



University of Pennsylvania
ScholarlyCommons

Department of Physics Papers

Department of Physics


11-17-2022

Physics 516: Electromagnetic Phenomena (Spring 2023)

Philip C. Nelson

University of Pennsylvania, nelson@physics.upenn.edu

Follow this and additional works at: https://repository.upenn.edu/physics_papers

 Part of the [Astrophysics and Astronomy Commons](#), [Atomic, Molecular and Optical Physics Commons](#), [Biological and Chemical Physics Commons](#), [Condensed Matter Physics Commons](#), [Elementary Particles and Fields and String Theory Commons](#), [Optics Commons](#), and the [Statistical, Nonlinear, and Soft Matter Physics Commons](#)

Recommended Citation (OVERRIDE)

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/physics_papers/644
For more information, please contact repository@pobox.upenn.edu.

Physics 516: Electromagnetic Phenomena (Spring 2023)

Abstract

These course 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.

Keywords

electrodynamics

Disciplines

Astrophysics and Astronomy | Atomic, Molecular and Optical Physics | Biological and Chemical Physics | Condensed Matter Physics | Elementary Particles and Fields and String Theory | Optics | Physical Sciences and Mathematics | Physics | Statistical, Nonlinear, and Soft Matter Physics

Comments

PHYS5516 notes
University of Pennsylvania P. Nelson
Nov 2022

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



As for me, I distrust the commonplace;
Demand and am receiving marvels, signs,
Miracles wrought in air, acted in space
After imagination's own designs.
The lion and the tiger pace this way
As often as I call; the flight of wings
Surprises empty air, while out of clay
The golden-gourded vine unwatered springs.
— Adrienne Rich

We should not identify the naked skeleton we assume Nature to be with
the only real thing, the gay-coloured vesture of Nature.
— Oliver Heaviside, scribbled in the margin of a journal article.

I am no poet, but if you think for yourselves, as I proceed, the facts will
form a poem in your minds.
— Michael Faraday

Brief Contents

PART I Preliminaries

Prologue 2

Chapter 1 | Warmup: Newtonian Gravitation 16

The same hyperbolic trajectories taken by comets can also be relevant when charged particles are fired into matter.

PART II Static and Almost-Static

Chapter 2 | Electrostatics Introduced 26

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

Chapter 3 | Electrostatic Multipole Expansion 36

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

Chapter 4 | Vista: Fluorescence Resonance Energy Transfer 56

Resonance energy transfer creates a “private communication channel” between two fluorophores with a characteristic dependence on distance and on the orientations of the donor, the acceptor, and the vector separating them.

Chapter 5 | Curvilinear Coordinates and Separation of Variables 65

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

Chapter 6 | Capacitors 74

A charged capacitor will pull dielectric material into its gap.

Chapter 7 | Vista: Electrohydrostatics 96

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 112

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 126

The capacitance of an object can be measured without placing electrodes on either side of it.

Chapter 10 | Statistical Electrostatics of Solutions 132

DNA falls apart into separate strands in pure water.

Chapter 11 | The Cable Equation 161

Small electrical disturbances on a nerve axon spread diffusively.

Chapter 12 | Vista: Nerve Impulses 174

A nerve axon carries signals that preserve their form and amplitude as they travel long distances.

Chapter 13 | Examples of 3-Tensors in Physics 189

Molecular polarizability is in general anisotropic.

Chapter 14 | Tensors from Heaven 201

Although nematic liquid crystals are made from complicated molecules, only a few physical constants are needed to describe their overall behavior.

Chapter 15 | Magnetostatics 213

Tiny magnetic field disturbances can reveal brain activity without requiring invasive probes.

Chapter 16 | Units and Dimensional Analysis 231

Eddy currents slow the fall of a magnet through a conducting tube.

Chapter 17 | Magnetostatic Multipole Expansion 243

Just three constants suffice to characterize the far fields of even a complicated stationary current distribution.

PART III Dynamic**Chapter 18** | Beyond Statics 255

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]] 283

[[Vista: Johnson noise]] 285

Chapter 20 | First Look at Energy and Momentum Transport by Waves 286

The expansion of the early Universe was faster than predicted from gas pressure alone.

Chapter 21 | Ray Optics and the Eikonal 294

A lens with appropriately graded index can minimize spherical aberration.

Chapter 22 | [[Diffraction]] 312**Chapter 23** | [[Vista: Rainbows and Other Caustics]] 321

Sometimes the color sequence in a rainbow stutters.

Chapter 24 | Partial Polarization 324

For optical purposes, a plane wave of light may be described by a point inside an abstract sphere.

Chapter 25 | Generation of Radiation: First Look 330

An antenna emits energy with a specific directional pattern.

PART IV Relativity: Low Tech

Chapter 26 | Galilean Relativity 343

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 357

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 364

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 375

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 387

Each star's apparent position is shifted relative to others, depending on Earth's momentary velocity.

Chapter 31 | Relativistic Momentum and Energy of Particles 409

Huge amounts of energy can be liberated in nuclear reactions.

PART V Relativity: High Tech

Chapter 32 | Four-Vectors 420

Electrons, protons, and other "material" particles also show wavelike behavior.

Chapter 33 | The Faraday Tensor 440

The orbital period of a charged particle in uniform magnetic field starts to depend on its energy, when that energy is high.

Chapter 34 | Manifestly Invariant Form of Maxwell 453

A suddenly accelerated charge emits a pulse of electromagnetic radiation.

Chapter 35 | Energy and Momentum of Fields 478

x

Superconducting magnets can fail catastrophically.

Chapter 36 | Vista: Faraday's Field Lines 491

Like charges, and like magnetic pole tips, repel; opposite charges and pole tips attract.

Chapter 37 | Plane Waves in 4D Language 497

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 506

Each far-field wavefront of a dipole spherical wave is isotropic, but the resulting energy flux is not.

Chapter 39 | Beams: Gaussian, Vortex, and Bessel 512

A structured beam of light can transfer angular momentum far greater than that of a circularly polarized plane wave.

Chapter 40 | Vista: Variational Formulation 524

[Not ready yet.]

PART VI Radiation and Scattering

Chapter 41 | Radiation Green Function Revisited 537

A charged particle in uniform motion carries fields along with it but does not radiate.

Chapter 42 | Vista: J. J. Thomson's Pictorial Explanation of Radiation 553

The pulse of radiation from a suddenly accelerated charge consists of fields that are transversely polarized, have maximal strength in the equatorial plane, and fall with distance as $1/r$.

Chapter 43 | Electric Dipole Radiation 560

Homonuclear molecules have little infrared activity, but more complex ones can be strong greenhouse gases.

Chapter 44 | Higher-Multipole Radiation 573

Some radiative nuclear transitions are much faster than others.

[[Vista: Transition radiation]] 580

Chapter 45 | Synchrotron Radiation 581

[[Vista: Radiation Reaction]] 589

Chapter 46 | The Microwave Polarizer 590

An array of aligned, linear conductors can act as a polarizing filter, and can even regenerate a missing polarization in an incoming beam.

Chapter 47 | Scattering by Free and Bound Charges 594

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]] 601

[[Vista: Scattering by a Dielectric Sphere]] 602

PART VII Light in Materials

Chapter 49 | Light in Isotropic, Linear Media 604

A solution of randomly oriented molecules is completely isotropic, yet nevertheless can rotate polarized light.

[[Vista: Microscopy]] 624

Chapter 50 | Anisotropic, Linear Media 625

A transparent, crystalline material can change the polarization of light, even without chirality.

[[Vista: Optical Solitons in Nonlinear Media]] 631

Chapter 51 | Čerenkov Radiation 632

When a charged particle moves through a medium faster than the local speed of light, it emits radiation even without accelerating.

Chapter 52 | [[Energy in Media]] 637

Chapter 53 | [[Vista: Photonic Bandgap Materials]] 638

x.

Chapter 54 | Waves in a Cold Plasma and the Faraday Effect 642

Light from a black-hole accretion disk has a pattern of polarization.

Chapter 55 | [[Vista: Metamaterials]] 651

x.

PART VIII *Plus Ultra*

Chapter 56 | Vista: Field Quantization 655

Insects and crustaceans can “see” the polarization state of light.

