equation. Sometimes referred to as “nerve impulse.”

**ACTIVE TRANSFORMATION**: See **SYMMETRY**.

**Æther**: Hypothesized material substance allowing the propagation of electromagnetic effects through vacuum; the longer name **Luminiferous Æther** once discriminated it from other now-forgotten Æthers.

**AFTERHYPERPOLARIZATION**: See **HYPERPOLARIZATION**.

**AlANINE**: An amino acid.

**AMORPHOUS**: Said of a material with a disorganized (noncrystalline) arrangement of constituent molecules, such as window glass.

**AMO**: Atomic, molecular, and optical (physics).

**ÅNGSTROM UNIT**: 0.1 nm, a non-SI unit of length.

**ANGULAR AREA**: Same as solid angle. **ANGULAR VELOCITY**: Rate of change of an angle in radians per time unit.

**ANION**: Ion attracted to an anode, that is, one carrying negative net charge.

**ANISOTROPIC, ANISOTROPY**: Said of an atom that is not spherical, or of a material with a tensor property that is not rotationally symmetric (such as susceptibility).

**ANISOTROPY**: Lack of **ISOTROPY**. An anisotropic medium is not rotationally invariant.

**ANNIHILATION**: Quantum-mechanical property in which two antiparticles, for example electron and positron, both disappear with a corresponding increase in the energy of the rest of the world, for example via the creation of two photons.
ANTIPARTICLES: Two species of particle having all additive conserved properties opposite to each other (for example, charge, lepton number, and so on) and able to annihilate each other.

ANTISYMMETRIC TENSOR: One whose component $T_{ijk}$ equals minus the corresponding one after a specified index pair have been exchanged, for example, “antisymmetric on $i$ and $k$.” TOTALY ANTISYMMETRIC: For any permutation $\sigma$ of the indices, let $(-1)^{|\sigma|}$ denote $-1$ for odd parity or $+1$ for even parity. Then $T_{\epsilon(\sigma ij)} = (-1)^{|\sigma|}T_{ij}$. ANTISYMMETRIZE a tensor: To force antisymmetry by adding/subtracting corresponding elements, for example, $T_{ijkl} = (T_{ij} - T_{ji})/2$ is antisymmetric on $i$ and $j$.

ARC LENGTH: Distance along a curve; what we’d get by laying a flexible tape measure along the curve.

AREAL DENSITY: When a quantity such as electric charge or capacitance is distributed over a 2D surface (or a small element of such a surface), the total divided by the area of the region in question.

AROMATIC: Describes a class of molecules with a ring structure.

AUTOMATICALLY SATISFIED: Property of being always true regardless of the values of physical quantities; said of mathematical identities (see IDENTITITY).

AXISYMMETRIC: Describes a shape that is unchanged upon rotations about an axis.

AXON: Long tube extending from a neuron and sharing its interior composition; nerve impulses travel along the cell membrane covering its axon. The “nerves” of anatomy are generally bundles of axons. MYELINATED, UNMYELINATED denote the presence or absence of a separate insulating sheath outside the cell membrane, interrupted by “nodes of Ranvier.”

AXOPLASM: Material contained inside a neuronal axon.

AZIMUTHAL ANGLE: Coordinate describing rotation around an axis (usually for an axisymmetric body), generally called $\varphi$; analogous to longitude in geography.

BANDGAP, PHOTONIC: Ability for a structured dielectric composite to reflect completely plane EM waves with certain frequencies, despite the fact that each constituent is transparent.

BASEPAIR: A DNA molecule consists of a stack of nearly planar atomic groups called basepairs, each covalently bound to the next by a “backbone.”

BEAM SPLITTER: Typically a glass optical element with a coating of silver that is too thin to reflect all incoming EM radiation (“half-silvered”). Instead, some fraction is reflected while the rest is transmitted, splitting an incoming beam.

BEAMING: Emission of EM radiation that at any one instant is tightly bunched into a narrow range of directions, in contrast to the broadly emitted radiation from a nonrelativistically moving charge distribution.

BEAM, BESSEL: Beam of light with very low diffractive spreading. GAUSSIAN: Beam with transverse profile of energy flux given by a gaussian function. LAGUERRE–GAUSSIAN: A particular family including gaussian and more complicated beams. VORTEX: Generic term for nonaxisymmetric beams. WAIST: Narrowest part of a beam, or that part’s diameter.

BIAXIAL: Said of an object for which the eigenvalues of the quadrupole moment tensor (mass or charge), or of the polarizability tensor, all have different values (no two are equal).

BIFURCATION: Sudden gain or loss of stationary (critical) points in a function as a parameter is changed. For example, in mechanics, this can include loss of stability as an object becomes too top-heavy (at “tipping point”).

BILAYER MEMBRANE: A self-assembled molecular structure: Two layers of molecules held together by noncovalent interactions. Each molecule has a nonpolar, hydrocarbon chain facing the opposite layer, and a polar headgroup facing outward into surrounding water.

BILINEAR: Said of a function with two arguments that is linear in each one. For example, both the metric and any spring tensor specify bilinear functions of two vectors. Similarly TRILINEAR, MULTILINEAR.

BIREFRINGENCE, CIRCULAR: Property of a material that the two circular polarizations of light each propagate unchanged but with different speeds. Generally associated with chiral materials, possibly disordered. ORDINARY BIREFRINGENCE: Property of a material that in certain directions, two linear polarizations of light each propagate unchanged but with different speeds. Generally associated with crystalline materials.

BJERRUM LENGTH: Length scale relevant for a pure solvent, depending only on its permittivity and the temperature.
**BOOST, LORENTZ OR GALILEAN:** Transformation from one inertial coordinate system to another that is in uniform, straight-line motion relative to the original one.

**BORN SELF-ENERGY:** Electrostatic self-energy of an object embedded in a medium.

**CAPILLARY TUBE:** Glass tube of very small diameter.

**CATION:** Ion attracted to a cathode, that is, one carrying positive net charge.

**CAUSALITY, CAUSAL:** Refer to the requirement that a causative agent must act earlier in time than its effect.

**CAUSTIC:** A point, line, or surface in space at which geometrical optics predicts infinite intensity of light. Special cases include the focal point of a perfect lens, or a rainbow.

**CELL, VOLTAIC:** Drives electrons against a potential gradient by releasing chemical energy.

**ČERENKOV RADIATION:** Electromagnetic radiation emitted when a charged particle passes through a transparent medium at speed greater than \( c/n \).

**CHARGE DENSITY, BOUND:** Net charge density attributable to permanent or induced charge separation over submolecular distances, of charges that are not free to move longer distances. **FREE:** Contribution attributable to charges that are free to move arbitrary distances.

**CHARGE FLUX, BOUND:** Net charge flux attributable to time-varying induced charge separation over submolecular distances, of charges that are not free to move longer distances. **FREE:** Contribution attributable to charges that are free to move arbitrary distances.

**CHIRAL, CHIRALITY:** Describes an object that cannot be made to coincide with its mirror image by any translation or rotation. Synonym “handed.” **MEDIUM:** One that either consists of molecules that are individually chiral, or that are arranged on a chiral crystal lattice. **CHIRAL FERMION:** Class of particles that can adopt only one helicity, typically represented by quantizing a Weyl field.

**CHROMOPHORE:** A molecule (or part of a molecule) that absorbs light preferentially in a particular wavelength band, imparting visual color by reflecting or transmitting other wavelengths. Visual chromophores also initiate neural signaling by undergoing a conformational change.

**CLOCKWISE, ANTICLOCKWISE:** When a rigid body rotates about an axis, that motion can be described unambiguously by an antisymmetric rank-2 tensor. If we also are told a handedness convention on space, then we can convert that tensor into an angular velocity vector. However, we still cannot say whether that motion is clockwise/anticlockwise until we further choose a direction along the axis from which to “view” the object. For example, the hands of an ordinary clock turn anticlockwise when viewed from behind the clock.

**CMBR:** See COSMIC MICROWAVE BACKGROUND RADIATION.

**COAXIAL:** Said of two cylindrically-symmetric objects whose axes of symmetry coincide. **COAXIAL CABLE:** One with a cylindrical inner conductor surrounded by a hollow-cylinder insulator, which itself is surrounded by a hollow-cylinder conductor, with all three elements coaxial.

**COION:** Small ion with charge agreeing in sign with that of a macroion or solid object.

**COLATITUDE:** Usually called POLAR ANGLE in this book. Angular distance to a point on a sphere from the north pole. Equals \( \pi/2 \) minus the geographer’s north latitude.

**COLOID OR COLOIDAL SUSPENSION:** Mesoscopic particles distributed in solution. The particles are much bigger than individual molecules, yet small enough that thermal motion prevents their settling under gravity.

**COMMUTE:** Two operations commute if they may be applied in either order with the same net result on any state. For example, all rotations about the \( z \) axis commute with each other, though not with rotations about any other axis. If an algebraic system has the property that all pairs of elements commute, then its combination operation is called commutative. For example, addition of ordinary numbers is commutative, whereas matrix multiplication is not.

**COMPLIANCE:** Reciprocal of spring constant, or tensor whose components form the inverse matrix of those of the spring tensor.

**COMPONENTS OF A TENSOR:** Set of ordinary numbers representing that tensor in a particular coordinate system.

**COMPTON SCATTERING:** Collision between one electron and one photon, leading to one electron and one photon but with altered 4-momenta.
Appendix G Glossary

C: Length scale formed from a particle's mass using the Planck constant and c.

CONDUCTIVITY VS CONDUCTANCE: Both describe an ohmic material. The former is a material property, a tensor giving the linear relation between charge flux and applied electric field (often assumed to be isotropic). The latter depends on the size and shape of a particular object (total current divided by electric potential drop).

CONE CELLS: Photoreceptors in a vertebrate eye responsible for generating color-discrimination signals.

CONSERVATIVE FORCE: One that may be written as minus the gradient of a potential energy function. NONCONSERVATIVE FORCE: One that may not be so written, for example, a frictional or other dissipative force.

CONSTITUTIVE RELATION: A particular class of response function: Yields the \( \vec{D} \) or \( \vec{H} \) field in terms of the \( \vec{E} \) or \( \vec{B} \) field respectively.

CONTINUITY EQUATION: An expression of the local conservation of some quantity, for example, electric charge, energy, or momentum.

CONTRACTION OF A TENSOR: New tensor of lower rank obtained by taking a trace, that is, by setting two indices equal and summing them. OF TWO TENSORS: A contraction of their tensor product.