Chapter 57 | [[Vista: Einstein's Gravitation]] 670

Chapter 58 | [[Vista: Classical Yang–Mills theories]] 671

Epilogue 672

Appendix A | Units and Dimensional Analysis 678

Appendix B | Global List of Symbols 683

Appendix C | Numerical Values 694

Appendix D | Animated graphics 696

Appendix E | Formulas 697

Detailed Contents

Web resources xxxiv

To the student xxxv

To the instructor xli

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 Two useful lemmas 9
 - 0.3.6 Euler theorem 10
 - 0.3.7 Angle and solid angle 10
 - 0.3.8 Delta function 10
- 0.4 What Lies Ahead 11
 - 0.4.1 Einstein's critique 11
 - 0.4.2 Some more hanging questions 13
- Track 2 14
 - 0.1.2' The notion of point charge 14
- Problems 15

Chapter 1 | Warmup: Newtonian Gravitation 16

- 1.1 Framing: *Interplay* 16
- 1.2 Space carries a physical function called the newtonian potential 16
- 1.3 An immobile point mass yields the familiar $1/r$ potential 17
- 1.4 Newton's law unifies celestial, terrestrial, and even laboratory measurements 19
- 1.5 Extended objects can be handled by combining fundamental solutions 19
- 1.6 *Plus Ultra* 20
- 1.7 More Hanging Questions 21

Track 2	22
1.6'	XX 22
Problems	23

PART II Static and Almost-Static

Chapter 2 | Electrostatics Introduced 26

2.1	Framing: <i>Coequal Partners</i>	26
2.2	Rephrase in Terms of a Potential	27
2.2.1	A static electric field can be re-expressed via an integrability lemma	27
2.2.2	Force law	28
2.2.3	An integrability lemma underlies the success of the potential method	29
2.3	Differences from Gravitation	29
2.4	Basic Solutions	29
2.4.1	Point charge	29
2.4.2	Continuous charge distribution	30
2.5	Conductors	30
2.6	Upcoming	31
2.6.1	Reality of electric field	31
2.6.2	Quasi-static	32
2.6.3	Beyond static	32
Track 2	33	
2.2'	Falsifiable content of the equations	33
Problems	34	

Chapter 3 | Electrostatic Multipole Expansion 36

3.1	Framing: <i>Distillation</i>	36
3.2	The Electrostatic Multipole Formula	36
3.3	Some Taylor Expansions	38
3.4	Derivation of the Formula	39
3.5	Multipole Moments Organize the Features of a Distribution According to Importance	39
3.6	More Remarks	40
3.6.1	Summary so far	40
3.6.2	From potentials to fields	40
3.6.3	Apparent singularity	40
3.6.4	All moments after the first nonzero one depend on choice of base point	41
3.6.5	Spherical distributions	41
3.6.6	Symmetry may dictate that some moments equal zero	41
3.6.7	Pure dipole is an idealization arising as a limiting case	42
3.7	Force and Torque on a Fixed Charge Distribution	43
3.7.1	Potential energy depends both on position and on orientation	43
3.7.2	Force and torque arise as derivatives of potential energy	44
3.7.3	Several intermolecular forces are dipolar in character	45
3.7.4	Dipole moment can be induced by an external field	46
3.8	<i>Plus Ultra</i>	46
Track 2	48	
3.2'a	Counting moments	48

3.2'b	Connection to spherical harmonics	48
3.7.3'a	Electric dipole moments of fundamental particles	48
3.7.3'b	Nuclear quadrupole moments	49

Problems 50

Chapter 4 | Vista: Fluorescence Resonance Energy Transfer 56

4.1	Framing: <i>A Private Channel</i>	56
4.2	Fluorescence and an unexpected phenomenon	56
4.2.1	Fluorescence microscopy is a versatile tool to image specific molecular actors	57
4.2.2	Resonant energy transfer defies naïve expectations	57
4.3	Dipole-mediated Transfer	60
4.3.1	Electrostatic near fields can be strong	60
4.3.2	FRET as a “spectroscopic ruler”	60
4.3.3	FRET depends on donor and acceptor orientation	61
4.4	<i>Plus Ultra</i>	63

Problems 63

Chapter 5 | Curvilinear Coordinates and Separation of Variables 65

5.1	Framing: <i>Level Sets</i>	65
5.2	Separation of Variables in the Laplace Equation	65
5.3	Familiar Examples	66
5.3.1	Cartesian coordinates	66
5.3.2	Plane polar coordinates	66
5.3.3	Plane polar payoff	68
5.3.4	Another hint about general relativity	68
5.3.5	Three dimensions	68
5.4	A Spherical Conductor in a Uniform Field	68
5.5	Lightning Rod Via Ellipsoidal Coordinates	69
5.6	Other Vector Operators	70
5.6.1	Hard way	70
5.6.2	Easy way	70
5.7	<i>Plus Ultra</i>	71

Problems 71

Chapter 6 | Capacitors 74

6.1	Framing: <i>Dielectrics</i>	74
6.2	Parallel Plates in Vacuum	74
6.3	The Energy Stored is Proportional to Volume	75
6.4	Cylindrical Conductors in Vacuum	76
6.5	Parallel Plates With Medium	76
6.5.1	Dielectric susceptibility describes the response of a material in linear approximation	76
6.5.2	An energy puzzle	78
6.5.3	Dense media roughly follow the Clausius–Mossotti relation	80
6.6	Nonuniform Polarization Leads to a Bound Charge Distribution	82
6.7	Charge Neutrality Breaks Down on the Nanoscale	83
6.8	Polar Fluid Media Can Be Highly Polarizable	84
6.9	Partitioning of Ions	85

6.9.1	Solubility of ionic solids follows a simple quantitative rule	85
6.9.2	Partitioning at a fluid interface or cell membrane; permeability	86
6.10	Boundary Conditions	86
Track 2		89
6.5.1'	Ferroelectricity, electrostriction and piezoelectricity	89
6.8'	Electrorotation	89
Problems		90

Chapter 7 | Vista: Electrohydrostatics 96

7.1	Framing: <i>An Impossible Shape</i>	96
7.2	Some geometry of Curves and Surfaces	97
7.2.1	Curves in a plane can be characterized by a single curvature function	97
7.2.2	Mechanical equilibrium of an interface in a plane	99
7.2.3	Surfaces in space have two distinct curvature functions	100
7.2.4	The Young–Laplace formula describes a trade-off between surface tension and pressure	102
7.3	Effect of Electric Field	103
7.3.1	An electric field jump across an interface modifies the energy balance	103
7.3.2	The modified mechanical equilibrium admits a conical point solution	104
7.4	Technological applications	105
7.5	<i>Plus Ultra</i>	105
7.5.1	A look ahead	106
7.5.2	Other physical surfaces	106
7.5.3	A glimpse of general relativity	106
Track 2		108
7.2.3'	Metric and second fundamental form	108
7.2.4'	Derivations of variational formulas	108
Problems		110

Chapter 8 | Charge Flux, Continuity Equation, and Ohmic Conductors 112

8.1	Framing: <i>Conservation</i>	112
8.2	A Graphical Argument for the 1D Continuity Equation	112
8.3	Two or More Dimensions	113
8.3.1	Any local conservation rule leads to a continuity equation	113
8.3.2	The continuity equation bridges local and global conservation	114
8.4	Remarks	114
8.5	Nonstatic Situations	115
8.5.1	Conductivity, resistivity, conductance, resistance	115
8.5.2	Salt water conducts electricity via the motions of ions	116
8.6	Quasi-static Situations	117
8.7	Electroencephalogram/Electrocardiogram	117
8.7.1	A steady current source in solution again leads to a Poisson equation	117
8.7.2	An isolated neuron creates an exterior potential	118
8.7.3	Electroencephalogram	121
8.7.4	Electrocardiogram	122
Problems		123

Chapter 9 | Vista: Cell Membrane Capacitance 126

9.1	Framing: <i>Noninvasive Measurement</i>	126
-----	---	-----

- 9.2 Fricke's Experiment 127
 - 9.2.1 Setup and solution 127
 - 9.2.2 The membrane stores electrostatic energy despite not being "in series" with the applied potential 129
 - 9.2.3 The experimentally measured phase lag determines the capacitance 129
- Problems 131