COSMIC MICROWAVE BACKGROUND RADIATION: Today's remnant of electromagnetic radiation emitted by matter in the early Universe around the time of recombination.

COULOMB GAUGE CHOICE: See GAUGE FIXING.

COUNTERION CLOUD: Counterions held in the neighborhood of a charged macromolecule or surface.

COUNTERION: Small ion with charge sign opposite to that of a macroion or solid object.

COVARIANT: Term not used in this book. (a) Some authors call an equation “covariant” when this book would call it “manifestly invariant,” because both sides “vary” the same way upon Lorentz transformation. (b) Some authors call a tensor of rank \( (1,0) \) a “covariant vector” to distinguish it from \( (0,1) \) (“contravariant”).

COVECTOR: 4-tensor of rank \( (0,1) \) (components have one lower index); quantity transforming the same way as a 4-gradient.

CREATION OPERATOR: Quantum-mechanical operator that increases the number of photons in a state.

CROSS-SECTION, DIFFERENTIAL: Energy per time per solid angle, divided by incoming energy flux. RAYLEIGH: Case of EM radiation on a bound charge. THOMSON: Case of EM radiation on a free charge.

CURL OPERATOR: Cross product of 3D gradient with a vector field. CURL-FREE VECTOR FIELD: One whose curl is zero throughout some region.

CURRENT DENSITY: Not used in this book, but used by some authors to mean what we call “charge flux.”

CURRENT DENSITY: Term not used in this book; see CHARGE FLUX.

CURVATURE, OF A CURVE IN A PLANE OR IN SPACE: A measure of second-order perpendicular deviation of a curve from its tangent at a point (Chapter 7). Equivalently, first-order change of the tangent as we move away from selected point (Chapter 21).

CURVATURE, SURFACE, PRINCIPAL: Eigenvalues of the normalized matrix that describes the second-order perpendicular deviation of a surface from its tangent plane at a point. GAUSS CURVATURE (also called INTRINSIC): Product of two principal curvatures. MEAN CURVATURE (also called EXTRINSIC): Average of the two principal curvatures.

CURVILINEAR COORDINATES: Coordinate system on space for which level sets of each coordinate are not parallel planes.

CUTOFF: A point beyond which some function is negligibly small. For example, a value frequency above or below which a transmission line will not transmit signals.

CYCLIC PERMUTATION: See PERMUTATION.

CYCLOTRON: Apparatus to accelerate charged particles; limited to nonrelativistic speeds. CYCLOTRON MOTION: Circular or helical motion of a charged particle in a uniform magnetic field.

DALEMBERTIAN, OR D’ALEMBERT OPERATOR: See WAVE OPERATOR.

DEBYE SCREENING LENGTH: Length scale over which electrostatic fields decay exponentially in salt solution.

DEBYE: Non-SI unit for electric dipole moment.

DELTA FUNCTION: The linear operator that takes a function \( f \) and returns its value \( f(0) \).

DENDRITE: Structure on a neuron specialized for receiving signals from other neurons.

DEPLETION LAYER: Thin layer of charge non-neutrality, for example, at a p-n semiconductor junction.
DEPOLARIZATION: Change in the normal electric potential jump across a cell membrane, with the interior becoming less negative relative to the exterior than in the “resting” state, typically via channel opening during the passage of a nerve impulse. Not necessarily complete (membrane potential may still be nonzero). Opposite HYPERPOLARIZATION.

DESTRUCTION OPERATOR: Quantum-mechanical operator that reduces the number of photons in a state, or returns zero if no photons are present.

DIAMAGNETIC, DIAMAGNETISM: Refers to a material whose induced magnetic dipole moment density is antiparallel to the applied magnetic field that gave rise to it.

DICHROISM, ORDINARY: Different absorption in media for light of different linear polarizations. CIRCULAR: Same but for light of different circular polarizations.

DIELECTRIC CONSTANT: Permittivity relative to $\varepsilon_0$ (term not used in this book).

DIELECTRIC: Insulating (nonconducting) material.

DIFFERENTIAL FORM: Totally antisymmetric tensor field of some rank $p$ (also $p$-FORM). CLOSED $p$ FORM: One that gives zero when the exterior derivative is applied. EXACT $p$-FORM: One that can itself be expressed as the action of the exterior derivative on some $(p-1)$-form.

DIFFRACTION, DIFFRACTION PATTERN: Resulting pattern of light when an EM wave impinges on an inhomogeneous medium under conditions where geometrical optics (wavelength $\rightarrow 0$) is a poor approximation.

DIFFRACTION: Deviations from geometrical-optics behavior when light encounters objects that are larger than the light’s wavelength, generally describable via the combination of continuously-many waves. (Combination of just two or a few waves is generally called “interference.”)

DIFFUSE CHARGE LAYER: Synonym for counterion cloud.

DIFFUSION EQUATION: Partial differential equation describing the spread of a molecular species in water purely by random, thermal motion.

DILATION: Coordinate transform that rescales all spatial coordinates by the same factor.

DIMENSIONS: Abstract characterization of a class of physical quantities in terms of a symbolic product of powers of $L$, $M$, $T$, and $Q$, distinct from a unit, which is a particular value with specified dimensions.

DIPOLE, CMBR: Systematic dependence of apparent temperature in the cosmic microwave background radiation, featuring antipodal “hottest” and “coldest” points. CURRENT: Distribution of current sources and sinks in a medium. See also DIPOLE, ELECTRIC and DIPOLE, MAGNETIC.

DIPOLE, ELECTRIC: Distribution of charge with nonzero electric dipole moment. MAGNETIC: Distribution of current with nonzero magnetic dipole. PURE OR POINT: Mathematical limit where physical size $\rightarrow 0$ holding fixed dipole moment, also called “point dipole.” RADIATION: EM radiation generated at a leading order of the multipole expansion.

DIPOLE MOMENT, ELECTRIC: First moment of charge. MAGNETIC: First moment of charge flux. MAGNETIC, VECTOR OR TENSOR: two ways of packaging the same three components. INDUCED: Response of a polarizable object to an imposed field, changing its dipole moment.

DISPERSION INTERACTION: Same as LONDON INTERACTION.

DISPERSION: The phenomenon of wave velocity depending on frequency in certain media (absent in vacuum). In optics, it is quantified via the “Abbe number,” $(n_{589\text{ nm}} - 1)/(n_{486\text{ nm}} - n_{656\text{ nm}})$. Also, the spreading-out of signals in a cable or medium brought about by dispersion. DISPERSION RELATION: Relation between $\omega$ and allowed $\tilde{k}$; if it is linear, then the material is not dispersive. OPTICAL ROTATORY DISPERSION: The spectrum of circular birefringence as a function of wavelength.

DISPLACEMENT: Vector describing a change of position (translation) in space or spacetime. ELECTRIC: Formal name for the $\vec{D}$ field.

DISSIPATION, DISSIPATIVE: The property of a process that it irreversibly transforms energy into heat from some other form, such as mechanical or electromagnetic energy.

DIVERGENCE OPERATOR: Contraction of 3D gradient with a vector field. 4-DIVERGENCE: Contraction of 4D gradient with a 4-vector field.

DONOR FLUOROPHORE: One whose emission peak is matched to another (acceptor) fluorophore’s excitation peak.
DOPPLER EFFECT OR SHIFT: GALILEAN: Effect on the frequency of a received wave from the component of the receiver’s velocity along the line of sight. EINSTEIN, LONGITUDINAL: Effect from the component of the receiver’s velocity relative to the transmitter’s velocity along the line of sight. EINSTEIN, TRANSVERSE: Effect from the component of the relative velocity perpendicular to the line of sight.

DOT PRODUCT: Contraction (invariant product) of two 3-vectors.

DOUGHNUT, DIPOLE: Pattern made by the magnitude of energy flux far from an oscillating electric dipole.

DUALITY, ELECTRIC–MAGNETIC: A property of the vacuum Maxwell equations: Their form is unchanged when they are reexpressed in terms of the Hodge dual of the Faraday tensor.

DYAD PRODUCT: Special case of tensor product, for two vectors. The rank-two tensor whose entries are products of one component from each of the two vectors.

DYNAMIC DYNAMICS: versus kinematics: See KINEMATICS.

E-INERTIAL COORDINATE SYSTEM: Coordinate system on spacetime in which mechanics and electrodynamics obey equations of the usual Einstein form. For example, in such coordinates the trajectory of any material object that is not acted on by any force will be a straight line in spacetime.

ECG: See ELECTROCARDIOGRAM.

EDDY CURRENTS: Charge flux with closed streamlines induced in an ohmic material via the Faraday law.

EEG: See ELECTROENCEPHALOGRAM.

EIKONAL (ALSO CALLED EIKONAL FUNCTION): A function on space. It obeys the EIKONAL EQUATION, and its gradient is everywhere parallel to the flux of energy in geometrical-optics approximation. Also an approximate solution to the Maxwell equations involving an eikonal function.

EINSTEIN RELATION: Linear connection between photon energy and frequency. Later combined with de Broglie’s relation to be a linear connection between 4-momentum and 4-wavevector.

“EINSTEIN THINKING”: A style of generating hypotheses driven by symmetry/invariance properties and specifically the Rules for tensor manipulation.

EKG: See ELECTROCARDIOGRAM.

ELECTRET: See FERROELECTRICITY.

ELECTRIC DOUBLE LAYER: Thin layer of charge non-neutrality due to thermal motion when one charge carrier is more mobile than the other, for example, when a surface ionizes in aqueous solution.

ELECTROCARDIOGRAM: Measurement of heart activity by means of electrodes on the skin surface.

ELECTROCHEMICAL potential: Combined effect of electrostatic potential energy and entropic free energy (“chemical potential”) per ion charge for a given ion species in solution. ENERGY: Corresponding quantity per ion. FORCE: Minus its gradient.

ELECTROENCEPHALOGRAM: Measurement of brain activity by means of electrodes on the skin surface.

ELECTROLYSIS OF WATER: Splitting water molecules, ultimately into hydrogen and oxygen, driven by electric current. If the water has a dissolved solute, other atoms may be released, for example chlorine gas from NaCl solution.

ELECTROLYTE: Fluid that conducts electricity via migration of ions, for example, salt solution.

ELECTRON RADIUS, CLASSICAL: Length scale constructed from the charge and mass of the electron.

ELECTROPHORESIS: Net motion of ions in solution induced by an applied electric field.

ELECTROPLATING: Depositing metal onto an electrode under the influence of an electric current.

ELECTROSTATIC potential: Potential energy per charge of a test body in a static electric field. ELECTRIC potential: A quantity whose gradient contributes to the electric field; reduces to the electrostatic field in static situations.

ELECTROTHERM: The passive spread of charge inside a neuron.

ELECTROWEAK INTERACTION: A successor to electrodynamics in which three 4-vector fields (photon, W and Z bosons) jointly mediate electromagnetic and weak nuclear interactions.

ELLIPSE: Plane figure, for example, the points satisfying \((x/a)^2 + (y/b)^2 = 1\). ELLIPSOID: A surface in 3-space obtained by rescaling a sphere along each of three perpendicular axes. For example, a figure of revolution of an ellipse about one of its axes of reflection symmetry, either PROlate (cigar-shaped, stretched) or OBLATE (pancake-shaped, squashed).
ELLiptic Function: Special function arising in many contexts. Ellipse: Three-dots symbol representing something omitted.

EnanTiomers: A pair of molecular structures that are each other’s mirror image. Enantioselectivity: The capability to sort enantiomers or selectively react with just one of them.

EndoScope: Flexible bundle of optical fibers that guide light from every pixel of a visual scene to an eyepiece or camera.

Energy, Kinetic Newtonian: A scalar quantity that would be conserved in an imagined world governed by galilean invariance, but that is only approximately conserved in our relativistic world (and not even approximately in nuclear reactions). Energy, relativistic: A different quantity that is exactly conserved in electrodynamics and other nongravitational physics.

Energy–Momentum Flux 4-Tensor: Object combining energy density, momentum density, energy flux, and momentum flux.

Energy–Momentum Tensor: This book instead uses the phrase energy–momentum flux tensor.

Equipotential: See level set.

ESU: Generic label for an unspecified gaussian unit.

Euler Angles: A parameterization of the rotation group; not used in this book

Evanescent Fields: Fields present in the “forbidden” region outside a dielectric under conditions of “total” internal reflection.

Event: Point in spacetime. A particle interaction, such as radioactive decay, may occur at an event. Or the trajectory of a particle may be regarded as a chain of events at which nothing “happens” other than that the particle could have been observed there.

Evolution: In physics, simply means “change over time,” or by analogy as some time-like variable is changed. Colloquially can also mean “gradual change” as opposed to “revolution.” In biology, shorthand for “evolution by natural selection.”

Excitable Medium: One that is initially spatially uniform, trapped in a metastable state.

Extensive Quantity: Quantity describing a homogeneous system and proportional to its length (resistance, inductance of a solenoid), area (interfacial energy cost), or volume (total mass).

Exterior Derivative of a Differential Form: A \((p+1)\)-form obtained by computing all the partial derivatives of components of a \(p\)-form, then antisymmetrizing.

Extraordinary Wave: See ordinary wave.

Extremum, Extrema: Maximal or minimal value(s). Extremal: having the property of being an extremum.

Faraday Tensor: 4-tensor that includes electric and magnetic fields. Various versions are distinguished by index placement.

Faraday’s Magneto-optical Effect, Faraday’s Rotation Effect: Circular birefringence when light passes through an isotropic, nonchiral medium with a background magnetic field parallel to the light’s direction of propagation.

Farad: SI unit for capacitance.

Farad: SI unit of capacitance

Far Field Limit: In electrostatics or magnetostatics, a situation in which the observer distance is \(\gg\) source size. In electromagnetic radiation, a situation in which the observer distance is much greater than the wavelength, which itself is much greater than the source size. Far-field Approximation: Obtained by dropping terms that are subleading in the respective Taylor expansions.

Ferroelectric Material, Ferroelectricity: Material with a permanent bulk electric dipole moment density in its undisturbed state.

Ferromagnetic, Ferromagnetism: Refers to a material that can retain a permanent magnetic dipole moment density even in zero applied field, for example, steel.

Field Line: A streamline of the \(\vec{E}\) or \(\vec{B}\) field.

Field Point: Point in spacetime where and when an observer measures a field.

Fission, Nuclear: Splitting of a heavy atomic nucleus into lighter fragments.
FLAGELLUM: Corkscrew-shaped appendage projecting from certain bacteria and used for locomotion; pl. flagella.

FLUORESCENCE RESONANCE ENERGY TRANSFER: An electromagnetic mechanism that couples excitation on one atom to another without any actual electromagnetic radiation.

FLUORESCENCE: Process in which a molecule absorbs light, then after a delay emits light with a longer wavelength. The direction and polarization of the emitted light may differ from those of the incoming light.

FLUOROPHORE: Molecule, or group within a molecule, that when excited has a high probability of reemitting light.

FLUX, 1D: Rate of flow of some quantity per time. 2D: Per time per transverse length. 3D: Per time per transverse area.

4-VECTOR: Combined object containing both density and flux. ELECTRIC OR MAGNETIC: Terms not used in this book.

ELECTROPHORETIC: flux of ions in solution driven by an electric field. See also CHARGE FLUX, MOMENTUM FLUX.

FORM, DIFFERENTIAL: See DIFFERENTIAL FORM.

Förster radius: Parameter describing the coupling of two molecules via FRET.

FREQUENCY, PLASMA: See PLASMA.

FREQUENCY: ANGULAR: Radians per unit time, often denoted \( \phi \). CIRCULAR: Revolutions or complete periods per unit time, often denoted \( \omega \).

FRET: See FLUORESCENCE RESONANCE ENERGY TRANSFER. FRET efficiency: Probability that a donor fluorophore will actually transfer its excitation to a donor.

G-INERTIAL COORDINATE SYSTEM: Coordinate system on spacetime in which mechanics and gravitation obey equations of the usual newtonian form. (We now know that no such coordinates exist, because newtonian physics is only approximately correct, but we sometimes discuss an imagined world in which it does hold.)

GALILEAN GROUP: Consists of rotations, translations, galilean boosts, and reflections in space and time. See also GROUP.

GALILEAN RELATIVITY: Implements the Principle of Relativity by imposing invariance under the galilean group.

GAUGE FIXING OR GAUGE CHOICE: Procedure of imposing an auxiliary condition to specify a unique representative vector potential for a given magnetic field. GAUGE TRANSFORMATION: Addition of a gradient to the 3-vector potential or 4-vector potential. GAUGE INVARIANCE: The property that fields, or the form of the Lagrangian function, do not change upon gauge transformation.

GENERATOR: OF ROTATION, \( 3 \times 3 \) antisymmetric matrix, interpreted as the first-order difference between an infinitesimal rotation matrix and the identity. Similarly for GENERATOR OF LORENTZ TRANSFORM.

GEOMETRICAL OPTICS: An approximation scheme in which we consider light in dielectrics that either are piecewise homogeneous with sharp boundaries between regions with sizes much larger than the wavelength of light, or else that have gradually varying refractive index compared to the wavelength of light.

GRADIENT INDEX MEDIUM: One with gradually varying refractive index.

GRAND UNIFIED THEORY (GUT): One of a class of field theories hypothesized to describe Nature, unifying strong with electroweak interactions.

GREEN FUNCTION: The “inverse” of a differential operator, for example, the laplacian (for newtonian gravitation or magnetostatics in Coulomb gauge), or the wave operator on 4-vector fields (for electrodynamics in Lorenz gauge).

CAUSAL OR RETARDED: One with the property that effects must arise later in time than their causes.

GREENHOUSE GAS: Atmospheric constituent molecule having strong interaction with infrared radiation.

GRIN: See GRADIENT INDEX MEDIUM.

GROUP VELOCITY OF A PLANE WAVE: Derivative of angular frequency with respect to wavenumber.

GROUP: Mathematical system consisting of some elements (for example, coordinate transformations, matrices or permutations), a composition rule (for example, combination of boosts or matrix multiplication), and an inverse operation under that composition. See also LORENTZ GROUP, GALILEAN GROUP, and specific group name abbreviations.

GUT: See grand unified theory.

HANDEDNESS: Four senses: (a) The many coordinate for space fall into two categories. “Handedness” of a coordinate system refers to which of those two categories a particular system belongs to. (b) Selecting one class of coordinate systems and declaring it to be right handed is said to confer “handedness” on space itself. (c) Synonym for CHIRALITY, that is, the property of an object being superimposable on its mirror image. By extension, the analogous property for...
a plane wave of EM radiation. (d) A choice of which version of a handed object is under discussion, for example, the “handedness” of a particular helix.

HEATMAP: Representation of a function of two variables in which the function’s value is encoded as color.

HELCITY: Choice of whether circularly polarized light is right- or left-handed. HELCITY BASIS: A complex basis for the 2D space of vectors transverse to the wave propagation direction.

HENRY: SI unit for inductance.

HERMITIAN CONJUGATE: Operation that returns the complex conjugate of the transpose of a matrix. HERMITIAN MATRIX: One that is equal to its hermitian conjugate.

HOMOGENEITY: Property that some or all properties are constant throughout a region of space (“homogeneous”) (possibly after some kind of local averaging).

HOMONUCLEAR MOLECULE: One consisting of two identical atoms.

HYDROLYSIS: Splitting of a molecule by breaking a covalent bond and attaching fragments of a water molecule to the two atoms previously bonded to each other. In particular, splitting of ATP which releases stored energy.

HYDROSTATIC EQUILIBRIUM: Situation where the forces on one or more fluids balance, leading to zero bulk motion.

HYPERBOLA: A plane curve, for example, graph of the solutions of $xy = 1$. HYPERBOLIC TRIG FUNCTION: Class of functions related to sin, cos, … and denoted sinh, cosh, …. HYPERBOLOID: A surface of revolution in 3D obtained by rotating a hyperbola about its symmetry axis.

HYPERCUBE: The analog of a cube in $D > 3$ dimensions; region where all of a set of cartesian coordinates lie with the same range.

HYPERPOLARIZATION: Condition in a living cell where the interior electric potential is even more negative relative to the exterior than in the resting state.

IDENTITY: A formula with very general validity, usually expressing a mathematical relation not contingent on physical measurement or theory. IDENTITY MATRIX: In $D$ dimensions, the $D \times D$ array of numbers, all equal to 1 on the diagonal or 0 off the diagonal. IDENTITY TENSOR: In 3D, has components $\delta_{ij}$ that are equal to the corresponding elements of the identity matrix in any coordinate system. In 4D, has components $\delta^*_{ij}$ or $\delta^*_i$, with a similar property. In any dimensions, this tensor may also be regarded as the linear machine that eats any vector and returns the same vector.