Chapter 10 | Statistical Electrostatics of Solutions 132

- 10.1 Framing: *Ion Clouds* 132
- 10.2 Solution is Different from Vacuum 132
 - 10.2.1 The Nernst relation sets the scale of membrane potentials 132
 - 10.2.2 The electrical conductivity of a solution reflects frictional dissipation 135
- 10.3 A Repulsive Interlude 135
 - 10.3.1 Electrostatic interactions are crucial for proper functioning of living cells 135
 - 10.3.2 The Gauss law 138
 - 10.3.3 Detailed form of the neutralizing ion cloud outside a charged surface in pure water 139
 - 10.3.4 Excess salt shrinks the electric double layer 143
 - 10.3.5 The repulsion of like-charged surfaces arises from compression of their ion clouds 144
 - 10.3.6 DNA denatures in pure water 146
- 10.4 Oppositely Charged Surfaces Attract by Counterion Release 146
- 10.5 *Plus Ultra* 147
- Track 2 148
 - 10.2.2' Electric currents in metals 148
 - 10.3.4'a Solutions with added salt or acid 148
 - 10.3.4'b Low-salt limit 149
 - 10.3.4'c Far field limit 149
 - 10.3.4'd Weakly charged limit; linearized Poisson-Boltzmann equation 149
 - 10.3.4'e Stored energy 150
 - 10.3.4'f How voltaic cells push electrons 150
 - 10.3.5' Alternative derivation of force 155
- Problems 158

Chapter 11 | The Cable Equation 161

- 11.1 Framing: The Ill-Fated Transatlantic *Cable* 161
- 11.2 Coaxial Cable 162
 - 11.2.1 A mathematical hyperlink to heat conduction 163
 - 11.2.2 Discrete-element models as stepping-stones to distributed elements 164
 - 11.2.3 The linear cable equation explains the observed dispersion of signals 165
- 11.3 Neurons 167
 - 11.3.1 Nerve impulses propagate without dispersion or attenuation 167
 - 11.3.2 Some ion species are far out of equilibrium 168
 - 11.3.3 Linear cable equation for an axon 170
 - 11.3.4 Threshold behavior foreshadows a role for nonlinearity 171
- Problems 173

Chapter 12 | Vista: Nerve Impulses 174

- 12.1 Framing: *Nonlinearity* 174
- 12.2 The Time Course of an Action Potential Confirms the Hypothesis of Non-Ohmic Conductance 174

12.3	Voltage Gating Leads to a Nonlinear Cable Equation With Traveling Wave Solutions	178
12.3.1	A purely mechanical system with traveling, solitary waves	178
12.3.2	Voltage gating leads to bistability	179
12.3.3	The nonlinear cable equation	181
12.3.4	Solution	181
12.3.5	Interpretation	182
12.4	<i>Plus Ultra</i>	183
Track 2		185
12.3'	Velocity selection in more general models	185
12.4'a	More detailed models	186
12.4'b	FitzHugh–Nagumo model	186
12.4'c	Solitons	186
Problems		188
Chapter 13	Examples of 3-Tensors in Physics	189
13.1	Framing: <i>Anisotropy</i>	189
13.2	Rank Zero; Rank One	189
13.3	Rank Two	190
13.3.1	A tensor can represent a vector-valued, linear function of vectors	190
13.3.2	More general examples of rank-two tensors	192
13.3.3	Tensors arise naturally throughout physics: some examples	192
13.3.4	A symmetric tensor can also represent a scalar-valued quadratic function of a vector	195
13.3.5	Some linear vector functions, but not all, arise as the derivative of a quadratic scalar function	196
13.4	Rank Three	196
13.4.1	Levi-Civita as a vector-valued bilinear function of vectors	196
13.4.2	Levi-Civita as a scalar-valued trilinear function of vectors	197
13.5	Tensor Fields	198
Track 2		199
13.2'a	Vectors and their duals	199
13.2'b	Tensor properties of probability density functions	199
13.3'a	Tensors in quantum mechanics	199
13.3'b	Another concept of rank	199
Problems		200
Chapter 14	Tensors from Heaven	201
14.1	Framing: <i>Intrinsic Structures</i>	201
14.2	The Components of a Tensor Transform Upon Linear Change of Coordinates	201
14.2.1	An example from mechanics	201
14.2.2	Cartesian coordinates are connected via orthogonal matrices	202
14.2.3	The components of the 3D metric are the same in any cartesian system	203
14.3	Components of the Levi-Civita Tensor	203
14.3.1	The components of ϵ are the same in any right-handed cartesian system	203
14.3.2	Components only specify a unique ϵ after a right-hand convention is chosen	204
14.3.3	<i>Plus Ultra</i>	205
14.4	Connect to Familiar Things	205
14.4.1	Dot product	205
14.4.2	Cross product	205
14.5	Useful Identities	206
14.5.1	Swap dot and cross	206

14.5.2	Triple cross product	206
14.6	<i>Plus Ultra</i>	207
Track 2		209
14.3'	Spatial inversion invariance	209
14.4'	Twisted tensors	209
Problems		212

Chapter 15 | Magnetostatics 213

15.1	Framing: <i>Integrability</i>	213
15.2	A New Force Awakens	214
15.3	Vector Potential	215
15.3.1	No scalar potential this time	215
15.3.2	Lemma to a lemma	215
15.3.3	Revisit electrostatics	216
15.3.4	The magnetic Gauss law expresses an integrability condition	217
15.3.5	The Poincaré lemma applies in any number of dimensions, and to tensors of any rank	218
15.4	Gauge Invariance and Coulomb Gauge	219
15.5	Back to Physics	219
15.5.1	Steady currents	219
15.5.2	Axial symmetry suggests a solution to the Oersted problem	220
15.5.3	The electrostatic Green function also solves the magnetostatic equations	220
15.5.4	Self-consistency	221
15.5.5	Some of the equations are vacuous, resolving a counting puzzle	221
15.6	Biot–Savart Formula	222
15.6.1	Second solution to Oersted, via vector potential	222
15.6.2	\vec{B} for a general current distribution	222
15.6.3	More about thin wire approximation	223
15.7	Boundary Conditions	224
15.8	Magnetoencephalography	225
15.9	<i>Plus Ultra</i>	225
Track 2		226
15.2'	Puzzle about angular momentum conservation	226
15.9'a	About magnetic monopoles	226
15.9'b	Elimination of pseudovectors	226
15.9'c	Differential forms	227
Problems		228

Chapter 16 | Units and Dimensional Analysis 231

16.1	Framing: <i>Communication</i>	231
16.2	Time, Length, and Mass	231
16.2.1	Base units in mechanics	231
16.2.2	Elimination of units is an abbreviation	232
16.3	Units in Electrodynamics	233
16.3.1	The SI base unit of charge is the coulomb	234
16.3.2	Derived SI units	235
16.3.3	The gaussian base unit of charge is the statcoulomb	235
16.3.4	The gaussian system involves two additional conventions	236
16.3.5	What is an “esu”?	238
16.4	Remarks	238

xx

Track 2 240

16.3.1' Why base the SI on the proton? 240

16.3.3'a Planck units 240

16.3.3'b Elimination of more units 240

Problems 241

Chapter 17 | Magnetostatic Multipole Expansion 243

17.1 Framing: *Distillation Again* 243

17.2 Tensor Preliminaries 243

17.3 Far Fields of a Steady, Localized Current Distribution 244

17.3.1 The magnetic dipole vector potential is the leading term in a series expansion 244

17.3.2 A familiar example 245

17.4 Higher Moments 245

17.4.1 The magnetic quadrupole potential falls faster than the dipole 245

17.4.2 All moments after the first nonzero one are basepoint-dependent 247

17.5 Magnetic Dipole in an External Field 247

17.5.1 Force and torque on a dipole of fixed strength 247

17.5.2 Diamagnetism, paramagnetism, ferromagnetism 249

17.5.3 Purification of oxygen via diamagnetic forces 249

17.5.4 Magnetic levitation of macroscopic objects at room temperature 249

Problems 249

PART III Dynamic

Chapter 18 | Beyond Statics 255

18.1 Framing: *Self-consistency* 255

18.2 Review 255

18.2.1 Field equations 255

18.2.2 A coil carrying constant current 255

18.3 Time-dependent currents 257

18.3.1 Faraday observed an \vec{E} field associated to a time-varying magnetic field 257

18.3.2 Work must be done to increase current through a solenoid 258

18.3.3 Self-inductance also affects signal propagation along a cable 259

18.3.4 Magnetic field energy is proportional to volume 260

18.4 Maxwell's Modification to Ampère's Law 261

18.4.1 Mathematical consistency hinges on the continuity relation for charge 261

18.4.2 Boundary conditions 262

18.5 Wave Solutions 263

18.5.1 About traveling plane waves 263

18.5.2 The final form of the vacuum Maxwell equations have plane wave solutions 264

18.6 Points Remaining 265

18.7 Complex Exponential Notation for Waves 266

18.7.1 Electric Gauss law 266

18.7.2 Faraday law 267

18.7.3 Magnetic Gauss law 267

18.7.4 Ampère law 267

18.7.5 Traveling wave with attenuation 268

18.7.6 Summary 268

18.8	Potentials Beyond Statics	268
18.8.1	\vec{E} and \vec{B} can still be represented by using potentials	268
18.8.2	Gauge invariance and Coulomb gauge also extend beyond statics	268
18.8.3	Coulomb gauge can be augmented if charge density is zero	269
18.9	Waves via Potentials	270
18.10	Complex Polarizations	271
18.10.1	Linear, circular, elliptical	271
18.10.2	Helicity basis for circular polarization	271
18.10.3	Spherical waves foreshadowed	272
18.11	<i>Plus Ultra</i>	272
Track 2		274
18.4.1'a	Connection to ohmic materials	274
18.4.1'b	Stumbling yet pulled forward	274
18.6'a	On the speed of light	275
18.11'	On the guidance of mathematical consistency	276
Problems		277
Chapter 19	[[Vista: Waveguides]]	283
Problems		283
	[[Vista: Johnson noise]]	285
Chapter 20	First Look at Energy and Momentum Transport by Waves	286
20.1	Framing: <i>Pressure</i>	286
20.2	Linear Polarization	286
20.2.1	Electromagnetic waves transport energy	286
20.2.2	Although momentum is a vector, its transport in a wave does not time-average to zero	287
20.2.3	Radiation pressure underpins many electromagnetic phenomena	288
20.3	Light Cannot Be Interpreted As a Stream of Newtonian Particles	289
20.4	Circular and Elliptical Polarizations	289
20.5	Electromagnetic Waves can Also Transport Angular Momentum	290
Track 2		291
20.2.2'	[[Ponderomotive force and the Paul trap]]	291
Problems		292
Chapter 21	Ray Optics and the Eikonal	294
21.1	Framing: <i>Almost-Plane Waves</i>	294
21.2	Light Still has Plane Wave Solutions in a Uniform Medium	295
21.3	Piecewise-uniform Medium	295
21.3.1	The refraction law arises from matching fields across a planar boundary	295
21.3.2	Optical tweezers exploit the momentum transfer implied by refraction	296
21.3.3	Spherical aberration limits the practical focusing power of glass lenses	296
21.3.4	Total internal reflection arises when there is no solution to the refraction equation	298
21.4	Gradient-index Medium	299
21.4.1	Rays of light can be regarded as streamlines of energy flux	299
21.4.2	Almost-plane waves are a useful idealization when there is a separation of length scales	299
21.4.3	The eikonal equation controls propagation of an almost-plane wave	300
21.4.4	Rays in vacuum	301

21.4.5	Rays bend continuously in a gradient-index medium	302
21.4.6	Shortwave radio skip (skywave transmission)	303
21.5	More Phenomena	304
21.5.1	Mirages rely on our brains' assumptions about light propagation	304
21.5.2	A spherical gradient-index lens can minimize spherical aberration	305
21.5.3	Gravitational fields can bend light rays even in vacuum	305
21.6	<i>Plus Ultra</i>	306
	Problems	307

Chapter 22 | [[Diffraction]] 312

22.1	Waves	312
22.2	One Thin Slit	315
22.3	Two or More Thin Slits	315
22.4	One Fat Slit	317
22.5	Ray Optics	318
22.6	<i>Plus Ultra</i>	320

Chapter 23 | [[Vista: Rainbows and Other Caustics]] 321

23.1	Framing	321
	Problems	323

Chapter 24 | Partial Polarization 324

24.1	Framing: <i>Stokes Parameters</i>	324
24.2	Light as an Ensemble	324
24.2.1	Most sources give chaotic light	324
24.2.2	Optical instruments ultimately measure energy deposition	324
24.2.3	Steady sources: Replace time average by ensemble average	325
24.3	Some Convenient Models of Light	326
24.3.1	Fully polarized light corresponds to the periphery of the Poincaré sphere	326
24.3.2	Simplified model of unpolarized light	327
24.3.3	Partial Polarization	327
24.4	How to Measure the Stokes Parameters	328
	Problems	328

Chapter 25 | Generation of Radiation: First Look 330

25.1	Framing: <i>Slow Falloff</i>	330
25.2	Review: Green Function Solutions to Electro- and Magnetostatics	331
25.3	A Physically Motivated Guess for the Radiation Green Function	331
25.4	Check the Guess	332
25.5	Our First Antenna	333
25.5.1	A closed current loop can carry current without charge building up anywhere	333
25.5.2	Far from the source, the fields fall as $1/r$	333
25.5.3	Net energy escapes to infinity	336
25.5.4	The loop antenna is directional	336
25.6	<i>Plus Ultra</i>	336
	Track 2	337
25.5.1'	Realistic antenna theory requires a self-consistent solution	337
	Problems	338

PART IV Relativity: Low Tech

Chapter 26 | Galilean Relativity 343

- 26.1 Framing: The *Principle of Relativity* 343
- 26.2 An Illustration from Mechanics 344
- 26.3 Active Viewpoint: Symmetry 344
- 26.4 Passive Viewpoint: Invariance 345
 - 26.4.1 Alternative representations of the same physical situation 345
 - 26.4.2 Relation between active and passive 345
- 26.5 Rotations And Dilations are Both Linear, but Only Rotations are Invariances 346
- 26.6 Galilean Group 347
 - 26.6.1 Some coordinate systems on spacetime are preferred 347
 - 26.6.2 Boosts connect coordinate systems in relative motion 348
 - 26.6.3 Matrix notation 349
 - 26.6.4 Galilean transformations have a group structure 349
 - 26.6.5 The physical significance of invariance 350
 - 26.6.6 Light cannot be interpreted as a stream of newtonian particles, part 2 352
- 26.7 1905 and All That 353
- Track 2 354
 - 26.1' Complete isolation 354
 - 26.6.1'a Parity invariance 354
 - 26.6.1'b Time reversal symmetry 354
- Problems 356

Chapter 27 | Springs, Strings, and Local Conservation Laws 357

- 27.1 Framing: *Transport* 357
- 27.2 Equation of Motion 357
 - 27.2.1 Longitudinal vibration 357
 - 27.2.2 Transverse vibration 358
- 27.3 The wave equation seems to lack boost invariance 358
- 27.4 Invariance Regained 360
- 27.5 Connection to Electromagnetism 361
- 27.6 Continuity Relations for Energy and Momentum 361
 - 27.6.1 Energy and momentum each have local expressions for their density and flux 361
 - 27.6.2 Energy and momentum are both locally conserved 363
- 27.7 *Plus Ultra* 363
- Problems 363

Chapter 28 | Einstein's Version of Relativity: Overview 364

- 28.1 Framing: *Conservative Revolution* 364
- 28.2 The \AA ther Hypothesis 364
- 28.3 The No- \AA ther Hypothesis 366
 - 28.3.1 The vacuum is a unique state 366
 - 28.3.2 Follow the symmetry 367
- 28.4 Where We are Heading 367
- Track 2 369
 - 28.2'a Michelson–Morley 1887 369

- 28.2'b More about uniqueness of the vacuum state 372
- 28.2'c In praise of æther 373
- 28.3' Poincaré's work 373

Problems 374

Chapter 29 | Provisional Lorentz Transformations and the Fizeau Experiment 375

- 29.1 Framing: *Dragging Light* 375
 - 29.2 Review 375
 - 29.2.1 Galilean invariance predicts simple addition of velocities 375
 - 29.2.2 Æther skeptics have some explaining to do 376
 - 29.3 Graphical Explorations Suggest a Form for Boost Transformations 377
 - 29.4 The Wave Equation is Invariant Under Provisional Lorentz Transformations 378
 - 29.4.1 Coordinate transformation 378
 - 29.4.2 Active viewpoint 378
 - 29.4.3 Passive viewpoint 379
 - 29.5 Einstein's Velocity Addition 379
 - 29.6 A Nonnull, Falsifiable Prediction 381
 - 29.7 *Plus ultra* 383
- Problems 386