INACTIVATION: Behavior of sodium ion channels that upon sustained depolarization, they eventually re-close, helping to terminate an action potential.

INDEX OF REFRACTION: See REFRACTIVE INDEX.

INDEX RAISING AND LOWERING: Standard ways to change the rank ($\frac{d}{d}$) of a tensor by interconverting upper and lower indices of its components.

INDEX: DOTTED, UNDOTTED: Associated with the two inequivalent 4D spinor types. DUMMY: One of a pair of tensor indexes that are contracted; may be renamed at will. INDEX RAISING AND LOWERING: Contraction of an index with the metric tensor, for example, to interconvert 4-vectors and 4-covectors. LOOSE: Tensor index that is not summed.

LOWER, UPPER: Associated with the two equivalent 4D vector types.

INDUCTANCE, SELF: Expresses the linear relation between the time rate of change of current through a coil and the resulting work per charge required to push charges through it. Mutual inductance: Similar but for the work per charge in one coil versus current in a different coil. INDUCTOR: Circuit element designed to contribute self-inductance but negligible resistance.

INDUCTION, MAGNETIC: Appearance of an electric field in a conductor associated with a time-dependent applied magnetic field. See also inductance.

INDUCTION, MAGNETIC: Formal name for the $\tilde{B}$ field.

INDUCTOR: Circuit component, typically a helical coil, designed to have substantial self-inductance but little resistance.

In circuit diagrams, an idealized form with zero resistance.

INERTIAL COORDINATE SYSTEM: In electrodynamics, abbreviation for E-INERTIAL. In newtonian physics, abbreviation for G-INERTIAL.
INHOMOGENEOUS: Not homogeneous.

INTEGRABILITY LEMMA: Preparatory result that given some conditions, some quantity can be expressed as a particular kind of derivative of another quantity. We develop such results for the electrostatic potential, the magnetostatic vector potential, and the full electromagnetic 4-vector potential.

INTEGRAL CURVE: Same as STREAMLINE.

INTENSITY, ELECTRIC: Formal name for the $\vec{E}$ field.

INTERFERING-RAY OPTICS: x

INVARINACE: (a) A "passive" transformation of variables (does not change any physical situation) that leaves the form of the dynamical equations unchanged. (b) More broadly, a relation such as an equation of motion, or an element such as a wave operator may be "invariant" such a transformation, for example, "Lorentz invariant." (The synonym "form invariant" just emphasizes that what is unchanging is the mathematical form.) (c) More broadly still, a property of an object (such as the principal curvatures of a surface) may be invariant under some group of transformations. Such a quantity is often instead called 3-SCALAR if the group is rotations or 4-SCALARS if it is Lorentz transformations (or simply "constant"). (d) A mathematical operation, such as extracting the symmetric part of a tensor, is "invariantly defined" if it can be specified without any choice of coordinate, or if different coordinate choices in some class (for example, cartesian in 3D or E-inertial in 4D) lead to the same result. (e) See GAUGE INVARIANCE.

INVARINANT INNER PRODUCT: Double contraction of two four-vectors with the metric; a scalar.

INVERSION, SPATIAL: One of several kinds of linear transformations with determinant $-1$. POINT INVERSION: Matrix is $-\vec{I}$. INVERSION THROUGH A PLANE: For example, the $yz$ plane: $x \rightarrow -x, y \rightarrow y, z \rightarrow z$.

IONOSPHERE: Atmospheric layer that is partially ionized.

ION: Atom or molecule that has net charge because one or more electrons has been added or removed. IONIZE: Form an ion from a neutral atom or molecule. ION CHANNEL: A protein complex embedded in a bilayer membrane, typically containing a water-filled passage that controllably allows the passage of ions.

IRRADIATE: Expose to radiation, for example EM radiation.

ISOSURFACE: See LEVEL SET.

ISOTROPY: Property that some or all behaviors are rotationally invariant ("isotropic") at some point of space (possibly after some kind of local averaging). In vacuum or another isotropic medium, any vector expressing a preferred direction must equal zero; any rank-2 symmetric tensor must be a constant times the identity; and tensors of higher even rank must be built from tensor products of identity tensors.

JACOBIAN: A factor needed when changing variables of integration or in the coordinate transformation of a delta function: absolute value of the determinant of the matrix of partial derivatives of one coordinate set w.r.t. the others.

JONES VECTOR: Synonym for the complex vector amplitude (polarization vector) of a pure (hence polarized) plane EM wave. Specifies the amplitude and polarization of the wave. JONES MATRIX: Specifies the transformation of the Jones vector produced by an optical advice, for example, absorption of various polarization states, rotation of polarization, and so on.

JOULE HEATING: Same as ohmic heating.

KEESEMER INTERACTION: Interaction between two dipoles arising when their orientations are statistically correlated.

KINEMATIC, KINEMATICS: versus dynamics: Having to do with clear description of a system and its symmetries but not with the detailed equations of motion governing its time development.

KRONECKER SYMBOL: See IDENTITITY MATRIX.

LAGRANGIAN DENSITY: A function of positions and velocities of particles, and/or of fields and their derivatives, that characterizes a system by determining its equations of motion. Used when defining an action functional.

LAPLACIAN OR LAPLACE OPERATOR: Dot product of the 3D gradient with itself; a second-order differential operator.

LATITUDE: $\pi/2 - \vartheta$ as a coordinate for Earth’s surface. See also PARALLELS.

LENSING, GRAVITATIONAL: A compact, heavy object’s ability to bend rays of light passing close to it.

LENZ’S LAW: The direction of an induced electric field opposes the time rate of change of the magnetic field that is the electric field’s source.
LEPTON: Class of fundamental particles including electron, muon, tau, and their antiparticles. LEPTON NUMBER: A conserved quantity in the Standard Model equal to net number of all leptons minus antileptons.

LEVEL SET: Set of all points where a function has one fixed value. For example, on a topographic map the contour lines are level sets of altitude. Equipotentials are level sets of electric potential. The special case of a 2D level set in 3D space is sometimes called an “isosurface.”

LEVI-CIVITA SYMBOL: In 3D, a $3 \times 3 \times 3$ array of numbers, each $\pm 1$ or 0. In 4D, a $4 \times 4 \times 4 \times 4$ array of numbers, each $\pm 1$ or 0. LEVI-CIVITA TENSOR: In 3D, an antisymmetric tensor of rank 3 whose components in any right-handed cartesian coordinate system equal the corresponding elements of the Levi-Civita symbol, and similarly in 4D.

LIGHT CONE OF A SPECIFIED POINT: Locus of all points lightlike separated from that one. Similar to the surface of a cone in ordinary solid geometry, but of one higher dimension. LIGHT CONE COORDINATES, in one space and one time dimension, coordinates for which the two branches of the origin's light cone are level sets.

LIGHTLIKE FOUR VECTORS: One whose invariant inner product with itself is zero.

LONDON FORCE: Interaction between two dipoles arising when their orientations are quantum-mechanically anticorrelated.

LONGITUDE: See MERIDIAN.

LORENTZ FORCE LAW: Equation of motion for a point charge in the presence of electromagnetic fields. Time rate of change of the particle's momentum.

LORENTZ TRANSFORMATION, LORENTZ GROUP: Consists of rotations, reflections in space and time, Lorentz boosts, and all combinations of these. PROPER: The subgroup whose matrix has determinant $+1$. ORTOCHRONOUS: The subgroup that do not reverse the direction of time. RESTRICTED: The transformations that are both proper and orthochronous.

LORENZ GAUGE CHOICE: See GAUGE FIXING.

LOWERING OPERATOR: See DESTRUCTION OPERATOR.

MACROMOLECULE: Macromolecule, such as a protein or DNA, that has net electric charge in solution.

MAGNETIC FIELD INTENSITY: Formal name for $\mathbf{H}$.

MAGNETOENCEPHALOGRAPHY: Measurement of brain activity by means of magnetic field sensors on the skin surface.

MAGNETOSTATICS: Study of magnetic fields set up by moving charges when their charge flux is stationary (time-translation invariant).

MANIFEST INVARIANCE, MANIFESTLY INVARIANT: Written using notation that makes an invariance property clear at a glance.

MASS DEFECT: Difference between the total mass of incoming colliding particles and that of the outgoing products of the collision.

MATRIX, ORTHOGONAL: Square, real matrix with the property $\mathbf{M} \cdot \mathbf{M}^t = \mathbf{1}$. ROTATION: Orthogonal matrix with determinant $+1$. UNITARY: Square, complex matrix with the property $\mathbf{M} \cdot \mathbf{M}^\dagger = \mathbf{1}$.

MEAN-FIELD APPROXIMATION: Neglects statistical fluctuations, for example, by replacing local charge density by its expectation over an ensemble.

MEG: See MAGNETOENCEPHALOGRAPHY.

MEMBRANE POTENTIAL: The jump in electric potential across a cell membrane.

MERIDIAN OF LONGITUDE: Semicircle of points on Earth's surface with the same value of azimuthal angle.

METAMATERIAL: [[Not ready]]

METASTABLE: Not absolutely stable; trapped in a local minimum of energy. More generally, able to be dislodged from a steady state by outside perturbation or random thermal motion.

METRIC TENSOR: 2D, 3D: Tensor of rank two that gives the quadratic relation between spatial displacement and distance. Of FLAT 4D SPACE: Tensor of rank two that gives the quadratic relation between spacetime displacement and invariant interval.

MICELLES: Loose association of molecules into aggregates of fairly uniform, submicrometer size.