Chapter 30 | Aberration of Starlight and Doppler Effects 387

- 30.1 Framing: A *Greedy Principle* 387
 - 30.2 Again No Dilation Invariance 388
 - 30.3 Lorentz Transformations in One Space Dimension 389
 - 30.3.1 A subgroup that excludes dilations 389
 - 30.3.2 Rapidity parameter 391
 - 30.4 A Typical Paradox and its Resolution 392
 - 30.5 Lorentz Transformations in Three Space Dimensions 393
 - 30.6 More Key Experiments: Aberration of Starlight and Doppler Shift 394
 - 30.6.1 Light-speed trajectories bend while remaining at light speed 394
 - 30.6.2 Wave frequency transforms in an angle-dependent way 396
 - 30.7 An Enormous Generalization 397
 - 30.7.1 Lorentz invariance must apply to all of physics 397
 - 30.7.2 Muon lifetime, galactic redshifts, CMBR dipole, and more 398
 - 30.8 What's Next 401
- Track 2 402
- 30.3'a Light-cone coordinates 402
 - 30.3'b Reformulation of the invariant interval 402
 - 30.3'c Velocity addition in light-cone coordinates 402
 - 30.3'd Relation to rapidity 402
 - 30.6.2' Another view of the longitudinal Doppler shift 403
 - 30.7.2' More about muon lifetime 403

Problems 404

Chapter 31 | Relativistic Momentum and Energy of Particles 409

- 31.1 Framing: *Inseparable Aspects* 409
- 31.2 Conservation of Newtonian Energy and Momentum is not Compatible with Lorentz Invariance 409
- 31.3 Conservation Laws Recovered 411

- 31.3.1 “Einstein thinking” places symmetry first 411
- 31.3.2 What has/has not been shown 413
- 31.3.3 A geopolitical consequence 414
- 31.4 Particles with Speed at or Near c 415
 - 31.4.1 Any particle interpretation of light must involve the limit of zero mass 415
 - 31.4.2 Interactions involving massless particles 415
- 31.5 *Plus Ultra* 416
- Problems 417

PART V Relativity: High Tech

Chapter 32 | Four-Vectors 420

- 32.1 Framing: *Unification* 420
- 32.2 How to Avoid Reading This Chapter 420
- 32.3 3D Review 422
 - 32.3.1 Rotations preserve the form of the metric 422
 - 32.3.2 Equations of the form (3-vector) = 0 are rotationally invariant 423
 - 32.3.3 The 3-tensor transformation rule 424
 - 32.3.4 Symmetric and antisymmetric 3-tensors 425
 - 32.3.5 3D contraction is another invariant operation 426
- 32.4 Other Rotationally Invariant Systems in Mechanics 426
 - 32.4.1 Newtonian gravitation 426
 - 32.4.2 Field equations: the gradient operator 427
- 32.5 Summary: The Rules in 3D 427
- 32.6 Four Dimensions 429
 - 32.6.1 Packaging 429
 - 32.6.2 The Lorentz group and its main subgroups 430
 - 32.6.3 The invariant interval 432
 - 32.6.4 4D contraction 432
 - 32.6.5 The four-velocity invariantly characterizes a particle trajectory 433
 - 32.6.6 Summary and first payoff: the 4-wavevector and its transformation rule 434
- 32.7 Momentum and Energy Revisited 435
 - 32.7.1 Beyond $\mathcal{E} = mc^2$ 435
 - 32.7.2 The 4-momentum of a massless particle is a null 4-vector 436
 - 32.7.3 de Broglie’s prediction for electron wavelength was dictated by Lorentz invariance 436
 - 32.7.4 Particle creation and destruction 436
- Problems 437

Chapter 33 | The Faraday Tensor 440

- 33.1 Framing: *Symmetries as Drivers* 440
- 33.2 4-tensors 440
 - 33.2.1 Tensor products of Lorentz transformations 440
 - 33.2.2 An extended Tensor Principle 441
- 33.3 Lorentz Force Law 442
 - 33.3.1 The Faraday tensor unifies electric and magnetic force laws 442
 - 33.3.2 Relate to traditional form 443
 - 33.3.3 More on the marriage of \vec{E} and \vec{B} 444
 - 33.3.4 On beauty 445

33.3.5	Better than beauty: an experimental consequence for cyclotron motion	445
33.4	Transformation of the Faraday Tensor	446
33.4.1	Electric and magnetic fields mix under Lorentz boosts	446
33.4.2	A charge in uniform, straight-line motion	447
33.5	Summary	448
33.6	<i>Plus Ultra</i>	449
Track 2		450
33.3.4'	More on beauty	450
33.6'a	Bigger symmetry groups	450
Problems		451

Chapter 34 | Manifestly Invariant Form of Maxwell 453

34.1	Framing: the <i>Rules</i>	453
34.2	Field Equations in 4D	453
34.2.1	The 4-gradient transforms differently from any 4-vector	454
34.2.2	The wave operator is the invariant contraction of two 4-gradients	456
34.3	General 4-tensors	456
34.3.1	Rank	456
34.3.2	Symmetry	456
34.3.3	The metric is itself a tensor	457
34.4	Summary: The Rules in 4D	457
34.5	Vacuum Maxwell Equations	458
34.6	The Charge Flux 4-Vector	460
34.6.1	The graphical formulation unifies charge density and flux	460
34.6.2	\underline{J} is a 4-vector	462
34.7	Complete, Invariant Maxwell Equations	462
34.8	Four-vector Potential	463
34.8.1	The Poincaré lemma again implies the existence of a potential	463
34.8.2	Particle in uniform motion revisited	464
34.9	More About \underline{J}	464
34.9.1	The delta function composed with an ordinary function is still a delta function	465
34.9.2	\underline{J} may alternatively be formulated in terms of individual trajectories	465
34.9.3	Another proof that \underline{J} is a 4-vector	466
34.10	A Dizzying Vista	467
34.11	<i>Plus Ultra</i>	468
Track 2		469
34.7'a	Degeneracy of Maxwell equations	469
34.7'b	Spatial inversion (parity) invariance	469
34.7'c	Time-reversal invariance	469
34.8.1'a	Counting equations, again	470
34.8.1'b	p -form gauge fields	470
34.9'	Geometric status of the charge flux	470
34.11'	Spinors	471
Problems		476

Chapter 35 | Energy and Momentum of Fields 478

35.1	Framing: <i>Local Conservation</i>	478
35.2	What Needs to Be Shown and Why	478

35.3	Continuity Equation for Energy and Momentum in the Absence of Long-Range Forces	479
35.4	Interactions Seem to Spoil Local Conservation	481
35.4.1	Long-range forces	481
35.4.2	Nonconservation of particle energy and momentum	482
35.5	Accounting for Field Contributions Restores Local Conservation of Energy and Momentum	483
35.6	What Has Been Accomplished	484
35.6.1	Poynting's theorem fits with older ideas	485
35.6.2	Magic without magic	486
Track 2		487
35.5'	Angular momentum flux tensor	487
35.6.1'	Connect to other idealized circuit elements	487
Problems		489

Chapter 36 | Vista: Faraday's Field Lines 491

36.1	Framing: <i>Toolkit</i>	491
36.2	Field Lines are Mathematically Similar to the Streamlines of an Incompressible Fluid	492
36.3	Electric and Magnetic Forces via Derivatives of Field Energy	493
36.4	Forces Via the Stress 3-Tensor	494
36.4.1	Electrostatic forces can be pictured as tension along, or pressure among, field lines	494
36.4.2	Magnetostatic forces have similar pictorial expressions	495
36.5	Magnetic Induction	495
Problems		495

Chapter 37 | Plane Waves in 4D Language 497

37.1	Framing: <i>Independent Channels</i>	497
37.2	Lorenz Gauge Choice	497
37.2.1	It's useful	497
37.2.2	It's permitted	498
37.3	Plane Waves and the Polarization 4-Vector	498
37.4	Energy and Momentum Carried by a Plane Wave Confirm Earlier Expectations	499
Track 2		502
37.2'a	Gravitational waves	502
37.2'b	Spin versus polarization	502
Problems		503

Chapter 38 | A Simple Spherical Wave 506

38.1	Framing: <i>Dipole Doughnut</i>	506
38.2	An Exact Solution With Spherical Wavefronts	506
38.2.1	Analogy to acoustics	506
38.2.2	Far fields carry energy in an anisotropic pattern	507
38.2.3	Near fields resemble a time dependent electric dipole	508
38.3	A Circularly Polarized Spherical Wave?	508
38.4	Interference	508
38.5	Summary	509
Problems		510

Chapter 39 | Beams: Gaussian, Vortex, and Bessel 512

39.1	Framing: <i>Diffractive Spreading</i>	512
39.2	Gaussian Beams	512
39.2.1	Paraxial approximation creates a mathematical analogy to the Schrödinger equation	513
39.2.2	The gaussian beam spreads slowly, although its wavefronts curve	514
39.3	Optical-Vortex Beams Transport Angular Momentum Even When Linearly Polarized	515
39.4	Transfer of Angular Momentum to a Sphere	515
39.5	Bessel Beams	516
39.5.1	An idealized solution with no diffractive spreading at all	516
39.5.2	A physically realizable approximation to the ideal	516
39.5.3	Application to microscopy	519
	Problems	522

Chapter 40 | Vista: Variational Formulation 524

40.1	Framing: <i>Noether Theorem</i>	524
40.2	Variational Formulation of Newtonian Mechanics	525
40.3	Variational Formulation of Field Equations	526
40.3.1	Local lagrangian densities and their variational equations	526
40.3.2	A simple lagrangian density leads to the Maxwell equations in vacuum	527
40.3.3	Fields plus charged particles	528
40.4	Continuous Invariances Lead to Conservation Laws	530
40.4.1	Scalar field example	530
40.4.2	Consequences of invariance: the Noether theorem	530
40.4.3	Translational invariance of electrodynamics leads to the same \underline{T} as was found previously	532
40.5	<i>Plus Ultra</i>	533
	Track 2	534
40.4'a	Angular momentum	534
40.4'b	Classical fermion fields and supersymmetry	534
	Problems	535

PART VI Radiation and Scattering

Chapter 41 | Radiation Green Function Revisited 537

41.1	Framing: <i>Lookback</i>	537
41.2	The Relativity of Time Ordering Constrains Causality	537
41.3	Green Function for the d'Alembert equation	539
41.3.1	Lorentz invariance tightly constrains the Green function	539
41.3.2	Reformulate and confirm the trial solution	541
41.4	Remarks	541
41.4.1	Upgrade to 4-vector fields	541
41.4.2	Check self-consistency	542
41.5	Point Particle Executing Specified Motion	543
41.5.1	The Liénard–Weichert potentials follow from the Green function solution	543
41.5.2	Uniform motion once again	545
41.5.3	Coda	549
	Track 2	550
41.3'	Alternative derivation in Fourier space	550
41.4'	Gravitational radiation	551
	Problems	552

Chapter 42 | Vista: J. J. Thomson's Pictorial Explanation of Radiation 553

- 42.1 Framing: *Kinks* 553
- 42.2 Electric Fields From a Suddenly Accelerated Charge 554
- 42.3 Magnetic Fields Also Have a Radiation Contribution 556
- Problems 558

Chapter 43 | Electric Dipole Radiation 560

- 43.1 Framing: *Double Expansion* 560
- 43.2 Three Length Scales 560
 - 43.2.1 Small source 561
 - 43.2.2 Harmonic time variation 562
 - 43.2.3 In many applications, the multipole parameter is small 563
- 43.3 Electric Dipole Radiation 563
 - 43.3.1 A time-varying ED moment leads to $1/r$ potentials 563
 - 43.3.2 Pure dipole limit 564
- 43.4 The Electric and Magnetic Fields Fall Slowly With Distance 564
- 43.5 Concrete Examples 565
 - 43.5.1 Electric dipole antenna 565
 - 43.5.2 Greenhouse gases absorb and radiate via molecular dipole moments 567
- 43.6 Energy Flux and Total Power Scale as ω^4 567
- 43.7 Linear Polarization Recovers the Dipole Doughnut Pattern 569
- Problems 571

Chapter 44 | Higher-Multipole Radiation 573

- 44.1 Framing: *Suppression* 573
- 44.2 Next-order Terms 573
 - 44.2.1 Order-one terms in ϵ_{multi} can be divided into two tensor structures 573
 - 44.2.2 Antisymmetric part of the moment: magnetic dipole radiation 574
 - 44.2.3 Symmetric part of the moment: electric quadrupole radiation 575
- 44.3 Higher Order Terms can Contribute Even if Lower Ones are Zero 576
 - 44.3.1 Magnetic dipole and electric quadrupole contributions can also transport energy to infinity 576
 - 44.3.2 Qualitative approach to nuclear radiative transitions 577
- 44.4 *Plus Ultra* 578
- Problems 578

[[Vista: Transition radiation]] 580

Chapter 45 | Synchrotron Radiation 581

- 45.1 Framing: *Beaming* 581
- 45.2 The Liénard–Weichert fields 581
- 45.3 Uniform circular motion 583
 - 45.3.1 Kinematics 583
 - 45.3.2 Electric field 584
 - 45.3.3 Magnetic fields 585
 - 45.3.4 Energy flux 586
- Problems 588

xxx

[[Vista: Radiation Reaction]] 589

Chapter 46 | The Microwave Polarizer 590

- 46.1 Framing: *Jones Tensor* 590
- 46.2 Idealization as a continuous, anisotropic conducting sheet 590
- 46.3 Effect on an Arbitrarily Polarized Wave: Jones Tensor 592
- 46.4 A Tilted Polarizer can Regenerate a Missing Polarization 592
- 46.5 Extension to Optical Polarizers 592
- Problems 593

Chapter 47 | Scattering by Free and Bound Charges 594

- 47.1 Framing: *Reradiation* 594
- 47.2 Weak-Field Limit 594
- 47.3 Scattering Cross Section and the Thomson Formula 595
- 47.4 Light Propagates Diffusively in a Stellar Interior 596
- 47.5 Polarized Incoming Light Retains its Polarization Upon Scattering 597
- 47.6 Unpolarized Incoming Light Acquires Partial Polarization 597
 - 47.6.1 Selective scattering can create polarization 597
 - 47.6.2 Polarization of the cosmic microwave background as a reporter for early-Universe conditions 597
- 47.7 The Case of Bound Charges 597
 - 47.7.1 Rayleigh scattering cross section 597
 - 47.7.2 The blue, polarized sky 598
 - 47.7.3 Blue, polarized scattering from colloidal suspensions 598
- Track 2 599
 - 47.3' The transition to Compton scattering 599
- Problems 600

Chapter 48 | [[Vista: Scattering by Many Objects]] 601

[[Vista: Scattering by a Dielectric Sphere]] 602

PART VII Light in Materials

Chapter 49 | Light in Isotropic, Linear Media 604

- 49.1 Framing: *Cross-susceptibility* 604
- 49.2 Polarizable Media 604
 - 49.2.1 Induced electric dipole moment can be merged with \vec{E} , yielding the displacement field 604
 - 49.2.2 Induced magnetic dipole moment can be combined with \vec{B} to yield the \vec{H} field 605
 - 49.2.3 The Maxwell equations look simple in terms of the new fields 606
 - 49.2.4 Boundary conditions 606
- 49.3 Linear Regime 607
 - 49.3.1 Induced electric dipole moment effectively modifies ϵ_0 607
 - 49.3.2 Induced magnetic dipole moment effectively modifies μ_0 608
 - 49.3.3 The Maxwell equations then only involve free charge and charge flux 608
 - 49.3.4 Macroscopic physical realizations 609
 - 49.3.5 Remarks and further examples 609

49.4	“Total” Internal Reflection and the Evanescent Wave	610
49.5	Circular Birefringence Seems to Present a Paradox	610
49.6	Cross-susceptibility	613
49.6.1	Macroscopic physical realizations give intuition about the reality of a new effect	613
49.6.2	The general constitutive relation has new, frequency-dependent terms	614
49.7	The Origin of Circular Birefringence	615
49.8	How to Observe Circular Birefringence	616
49.9	More remarks	616
49.10	<i>Plus Ultra</i>	618
Track 2		619
49.2'	More realistic treatments of polarizable material	619
49.2.1'	Dissipation and frequency dependence	619
49.3'	More general response functions	620
49.6'a	Just two enantiomers	620
49.6'b	Relativistic formulation	620
Problems		622