MICHELSON–MORLEY EXPERIMENTS: We discuss two: (a) unobservability of “aether wind”; and (b) dragging of light by flowing water.
MICROWAVE RADIATION: EM radiation with frequency in the range 300 MHz — —300 GHz.
MOLAR: Unit of concentration (number density), equal to the numerical part of Avogadro’s number divided by one liter.
MOLE: In this book, considered as a very large integer $\approx 6.02 \cdot 10^{23}$.
MOMENT OF INERTIA TENSOR: The second moment of a rigid body’s mass, with its trace modified but not eliminated. It describes the linear relation between the body’s angular momentum and its angular velocity.
MOMENT: Tensor obtained by integrating over a body some number of factors of position vector components, weighted by mass density (for gravitational multipole moments or moment of inertia), charge density (for electric moments), or charge flux (for magnetic moments). Modified forms include MULTIPOLE MOMENTS and MOMENT OF INERTIA.
MOMENTUM, NEWTONIAN: A vector quantity that would be conserved in an imagined world governed by galilean in-variance, but that is only approximately conserved in our relativistic world. MOMENTUM, RELATIVISTIC: A different quantity that is exactly conserved in electrodynamics and other nongravitational physics. 4-MOMENTUM: The 4-vector that combines relativistic energy and momentum.
MONOCHROMATIC LIGHT: A single plane wave, hence with definite frequency, wavevector, and polarization vector. Not a superposition of many such waves. But some authors extend to mean a superposition with a narrow range of frequencies and no restriction on polarization.
MONOVALENT ION: Ion with charge $\pm e$.
MÖSSBAUER EFFECT: Nuclear transition leading to emission of a photon with no recoil of the emitting nucleus.
MULTILINEAR: See BILINEAR.
MULTIPOLE EXPANSION/APPROXIMATION: Procedure to organize the calculation of fields far from a localized source in a power series.
MULTIPOLE MOMENTS: Modified moments used in a far-field expansion of static or radiation fields. Various cases (mono-, di-, quadru-, octupole, …) are named for the quantity $2^p$ where $p$ is the order of the moment. For example, the quadrupole moment is the second moment of mass or charge with its trace removed.
MULTIPOLE POTENTIALS: A small number of standard electric potential functions, to be combined weighted by multipole moments to obtain the far electric potential function of an arbitrary localized charge distribution. (Similarly for magnetostatics.) Various cases (mono-, di-, quadru-, octupole, …) are named for the quantity $2^p$, where $p$ is the order of the corresponding moment.
MUON: Fundamental particle analogous to the electron but heavier and unstable.
NERNST RELATION, NERNST POTENTIAL: Equilibrium relation giving the concentration jump needed to cancel electrically driven ion migration. NERNST–PLANCK FORMULA: Nonequilibrium relation giving the net flux of an ion species under the combined effect of a concentration gradient and an electric field.
NERVE: See NEURON. NERVE IMPULSE: See ACTION POTENTIAL.
NEURON: Living cell in an animal specializing in electrochemical signaling.
NEUTRINO: Any of three distinct fundamental particles (with corresponding antiparticles). All are electrically neutral leptons, associated respectively with electrons, muons, or tau leptons.
NEWTONIAN POTENTIAL: Gravitational potential energy per mass of a test body.
NONCHIRAL: Not chiral; describes an object that can be made to coincide with its mirror image by a translation and/or rotation.
NONDIMENSIONALIZE: Replace a quantity by its quotient by a standard quantity with the same dimensions.
NORMAL VECTOR: A vector that is perpendicular to something, for example, a surface.
NORMALIZATION OF A 3-VECTOR OR 4-VECTOR: Procedure of computing the scalar product of the vector with the inverse of its norm.
NORTH POLE OF MAGNET: A permanent magnet experiences a torque tending to align its dipole moment with Earth’s magnetic field. If the magnet is free to pivot in a horizontal plane, then the point on its surface with maximal field strength and with $\vec{B}$ pointing outward will align roughly toward the Arctic, and so is called the magnet’s “north pole.”
Because opposite poles attract, when we regard Earth as itself a magnet then its own south magnetic pole is therefore located in the Arctic, and its north magnetic pole in the Antarctic.

NSOM: Near-field scanning optical microscopy.

NUCLEON: Class of composite subnuclear particles including proton and neutron and their antiparticles. NUCLEON NUMBER: A conserved quantity in the Standard Model equal to net number of all nucleons minus antinucleons.

NULL 4-VECTORS: Same as LIGHTLIKE 4-VECTORS.

O(3), O(3,1), O(3,1), O(4): Groups of real orthogonal or Lorentz matrices.

OCCUPATION NUMBERS: Integers characterizing one basis of the quantum-mechanical state space for the electromagnetic field.

OCTAHEDRON: Regular polyhedron with eight triangular faces.

OCTUPOLE: See MULTIPole MOMENTS and POTENTIALS.

OHMIC MATERIAL: One that conducts electricity (with free charges that can migrate anywhere) with a linear relation between applied $\vec{E}$ field and charge flux. OHMIC HEATING: Process in such a material where electric potential energy is degraded to heat. OHMIC HYPOTHESIS: The claim that the bilayer membrane surrounding a cell allows ions to pass through via an ohmic relation, where the role of electric potential is played by electrochemical drive.

OPTICAL ROTATORY POWER OR OPTICAL ACTIVITY: See CIRCULAR BIREFRINGENCE.

ORDINARY WAVE: In a biaxially anisotropic medium, one possible polarization of EM plane waves that lies in the plane perpendicular to the special axis. EXTRAORDINARY WAVE: The other polarization, perpendicular to ordinary.

ORIENTATION: The text uses this term in two senses. (a) The everyday sense is “which way an object points in space.” This continuous variable changes when we rotate the object. (b) But the mathematical sense is a binary choice of which direction we traverse a curve, or which coordinate systems on 3-space are deemed to be right-handed, or the analogous choice on 4D space. Orientation in any of these senses is discrete and not affected by rotations. For example, no matter how you rotate a hand, you won’t change it into the other hand.

ORTHOCHRONOUS: See LORENTZ TRANSFORMATION.

ORTONORMAL: Said of a set of vectors that are all mutually perpendicular, and also normalized, in some metric.

OUTER PRODUCT: Same as dyad product.

OVERDETERMINED: Said of variables that are required to satisfy constraints with no solution.

PARALLAX: Optical phenomenon in which the apparent relative positions of objects shift according to the observer’s position.

PARALLELEPIPED: 3D analog of a parallelogram: A polyhedron with three pairs of parallel faces, not necessarily meeting at right angles.

PARALLELS OF LATITUDE: Circular curves of fixed polar angle on Earth.

PARAMAGNETIC, PARAMAGNETISM: Refers to a material whose induced magnetic dipole moment density is parallel to the applied magnetic field that gave rise to it.

PARAMETERIZE: Assign a unique numerical label to each point on a curve. PARAMETERIZATION: such an assignment.

PARAMETERIZED CURVE: A curve together with a particular choice of parameterization. PARAMETRIC SPECIFICATION OF A CURVE: A set of $D$ functions whose graph in $D$-dimensional space is the curve in question. We also say that the curve has been specified parametrically.

PARITY OF A PERMUTATION: The identity permutation (do nothing) has “even (or +1) parity.” The permutations that exchange two adjacent positions all have “odd (or –1) parity.” Upon composition of two permutations, parity multiplies. All other permutations can be obtained by combinations of exchanges.

PARSIMONIOUS EXPLANATION: One that uses the fewest possible assumptions.

PASSIVE-SPREAD SOLUTION: Spacetime course of an electric disturbance assuming ohmic membrane conductance (no voltage gating).

PASSIVE TRANSFORMATION: See INVARIANCE.

PATCH-CLAMP: Experimental method to measure membrane potential difference in a living neuron.

PATHOLOGICAL: Contrived; not arising in practice; containing a mathematical singularity. PATHOLOGY: mathematical
complication that rarely if ever arises in practice, even though it may arise in a pathological imagined situation.

PERMEABILITY, MAGNETIC, OF VACUUM: Constant of Nature that describes what $\mathbf{B}$ field is generated by a given current.

OF LINEAR MAGNETICALLY POLARIZABLE MEDIUM: Material property that describes what $\mathbf{H}$ field arises in a given $\mathbf{B}$ field; in general a tensor.

PERMEABLE: Said of a membrane or ion channel that allows passage of some specified atom, molecule, or ion. PERMEANT species: Class of atom, molecule, or ion that has significant ability to cross a membrane, or a membrane channel under discussion. Opposites IMPERMEABLE, IMPERMEANT.

PERMITTIVITY, ELECTRIC, OF VACUUM: Constant of Nature that describes what $\mathbf{E}$ field is generated by a given charge.

OF LINEAR DIELECTRIC MATERIAL: Material property that describes what $\mathbf{D}$ field arises in a given $\mathbf{E}$ field; in general a tensor.

PERMUTATION: Rule specifying a rearrangement of two or more distinct objects, typically in a linear list. For example, two objects (penny and nickel) have just two permutations: Do nothing, or exchange the objects. A permutation may be something like $\sigma = \text{"swap the second and third items in the list, leaving others in place"}$; it may be applied to objects that initially are in any order. Hence the net result of applying this $\sigma$ twice is to return all elements to their original order, whatever that may have been. CYCLIC PERMUTATION: a special subclass of permutations that push each object in a linear list forward, “cycling” those that “fall off” the end back to the front. If $\sigma'$ is the cyclic permutation that shifts a list of $N$ objects by one place, then the result of applying it $N$ times is to return all elements to their original order, whatever that may have been. See also PARITY.

PHASE VELOCITY OF A PLANE WAVE: Quotient of angular frequency divided by wavenumber.

PHENOMENOLOGICAL: Empirical or partly so; not derived from elementary interactions via a detailed model.

PHOTORECEPTOR CELL: Living cell in an animal eye that transduces incoming light to neural signals.

PIEZOELECTRIC MATERIAL, PIEZOELECTRICITY: Material whose bulk electric dipole moment density changes upon mechanical deformation, and which conversely deforms upon application of an electric field.

PI: Class of composite (not fundamental) subnuclear particle with overall spin zero.

PLASMA: Fully or partially ionized gas. COLD: One in which thermal motions may be ignored. PLASMA FREQUENCY:

Inverse time scale constructed from the charge, mass, and concentration of the most mobile species.

PLUS ULTRA: “More beyond.”

POINCARÉ GROUP: Consists of rotations, translations, Lorentz boosts, and reflections in space and time.

POINCARÉ LEMMA: Mathematical result outlined in Section 15.3.5, stating that an antisymmetric tensor whose antisymmetrized partial derivatives vanish may itself be expressed as the antisymmetrized partial derivatives of a “potential” tensor.

POINCARÉ SPHERE: A geometrical construction for cataloging the possible polarization and intensity states of a beam of light (Section 24.2.2, page 385).

POINT CHARGE: Imagined limiting case of an object with charge and mass but whose higher multipole moments, if any, may be neglected for a particular application.

POLAR ANGLE: Coordinate describing the tilt away from an axis (usually for an axisymmetric body), usually called $\hat{\phi}$; analogous to colatitude in geography.

POLAR MOLECULE: One with a strong permanent electric dipole moment, for example H$_2$O. Contrast NONPOLAR MOLECULE: For example, CH$_4$.

POLARIZABILITY TENSOR OF ATOM OR MOLECULE: Tensor (often assumed isotropic) giving the average dipole moment developed in response to imposed electric field. OF LINEAR, ISOTROPIC, DIELECTRIC MEDIUM, BULK: Dipole moment density developed in response to imposed electric field, divided by field strength, or tensor describing their linear relation.

POLARIZABLE MOLECULE OR MEDIUM: One that is capable of generating an induced dipole moment or dipole moment density, respectively. POLARIZABILITY: That ability.

POLARIZATION, OF CELL MEMBRANE: The existence of a nonzero membrane potential, specifically a jump of potential across the membrane with the interior more negative than the exterior. See also DEPOLARIZATION, HYPERPOLARIZA-
POLARIZATION VECTOR: Complex amplitude of 3-vector potential in Coulomb gauge, or of 4-vector potential in any gauge.

POLARIZATION: Used in two distinct senses: (a) See polarized light. (b) See polarizability of molecule or medium, polarization of cell membrane.

POLARIZED LIGHT: LINEAR, CIRCULAR, ELLIPTICAL, PARTIAL: Represented by a plane or nearly-plane wave solution of Maxwell with polarization vector that is respectively (a) an overall constant times a real vector; (b) an overall constant times \( \frac{1}{\sqrt{2}} \); (c) something in between the extremes (a) and (b).

POLARIZER OF POLARIZING FILTER: Optical element that preferentially transmits one polarization state of light, partially extinguishing the other one. By extension, similar element acting on waves from other parts of the EM spectrum, for example, microwaves. POLAROID: Brand name for a thin plastic film used as a polarizer.

POR: See Principle of Relativity.

POSITRON: Fundamental particle with charge +e. Antiparticle of the electron. POSITRONIUM: Bound state of a positron and an electron.

POTENTIAL: See action potential, electrostatic potential, membrane potential, multipole potentials, Nernst relation, newtonian potential. 3-VECTOR POTENTIAL: Vector field whose curl equals the magnetic field. 4-VECTOR POTENTIAL: Combined object including electric potential and 3-vector potential; its antisymmetrized derivatives yield the Faraday tensor.

POYNTING VECTOR: Flux of electromagnetic energy. POYNTING THEOREM: A continuity equation for energy.

PRESSURE: Free energy per volume, or equivalently force per area, to reduce a region in space, for example, by compressing its contents.

PRINCIPLE OF RELATIVITY: States that no experiment done within an isolated system can determine whether or how fast that system is moving. The generator of hypotheses outlined in Idea 26.1 (page 405), both in newtonian and in Einstein physics.

PROBABILITY DENSITY: Describes the distribution of a continuous quantity via the probability for it to be found in a small range divided by the size of that range.

PROPAGATOR: See Green function.

PROPER TIME: Time as measured by a clock comoving with some object in question.

PULSAR: Rapidly rotating neutron star that emits radiation in a “searchlight” pattern, so that distant observers see short, intense, periodic pulses.

QUADROPOLE, ELECTRIC: Distribution of charge with nonzero electric quadrupole moment. MAGNETIC: distribution of current with nonzero magnetic quadrupole moment. RADIATION: EM radiation generated at a particular subleading order of the multipole expansion.

QUARK: Any of six classes of strongly-interacting fundamental particle, or their antiparticles. All quarks carry electric charge.

QUASI-STATIC: Regime in which charges may move, but do so slowly enough that the associated magnetic fields may be neglected.

QUASISTEADY: State variables, such as concentrations, are constant in time, but not necessarily due to thermodynamic equilibrium; fluxes in an out are matched, but not necessarily zero.

RACEMIC: Said of a solution containing equal amounts of each enantiomer of a chiral molecule.

RADIATION: Electromagnetic field created by a localized source and capable of carrying energy out to infinity (radiative energy transfer).

RAISING OPERATOR: See CREATION OPERATOR.

RANK OF A TENSOR: In 3D, an integer giving the number of indices needed to specify a component. In 4D, a pair of integers giving the number of up and of down position indices. OF A SPINOR: In 3D, the number of spin indices divided by two. In 4D, a pair of half-integers giving the number of undotted and dotted indices, each divided by two.

RAPIDITY PARAMETER: \( \tanh^{-1}(v/c) \) for a Lorentz boost with velocity \( v \).

RAY OPTICS: See geometrical optics. RAY OF LIGHT: A streamline of the energy flux in a geometrical-optics solution.
These all are solutions to the **Ray Equation**, itself a consequence of the eikonal equation.

**Rayleigh range**: Length scale characterizing the spreading of a beam.

**Ray of light**: Streamline of the energy flux in geometrical optics approximation. **Ray equation**, equation whose solutions are rays.

**Reactance**: Complex quantity describing a circuit with linear elements, with real contributions from resistance and imaginary contributions from capacitance and inductance.

**Redshift**: (a) Doppler effect when a source moves away from the observer. (b) Gravitational effect when light propagates from source to an observer at a less negative gravitational potential.

**Refraction**: Bending of a plane wave of light's wavevector as it cross an interface between two homogeneous media (one of which may instead be vacuum). **Double**: Different bending for two linear polarizations of incident light.

**Refractive index**: Phase velocity of light in a medium divided by velocity of light in vacuum.

**Relativistic effect**: One that is numerically small, though not necessarily negligible, if particle velocities are all \( \ll c \).

For example, the magnetic forces between from individual slowly-moving electrons are small, but may nevertheless be significant when many electrons combine. **Relativistic system**: One in which some particle speeds are not \( \ll c \).


**Renormalization of charge**: Phenomenon in which the fields created by a strong charge in solution fall exponentially, but with an overall prefactor different from the naïve expectation based on the weak-charge limit.

**Representation of a group, linear**: Associates a matrix to each group element, for example, the tensor product of two orthogonal matrices to their corresponding rotation, in a way that implements the group operations as matrix multiplication and inversion. **Fundamental**: A representation that, unlike the example just given, is as low dimension as possible.

**Resistivity vs resistance**: The former is the reciprocal of conductivity. The latter is the reciprocal of conductance.

**Response function**: Includes (a) The dielectric polarization and/or magnetization as functions of applied \( \vec{E} \) and/or \( \vec{B} \); (b) constitutive relations \( \vec{D} \) and \( \vec{H} \) in terms of \( \vec{E} \) and \( \vec{B} \); (c) charge flux in terms of driving electric field; and so on. May be frequency-dependent. A linear response function can be defined by a tensor, for example, susceptibility, permittivity or permeability, conductivity, and so on.

**Rest frame**: An inertial coordinate system in which a particular object is at rest, that is, its trajectory is the straight line \((ct, \vec{r})\) for variable \( t \) and fixed \( \vec{r} \).

**Retarded time**: For a point charge in motion, the time prior to observation at which the observer's past light cone intersects the particle trajectory, denoted \( t_r \). (In multipole approximation, the time prior to observation at which the observer's past light cone intersects the center of the radiating charge distribution is a different quantity, denoted \( t_r \).

**Rod cells**: Photoreceptors in a vertebrate eye responsible for monochromatic dim-light vision.

**Rotation matrix**: A \( 3 \times 3 \) orthogonal matrix with determinant \( +1 \). Or a \( 4 \times 4 \) Lorentz transform matrix with a 3D rotation matrix in its space-space block, 1 in its time-time block, and zero elsewhere.

**Scalar product, 3D**: Same as dot product. **Scalar product, 4D**: Contraction of two 4-vectors.

**Semiclassical approximation**: An approach to diffraction problems that is valid just beyond the geometrical-optics regime.

**Semiconductor**: Material with a small but nonzero concentration of mobile charge carriers, due either to thermal excitation or the intentional presence of impurity atoms.

**Semiinfinite**: Infinite in one direction, for example, extending from \( x = 0 \) out to \( +\infty \).

**Semimajor, semiminor axes**: Distances associated to an ellipse: one half of the largest, respectively shortest, diameter.

**Separation of variables**: Method of finding solutions to a partial differential equation by choosing coordinates in which product functions can be found that solve the equation, for example, \( A(\rho)B(\phi)C(z) \) in cylindrical coordinates.

**Shear stress**: See stress, shear.

**Shortwave radio**: EM radiation with frequency in the range 3–300 MHz.

**Sidereal year**: Time for Earth's center to complete an orbit around the Sun. A few minutes different from the time between successive fall equinoxes or other solar measures.
SIEMENS: SI unit for conductance (reciprocal of ohm).

SIMULTANEOUS EVENTS: In a particular coordinate system, two events that share the same value of the chosen time coordinate. RELATIVITY OF SIMULTANEITY: The observation that events that are simultaneous in one E-inertial coordinate system will generally not be so in another one.

SL(2,C): Group of complex matrices with determinant +1.

SMALL SOURCE LIMIT: In electromagnetic radiation, a situation less restrictive than far-field limit. SMALL SOURCE APPROXIMATION: Obtained by dropping terms that are subleading in the small-source limit.

SNEL LAW: Traditional name for the law of refraction. After precursor work by Ibn Sahl in 984 and Alhazen in 1021, the modern form was discovered by Thomas Harriot in 1602, rediscovered by Willebrord Snellius in 1621, but first published by René Descartes in 1637.


SOLENOID: Tight helically-wound coil of wire.

SOLITON, SOLITARY WAVE: Traveling nonlinear wave. Some mathematical authors reserve this word for the special case of an “exactly integrable” system.

SOURCE POINT: Point in spacetime whose contribution to a field is to be added to the others to obtain the full result.

SPACE CONSTANT OF CABLE OR AXON: Length scale determined by size of the axon, membrane conductance, and the interior and exterior solutions.

SPACELIKE 4-VECTORS: One whose invariant inner product with itself is positive.

SPECTRUM, EMISSION: Probability density function for light of various wavelengths to be emitted by a particular excited fluorophore. EXCITATION: Probability per time per incoming energy flux of inducing excitation in a fluorophore, as a function of incoming wavelength.

SPIN(3), SPIN(3,1): Covering groups of SO(3) or SO^{+}(3,1) respectively; same as SU(2) or SL(2,C) respectively.

SPINOR: Physical quantity transforming in a representation of Spin(3) or Spin(3,1) with half-integer spin rank. WEYL SPINOR: the case of spin rank (1/2,0) or (0,1/2). DIRAC SPINOR: both kinds of Weyl spinor taken together as representing a single physical object, for example, the field associated to electrons.