[[Vista: Microscopy]] 624

Chapter 50 | Anisotropic, Linear Media 625

50.1	Framing	625
50.2	The Susceptibility Tensor Defines Principal Directions	625
50.2.1	Half-wave plate	627
50.2.2	Linear dichroism	627
50.3	Optical Torque Wrench	628
50.4	[[Induced birefringence: The quadratic electro-optic (QEO, or Kerr) effect]]	628
Track 2		629
50.2'	Magnetic anisotropy	629
Problems		630

[[Vista: Optical Solitons in Nonlinear Media]] 631

Chapter 51 | Čerenkov Radiation 632

51.1	Framing: <i>Bow Shock</i>	632
51.2	In Vacuum, a Charged Particle has one Source Point in the Observer’s Past Light Cone	632
51.3	In a Dielectric Medium, a There Can Be Two Source Points, or None, in the Observer’s Past Light Cone	632
51.4	Interpretation	633
51.5	Application: Particle Identification in Underground Detectors	633
Problems		635

Chapter 52 | [[Energy in Media]] 637

Chapter 53 | [[Vista: Photonic Bandgap Materials]] 638

53.1	Framing	638
53.2	Wishlist	638

53.2.1	Low loss	638
53.3	1D Stack	639
	Problems	640

Chapter 54 | Waves in a Cold Plasma and the Faraday Effect 642

54.1	Framing: <i>Faraday Effect</i>	642
54.2	Approximations	643
54.3	Dispersion Relation for Transverse Waves	643
54.3.1	Induced free current obeys a nondissipative response function	643
54.3.2	The dispersion relation has a cutoff	644
54.3.3	Earth's ionosphere permits the passage of cosmic messengers	645
54.3.4	Pulsar chirp results from dispersion	646
54.3.5	Metals	646
54.4	Faraday's Magneto-optical Effect	646
54.4.1	A plasma becomes a chiral medium in the presence of a steady magnetic field	646
54.4.2	A one-way light valve	648
54.4.3	The accretion disk of M87*	648
54.4.4	The Faraday effect also appears in condensed matter	648
	Problems	649

Chapter 55 | [[Vista: Metamaterials]] 651

55.1	Framing	651
	Problems	651

PART VIII *Plus Ultra*

Chapter 56 | Vista: Field Quantization 655

56.1	Framing: <i>Ex Nihil</i>	655
56.2	Maxwell Equations as Decoupled Harmonic Oscillators	656
56.3	Quantization Replaces Field Variables by Operators	658
56.4	Photon States	660
56.4.1	Basis states can be formed by applying creation operators to the vacuum state	660
56.4.2	Coherent states mimic classical states in the limit of large occupation numbers	662
56.5	Interaction with Electrons	663
56.5.1	Classical interactions involve adding source terms to the field equations	663
56.5.2	Electromagnetic interactions can be treated perturbatively	663
56.5.3	The dipole emission pattern	664
56.6	Vistas	666
56.6.1	Many invertebrates can detect the polarization of light	666
56.6.2	Invertebrate photoreceptors have a different morphology from vertebrates'	667
56.6.3	Some transitions are far more probable than others	668
56.6.4	Lasers exploit a preference for emission into an already occupied state	669

Chapter 57 | [[Vista: Einstein's Gravitation]] 670

Chapter 58 | [[Vista: Classical Yang–Mills theories]] 671

58.1 The Atiyah-Hitchin-Drinfeld-Manin instanton 671

Epilogue 672

Acknowledgments 677

Appendix A | Units and Dimensional Analysis 678

A.1 Base Units 678

A.2 Dimensions Versus Units 679

A.3 About Graphs 681

A.3.1 Arbitrary units 681

A.3.2 Angles 681

A.4 Payoff 682

Appendix B | Global List of Symbols 683

B.1 Mathematical Notation 683

B.2 Units 687

B.3 Named Quantities 687

Appendix C | Numerical Values 694

C.1 Fundamental Constants 694

C.2 Optics 694

C.2.1 Refractive index for visible light 694

C.2.2 Miscellaneous 694

Appendix D | Animated graphics 696

Appendix E | Formulas 697

Bibliography 700

Credits 711

Index 713

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, integrate, consolidate.* 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 kinds of physics.* Do problems that make contact with those fields instead of working in a hermetically sealed silo.
3. *Meditate 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). Your understanding of physics deepens when you can draw analogies to a rich tapestry of phenomena. To give

Many more phenomena
are called out in the
margins of the text.

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. One way to think about this book is as a least-action trajectory that visits all of these phenomena. Each chapter head also hints at a physical idea that will be used to understand that phenomenon. What 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.

- 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 background not given in the main text. There are also Track 2 footnotes and problems, marked with the symbol $\boxed{T2}$. They are for readers who already know the basics, so some 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), a constant of Nature. The differential operation is indicated in roman type (d) to distinguish it from any variable called d , which could denote a distance. 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), \mathbb{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 and volts as volt.
- 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.

- “Why take this course a *fourth* time?”¹ 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.²
- “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

¹Typically high school, first-year undergrad, third-year undergrad, and again now.

²For 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 470).

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.³

- “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 56 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.

“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

³See Appendix D.

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:

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 can 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.

8 July. Electromagnetics, still without success.

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.⁴ It never gets

⁴One of mine involved spinor algebra.

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. Many systems are available. Your instructor may require you to use one in particular, but if not, I have found that Python can do everything requested, and developing your skills with Python pays dividends in other physics-related areas. On the few occasions when the text mentions Python specifically, I follow the widespread convention that `plt` is short for the module `matplotlib.pyplot` and `np` is short for the module `numpy`. Concerning animation, see Appendix D.

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. [\[Not ready yet.\]](#)

Plus Ultra—

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

Prologue

If one has a chance to ask one of the older generation how they felt at the time about the writings of Maxwell, there appears something in their eyes like the glint of the love of their youth, but at the same time they betray to us 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 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.¹ Each equation is named for somebody prior to Maxwell; besides systematizing everything, Maxwell also made a crucial modification to “Ampère’s” law (Chapter 18).

$$\vec{\nabla} \cdot \vec{E} = \rho_q / \epsilon_0 \quad \text{electric Gauss} \quad (0.1)$$

$$\vec{\nabla} \cdot \vec{B} = 0 \quad \text{magnetic Gauss} \quad (0.2)$$

$$\vec{\nabla} \times \vec{E} + \frac{\partial}{\partial t} \vec{B} = \vec{0} \quad \text{Faraday} \quad (0.3)$$

$$\vec{\nabla} \times \vec{B} - \mu_0 \epsilon_0 \frac{\partial}{\partial t} \vec{E} = \mu_0 \vec{J}. \quad \text{Ampère} \quad (0.4)$$

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

¹Even Einstein’s original relativity paper used different names for each cartesian component, today considered horrible.

The constants have numerical values $\mu_0 \approx 4\pi \cdot 10^{-7} \text{ m kg coul}^{-2}$ (the **magnetic permeability of vacuum**), and $\epsilon_0 \approx 8.85 \cdot 10^{-12} \text{ coul}^2 \text{ N}^{-1} \text{ m}^{-2}$ (the **electric permittivity of vacuum**).²

Later chapters will define the **charge density** ρ_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**. Some formulas are neater when expressed in terms of a 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 motions of an isolated, charged point under the influence of external fields:

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

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. \vec{f}_{other} represents any non-electromagnetic force acting on the charged bodies in the system.⁴ And the momentum $\vec{p}(t) = m\vec{v}(t)$, at least for speeds much smaller⁵ than 10^8 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.

[T2] Section 0.1.2’ (page 14) 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 $1/\epsilon_0$.”
- The magnetic Gauss law says, “No point sources nor sinks for magnetic fields.”
- The Faraday law says, “Time-dependent magnetic fields themselves also give rise to electric fields.”

²Chapter 16 will discuss units in greater detail, and explain why the value of μ_0 stopped being exact, and became only approximate, in 2019.

³Section 8.3 will define ρ_q and \vec{j} carefully. Some books call \vec{j} the “current density,” but that term risks confusion; we will use “density” to mean *volume* density (m^{-3}), or “areal density” to mean per area.

⁴Sometimes 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.

⁵Chapter 31 will generalize this relation.