SPRING TENSOR: Expresses the linear relation between restoring force and displacement of a point in an elastic system.

STATIC DISTRIBUTION OF CHARGE: One in which all charges are motionless; unchanged upon either time translation or time reversal. More restrictive than STATIONARY.

STATIONARY DISTRIBUTION OF CHARGE AND CURRENT: One in which appropriately averaged charge density and flux are unchanging in time; unchanged under time translation but not time reversal. Compare STATIC.

STERADIAN: Unit of angular area; see Appendix A.

STEREOSPECIFIC BINDING: Binding between two macromolecules that holds them in a particular relative orientation, with charge-complementary regions in apposition, and that arises only when those binding regions approach with that relative orientation.

STOKES PARAMETERS: Four real numbers describing the degree of polarization, overall intensity, and nature of the polarization of partially polarized light.

STOKES SHIFT: Difference between peak of excitation spectrum and emission spectrum for a particular fluorophore.

STRAIN RATE TENSOR: Twice the symmetric part of the tensor of derivatives of a fluid’s velocity vector field.

STREAMLINE OF A VECTOR FIELD: Parameterized curve in 3D space whose tangent vector is everywhere equal to a given vector field at each point along the curve. Or unparameterized but oriented curve whose tangent vector is everywhere proportional to a given vector field at each point.

STRESS, SHEAR: Off-diagonal component of the flux of momentum (force per area), typically generated by viscous drag, elasticity, or a combination.

STRESS TENSOR: This book instead uses the phrase “momentum flux tensor.” STRESS-ENERGY TENSOR: This book instead uses the phrase “energy–momentum flux tensor.”

SUBGROUP: Subset of a group that is itself a group; for example, the permutations of positive parity (“even” permutations).

SUPERPARAMAGNETIC BEAD: Submicrometer particle with magnetic susceptibility much larger than that of a bulk paramagnetic material.

SUPERPOSITION: Property of linear differential equations that a linear combination of solutions is a solution. INCOHERENT SUPERPOSITION: A linear combination of plane waves in a range of frequencies with random relative phases and polarizations.

SUPERSYMMETRY: Hypothetical property, involving a generalization of continuous group symmetry to one with anticommuting parameters.

SURFACE DENSITY: Term not used in this book; see AREAL DENSITY.

SURFACE, GAUSSIAN: A closed surface (no boundary, well defined interior) in 3-space to which we will apply Gauss’s theorem.

SUSCEPTIBILITY, DIELECTRIC: Tensor describing a linear relation between electric response and applied $\varepsilon_0\vec{E}$. That is, bulk electric dipole moment density per applied field of a linear dielectric medium, divided by $\varepsilon_0$. Usually applied to the numerical value, not to the qualitative property of being polarizable. MAGNETIC: Tensor describing a linear relation between magnetic response and applied $\vec{B}/\mu_0$. CROSS: Tensor describing a linear relation between electric response and applied $c\vec{B}$, or between magnetic response and $\vec{E}/c$.

SYMMETRIC TENSOR: One whose component $T_{ijk}$ equals the corresponding one after a specified index pair have been exchanged, for example, “symmetric on $i$ and $k$.” TOTALLY SYMMETRIC: symmetric on any rearrangement of the indices. SYMMETRIZE a tensor: To force symmetry by adding corresponding elements, for example $T^{[3]}_{ijk} = (T_{ijk} + T_{kij})/2$ is symmetric on $i$ and $k$.

SYMMETRY: Often used as a synonym for “invariance,” but in this book specifically an “active” transformation rule, one that turns one system trajectory into a different one without altering its status as a solution (or not) of the dynamical equations. Alternative meaning for tensors: see SYMMETRIC TENSOR.

SYNCHROTRON: Instrument that accelerates charged particles to relativistic velocity in a circular path. SYNCHROTRON RADIATION: EM radiation emitted by a charged particle traversing a curved path at relativistic speed.

TENSION, INTERFACIAL: Free energy per area, or equivalently force per length, to enlarge the interface between two regions in space. LINE: Force applied to an inextensible but flexible 1D object, for example, a floating thread separating two regions of an air–water interface, or a vibrating string in air. SPRING: Force (energy per extension to stretch the spring). SURFACE: The interfacial tension between a liquid and air or vacuum.

TENSOR "FROM HEAVEN": A tensor or tensor field that is not contingent on physical state, but instead reflects the geometry of space or spacetime. In electrodynamics the spacetime metric is such a tensor. The Levi-Civita tensor also has this property, but unlike the metric it depends on the arbitrary choice of which coordinate systems are deemed right-handed. (In Einstein gravity, both of these tensors instead encode information about the gravitational potentials, and hence are contingent on state.)

TENSOR PRODUCT: Generalization of DYAD PRODUCT to two tensors of any ranks; each component is the product of a component from the first times one from the second.

TESLA: SI unit for $\vec{B}$.

TEST BODY: Imagined limit of a point charge whose charge is too small to affect electromagnetic fields around it in a particular situation. Trajectories of test bodies may be used with the Lorentz force law to operationally measure electric and magnetic fields.

TETRAHEDRON: Regular polyhedron with four triangular faces.

TIME CONSTANT OF CABLE OR AXON: Time scale determined by the capacitance and conductance of the membrane.

TIMELIKE 4-VECTORS: One whose invariant inner product with itself is negative.

TIR: Total internal reflection.

TORUS: A two-dimensional surface in 3D space shaped like the surface of a doughnut.

TOTAL INTERNAL REFLECTION: Phenomenon where all the electromagnetic energy flux of a plane wave at high angle
Appendix G Glossary

of incidence is reflected back into a medium instead of being partially transmitted to surrounding air or vacuum.

**FRUSTRATED**: Phenomenon where energy jumps across such a “forbidden” region, so that a fraction of the incident flux does escape into a second “allowed” region nearby.

**TRACE OF A MATRIX**: Sum of its diagonal entries. Of a **TENSOR**: In 3D, the contraction on two specified indices, itself a tensor of rank reduced by two. In 4D, the contraction of a specified upper/lower index pair, or the contraction of an index pair with the appropriate metric. **TRACELESS MATRIX**: One whose sum of diagonal entries equals zero. **TRACELESS TENSOR**: One for which the contraction on two specified indices equals zero.

**TRAJECTORY**: Time history of one or more dynamical variables. For a single particle, a continuous chain of events, the locations of the particle at various times.

**TRANSFER MATRIX**: Mathematical tool for setting up the calculation of photonic bandgaps.

**TRANSITION DIPOLE**: Quantum-mechanical matrix element of the electric dipole moment operator.

**TRIGGER WAVE**: Nonlinear wave in which an excitable medium transitions from metastable to stable along a front which moves at constant velocity. A mathematically tractable simplification of the action potential.

**TRILINEAR**: See *BILINEAR*.

**TWEEZERS, OPTICAL**: Micromanipulation strategy relying partly on the dielectric contrast between a micrometer-scale bead and surrounding water; also called “laser tweezers.” **MAGNETIC**: Micromanipulation strategy relying on the force generated by a micrometer-scale bead’s induced or intrinsic moment by an external magnetic field gradient.

**TWINLEAD**: Simple waveguide consisting of two long parallel wires separated by air and plastic.

**UNIAXIAL**: Said of an object whose quadrupole moment tensor (of mass or charge) has two equal eigenvalues.

**UNIT MATRIX**: See *IDENTITY MATRIX*.

**UNIT TENSOR**: Same as identity tensor.

**UNITARITY**: Property of being unitary. See *MATRIX*.

**UNIT**: Standard quantity with specified dimensions. **BASE UNITS**: A particular choice of a unit for every dimension, for example, SI base units. **GAUSSIAN UNITS**: An non-SI set of base units combined with certain other notational conventions. **PLANCK UNITS**: Another alternative set of base units in terms of which some constants of Nature have simple numerical parts.

**UNORIENTED**: See *ORIENTATION*.

**UNPOLARIZED LIGHT**: Superposition of many different plane wave solutions with the same direction but varying polarization vectors, corresponding to the center of the Poincaré sphere.

**VELOCITY**: **ANGULAR**: see *ANGULAR VELOCITY*. **4-**: Derivative of 4-position along a trajectory with respect to proper time.

**VISCOS DRAG**: **FORCE**: Frictional force encountered by a body in viscous fluid. **TENSOR**: Expresses the linear relation between viscous drag force and velocity of the body relative to the fluid. Both of these quantities are proportional to **VISCOsITY**, a material parameter of the fluid describing internal friction.

**VOLTAGE**: Informal term for static or quasi-static electric potential. **VOLTAGE GATING**: Phenomenon in which individual ion channel proteins in a membrane change their conformation, and in turn alter their ion conductance, depending on applied electric field. **PROMPT V.G.**: Mathematical simplification in which the response to changes in field is instantaneous.

**VOLTAC CELL**: See cell.

**WAVE OPERATOR**: Contraction of 4-gradient with itself.

**WAVE PLATE, HALF-**: Optical device sometimes used to rotate polarization of light. **QUARTER**: Similar device, half as thick.

**WAVENUMBER**: Magnitude $k$ of the 3-wavevector.

**WAVEVECTOR**: In 3D, a vector that points along the direction of propagation of a plane wave, with magnitude (“WAVENUMBER”) equal to $2\pi$ divided by wavelength. In 4D, a 4-vector constructed from the wave’s frequency and 3D wavevector.

**WIRE**: One of a number of mathematical idealizations: In magnetism, charge flux is assumed to be everywhere zero except for a singularity along a mathematical curve (a “thin wire”). In circuit diagrams, a locus of points along which electric potential is assumed to be constant, neglecting self-inductance, mutual inductance, and resistance. (If self-inductance is included, this is represented by an inductor symbol on the diagram.)
$Z_2$: Group containing only two elements, such as $\pm 1$. 
Bibliography

Looking through this volume… was like roaming through an exquisite palace while its inhabitants slept.

— Orhan Pamuz

It is the possibility of our minds touching that makes writing a worthwhile endeavor at all.

— Ken Liu

Some of the articles listed below are published in “high-impact” scientific journals. It is important to know that frequently such an article is only the tip of an iceberg: Many of the technical details (generally including specification of any physical model used) are relegated to a separate document called Supplementary Information, or something similar. The online version of the article will generally contain a link to that supplement.


ADEE, S. 2023. We are electric: Inside the 200-year hunt for our body’s bioelectric code, and what the future holds. New York: Hachette.


Contents Index Notation Glossary November 14, 2023


IVANOV, D T, & NIKOLOV, S N. 2016. A new way to demonstrate the rainbow. Physics Teach., 54(8), 460–463.


Contents Index Notation Glossary November 14, 2023


Contents Index Notation Glossary November 14, 2023


SASLOW, W M. 2021. Voltaic cells: The good (Faraday), the bad (Volta), and the ugly (Galvani). *The Physics Teacher*, 59(1), 22–26.


Strutt, J W (Lord Rayleigh). 1887. XVII. On the maintenance of vibrations by forces of double frequency, and on the propagation of waves through a medium endowed with a periodic structure. *Phil. Mag. Ser. 5*, 24(147), 145–159.


---

Contents Index Notation Glossary November 14, 2023


Index

Bold references indicate the main or defining instance of a key term. Symbol names and mathematical notation are defined in Appendix B.

aberration
chromatic, 314, 314, 378, 694
spherical, 313, 329, 330, 332, 338, 377, 378
starlight, 458–460
Abraham, Max, 738
absolute temperature scale, 796
acceptor, 59, 60, 63
achiral object, 713, see nonchiral
action functional, 606
action potential, 90, 121, 124, 125, 133, 174, 175, 177, 179, 181, 189, 190
mechanical analog, 186
speed, 188
afterhyperpolarization, 182, 192
Airy
function, 379
integral (function), 374
Airy integral, 382
Airy, George, 373, 374
al-Farisi, 318
al-Shirazi, 318
Alexander of Aphrodias, 320
Alexander’s dark band, 320, 364, 368
ampere (unit), 797
Ampère law, 777
amplitude, complex, 279
angle, 798
azimuthal, 11
bipolar, 11
Angstrom (unit), 799
angular area, 11
angular momentum
flux tensor, 564, 616
angular velocity, 196
anion, 139
anomaly cancellation, 289, 522
antenna
dipole, electric, 647
theory, 399
aperture, 376
Arago, François, 577
arbitrary units, 800
arc length, 338, 339
arcminute (unit), 798
arcsecond (unit), 798
area, angular, 798
areal density, see density
arrow of time, 564
Aston, Francis, 480
ATP
dehydrogenase of, 144
Avogadro number, 813
axial symmetry, 50, 105, 107, 270, 339, 401, 473, 630, 637, 652, 741
axon, 95, 122, 174
giant, 182, 189
axoplasm, 95
Bacon, Roger, 318
Bartholin, Erasmus, 739
battery, 158, see voltaic cell
beam, 312, 315, 335, 619
Bessel, 623
gaussian, 622
Laguerre-gaussian, 623
waist, 622
beaming, 667, 672
beauty, 307, 379, 516, 521
visual, 316
benzene, 53
Berg, Howard, 92
Bessel beam, 623
direct, 623
Bessel function, 626
bipolar property, 44, 50, 54, 111
birefringence, 725
bifurcation, 351, 367, 369, 377, 381
Big Bang, see early Universe
bilayer membrane, 89, 95, 132, 133, 145
Biot-Savart formula, 231
birefringence
.circular, 731, 86, 688, 689
ordinate, 200, 713, 724
Birkhoff’s theorem, 42
Bjerrum length, 163
boiling point, 47
Boltzmann constant, 813
Boltzmann distribution, 89, 97, 141
boost
galilean, 410, 411, 422, 424, 428, 439, 442, 454, 462, 476
Lorentz, 415, 454, 454–457, 459, 460, 465, 467, 469, 472, 478, 483, 488, 498, 499, 503, 508, 517, 519, 580, 582
combination, 445, 467
provisional, 442, 448, 451–454
rapidity, 469
Born self-energy, 87, 89, 90, 144, 154
Born, Max, 87, 131, 334
Bradley, James, 451, 458, 470
Bremsstrahlung, 552, 709
de Broglie, Louis, 331, 505
cable equation
linear, 173
nonlinear, 188, 189
calcite, 323, 739–741
capacitance, 77, 150
capacitor, 121, 123, 133, 150
cartesian angle, see rainbow, caustic angle
catastrophe theory, 381
cation, 139
causality, 592
cautic, 316, 318–322, 335, 346, 362, 371, 373, 379, 381
angle, 320, 322, 362, 363, 367–370, 372
point, 375–379
Cavendish, Henry, 22, 30, 115, 119
CD, see circular dichroism
Celsius temperature scale, 796
Čerenkov radiation, 602, 708
Čerenkov, Pavel, 708
density, 3
areal, 145, 146
areal, bound, 80, 689
area, free, 691
bound, 86, 86, 688, 689
free, 689, 693
surface, bound, 85
surface, free, 80
electric, 797
flux, 3
1D, 115
2D, 117, 674, 676
3D, 117
4-vector, 534
bound, 689, 690
wire, 232
charge renormalization, 157
chicken-egg problem, 134, 146, 147, 488
chimera, 59
chiral medium, see medium
chirality, 713, 713
chirp, 757
chromophore, 780, 782
chutpah, 13
circular birefringence, 711, 712, 716–718, 724, 760
classical electron radius, 680
Clasius-Mosotti relation, 84, 85
clumping catastrophe, 143
CMBR, see cosmic microwave background radiation
Cockcroft, John, 480
coherent states, 775
coion, 146, 156
Cole, Kenneth S., 138
colloid, 139, 143, 287, 682
compliance tensor, 200
components of a tensor, 197, 209, 489–492, 497, 510, 528, 545, 802
3D, 195, 197
4D, 545
Compton scattering, 482, 683

858
strain rate, 495
Streamline, 8
streamline, iv, 7, 55, 239, 322–325, 338, 397, 401, 568–571, 652, 663
streamplot, 55, 403
stress-energy tensor, term not used in this book, see energy–momentum flux 4-tensor
shear, 201
tensor Maxwell, term not used in this book, 561
tensor, term not used in this book, see momentum flux 3-tensor
stress-energy tensor, term not used in this book, see energy–momentum flux 4-tensor
SU(2), SU(3), SU(5), see group subgroup, 452, 455, 490, 499
summation convention, 8
supergravity, 549
supernumerary bow, 317, 321, 362
superstationary phase point, 346, 368, 369, 371, 372, 377
supersymmetry, 616
surface attraction, electrostatic, 154
susceptibility 4-tensor operator, 721
dielectric, 81, 200, 692
anisotropic, 724
magnetic, 693
symmetry, 407
synchrotron radiation, iv, 663, 663
tau lepton, 549
Taylor cone, 108, 108
Taylor’s theorem, 799
tension interfacial or surface, 99, 100, 102, 102, 103, 105, 107, 113
line, 100, 102, 102, 103
tensor, 491
antisymmetric, 203, 493, 514
part, 226, 493
field, see also field
4D, 482, 510, 511, 528
energy–momentum flux, 613, see energy–momentum flux 4-tensor
Faraday, 513, 549
from Heaven, 209, 211, 213, 215, 217, 220, 498, 529, 579
spin, 547, 548
magnetic dipole moment, see multipole moments
magnetic field, see magnetic field
moment of inertia, see multipole moments, mass
operator, 207
product, 197, 492, 547, 548, 802, see also dyad
product, see also dyad
product group representation, 545, 547
Riemann, 109
strain rate, 495
symmetric, 197, 202–204, 492
part, 493
3D, 195, 490
compliance, 200
Levi-Civita, 204, 270
metric, 201
mobility, 200
moment of inertia, see multipole moments, mass
momentum flux, see momentum flux 3-tensor
polarizability, see polarizability
rank 2, 196, 492
rank 3, 204
spring constant, see spring tensor
totally antisymmetric, 204
viscous drag, see viscous drag tensor
trace, 202
traceless, 202
twisted, concept not used in this book, 217, 218
2D, 195
curvature, 109
metric, 111
Tensor Principle
3D, 495
4D, 511, 511, 522, 534, 536
tensor product, 4D, 496
4D, 530
tesla (unit), 244
test body, 3
the Rules
spinning, 4D, 547
4D, 548
Théodorich of Freiberg, 318
Thomson, J. J., 633
Thomson, William, 168–170, 173, 272, 299
threshold stimulus, 187, 189
time
proper, 478
retarded
distributed source, 393
point source, 597
time constant of cable or axon, 172
time-reversal invariance, 121
timelike separation, 195
transformation
active, 407
galilean, 412
Lorentz, see Lorentz transformation
passive, 407
transmission factor, 350, 742
traveling wave, 124, 173, 174, 181, 183, 185, 188, 189, 191, 192, 272, 275, 283, 299, 315, 421, 422, 424, 443
attenuated, 280
muscle, 126
trigger wave, 185
tweezers
magnetic, 263
optical, 48, 302, 312
twinkling, 379, 757
twinlead cable, 79
uncertainty relation, 783
uniaxial property, 44, 50, 54, 111, 739
birefringence, 725
uniform transparent medium, 338
unit tensor, see identity tensor
unitary matrix, 546, 815
units, 241–248, 796–800
base, 242, 796
dimensionless, 798
Système Internationale (SI), 796
vacuum state, 774
van der Waals interaction, 48, 143
Vavilov, Sergey, 708
velocity
3D, 4, 490
4D, 479, 497
null, see lightlike 4-vector
axial (term not used in this book), 215
contravariant (term not used in this book), 526
covariant (term not used in this book), 526
4D (4-vector), 511
polar (term not used in this book), 216
vector potential, 226, 783
velocity addition
Einstein, 444–445
galilean, 412
Verdet constant, 761
vesicle, 151
Victoria, Queen, 169
viscous drag
force, 199
tensor, 199
viscous drag tensor, 199
volt (unit), 244
voltage gating, 179, 181, 187–189
voltage-gating hypothesis
prompt, 184, 185
voltaic cell, 158–163
Walton, Ernest, 480
watt (unit), 797
wave equation, 421
inhomogeneous, see D’Alembert equation
Wave operator, 801
wave operator, 393, 454, 457, 467, 528, 528, 806
wavefront, 310, 622
wavelength, 276
wavenumber, 275
spectroscopic (alternate definition not used in this book), 276
wavevector, 279, 769
4D, 502, 575
weak interaction, 468
Weber, Wilhelm Eduard, 289
Weyl equation, 549
Wilkinson, David, 466
winding number, 623
world-line, 406
Yang–Mills theory, 518
Young, Thomas, 322, 362, 373, 577
Young–Laplace formula, 106
Young-Laplace formula, 109, 113
Zeeman, Pieter, 294

Contents Index Notation Glossary November 14, 2023