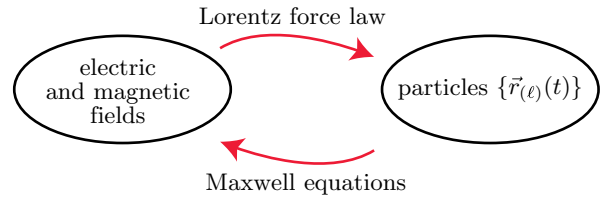


Figure 0.1: Reciprocal roles of the Maxwell field equations and the Lorentz force law.

- The Ampère law says, “Currents give rise to magnetic fields with some constant of proportionality μ_0 . Time-dependent electric fields themselves 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 system⁶) and quoting its components:

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

That is, the symbol \vec{r} can either represent an abstract geometric object (an arrow), or it can represent a set of three numbers, called $\vec{r}_1 = x$, $\vec{r}_2 = y$, and $\vec{r}_3 = z$, regarded as a column (3×1 matrix). Note that a subscript on a 3-vector indicates that only one of its components (an ordinary number) is meant. Again: The over-arrow notation implies that we mean specifically cartesian coordinates. Most authors drop the over-arrow when explicitly writing the index on a vector, but in this book we will retain it for clarity: \vec{r}_1 is a single number, but it’s not a scalar; it is a component of a vector.⁷ We won’t ever use the 3-vector notation \vec{r}^i (upper index).⁸

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

⁶Thus, 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 Maxwell’s equations look nice, and we’re studying Maxwell’s equations.” Section 5.6 will consider how the representation of a vector changes when we switch from one “good” system to another, or to a less “good” system.

⁷Later, 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 3-vectors is meant (see for example Section 1.2 later).

⁸Such 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 $\|\vec{r}\|^2$, \vec{r}^2 , or simply r^2 ; so $r \equiv \sqrt{\vec{r}^2}$. Section 14.4 (page 205) reviews why $\vec{a} \cdot \vec{b} = \|\vec{a}\| \|\vec{b}\| \cos \theta$, where θ 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 $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, and $\hat{\mathbf{k}}$).

The components of a vector *field*, such as $\{\vec{E}_i\}$, are themselves functions on space-time, that is, $\vec{E}_i(t, \vec{r})$ and so on. We differentiate them with the vector of operators⁹

$$\vec{\nabla}_i = \begin{bmatrix} \partial/\partial x \\ \partial/\partial y \\ \partial/\partial z \end{bmatrix}_i.$$

The dot product of $\vec{\nabla}$ with itself is the **Laplace operator** (or **laplacian**), written as¹⁰ ∇^2 .

The dot product of $\vec{\nabla}$ acting on a vector field is called the **divergence operator** and denoted $\vec{\nabla} \cdot \vec{V}$. Note that $\vec{\nabla} \cdot \vec{V}$ is an ordinary function, whereas $\vec{V} \cdot \vec{\nabla}$ is an *operator* that acts on whatever sits to its right and does not involve any derivatives of \vec{V} . In fact, $(\vec{V} \cdot \vec{\nabla})f$ 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 $\vec{v} = d\vec{r}/dt$.

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} , 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 human¹¹), and use it to classify coordinate systems as “right handed” or not.

⁹Again, 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.

¹⁰Mathematicians use the symbol Δ for the laplacian, but physicists don’t. It’s too easy to confuse that with Δ , the physicists’ symbol for a change in some quantity.

¹¹Less 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.

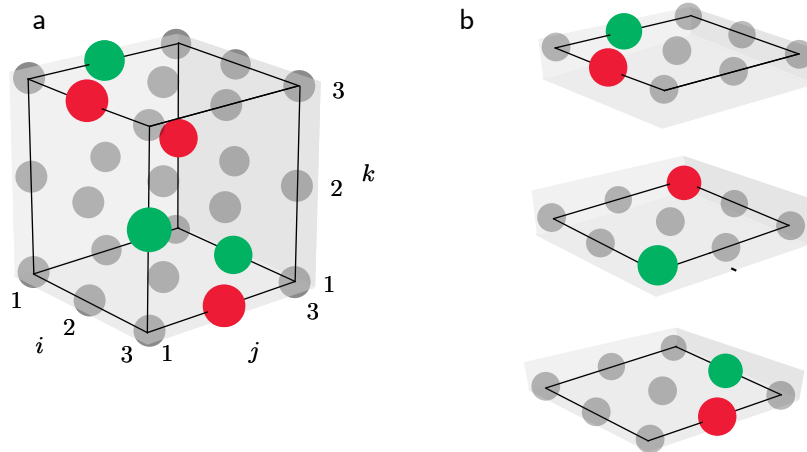
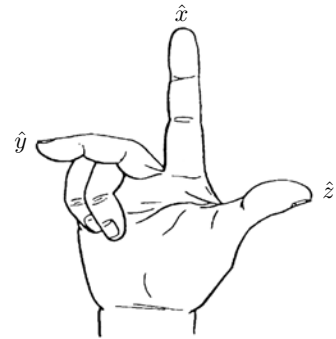


Figure 0.3: Structure of the Levi-Civita symbol. (a) The 27 numbers that make up the Levi-Civita symbol ε_{ijk} , represented as a stack of balls. Three entries are +1 (*green balls*), three are -1 (*red balls*), and 21 are zero (*transparent gray balls*). (b) Exploded view detailing individual layers.

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 θ 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 \vec{a}, \vec{b} , and \hat{c} form a right-handed triad in the sense of Figure 0.2.¹²

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

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

The formula involves the **3D Levi-Civita symbol** ε_{ijk} , which is shorthand for 27 num-

¹²If \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.

bers ($3 \times 3 \times 3$). Most of those entries equal zero: $\varepsilon_{ijk} = 0$ if any two of the indices match, for example, ε_{112} . If all three indices have different values, then they must be a permutation of 1, 2, 3; ε_{ijk} is then defined using the parity of that permutation. Thus, $\varepsilon_{123} = +1$, $\varepsilon_{231} = +1$, $\varepsilon_{132} = -1$ and so on (Figure 0.3).

The entries \vec{a}_j and \vec{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 geometric definition.¹³ But you can readily generate some evidence:

Your Turn 0A

- Use Equation 0.7 to show that $\vec{a} \times \vec{a} = 0$ for any vector, in agreement with the geometric definition.
- 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 \vec{a} will have components with $\vec{a}'_1 = \vec{a}_1$, $\vec{a}'_2 = \vec{a}_2$, and $\vec{a}'_3 = -\vec{a}_3$. Writing \times' for the alternate version, we find $(\vec{a} \times' \vec{b})'_3 = \vec{a}'_1 \vec{b}'_2 - \vec{a}'_2 \vec{b}'_1$ and so on. Are these the primed components of the vector $\vec{a} \times \vec{b}$ 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 more than three-dimensional spaces.¹⁴

The cross product of $\vec{\nabla}$ acting on a vector field \vec{V} is a new vector field called the **curl** of \vec{V} , denoted $\vec{\nabla} \times \vec{V}$. It enters in the Faraday and Ampère laws.

0.2.3 The Kronecker symbol

There’s also the more familiar **Kronecker symbol** δ_{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 $\vec{v}(\vec{r})$.¹⁵ We can ask about the **streamlines** of this vector field. The streamlines are curves in space that are everywhere tangent to \vec{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

¹³See Section 14.4.2.

¹⁴And even to curved spaces.

¹⁵We could define average velocity as the flux of mass divided by mass density.

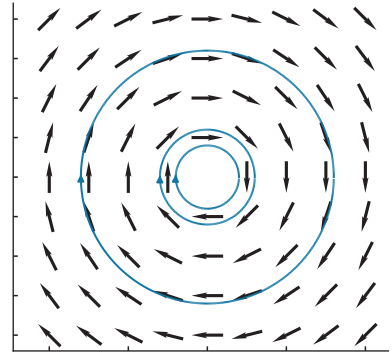


Figure 0.4: [Mathematical functions.] **Streamlines of a vector field.** Three typical streamlines (*curves*) are shown for the vector field with $\vec{v}_1 = y/r$, $\vec{v}_2 = -x/r$, where $r = (x^2 + y^2)^{1/2}$.

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**. Although we no longer think of them as physical objects like rubber bands, nevertheless Chapter 36 will explain why that imagery was useful to Michael Faraday.

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 $(\vec{a} \times \vec{b})_i = \varepsilon_{ijk} \vec{a}_j \vec{b}_k$. 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 $\varepsilon_{ijk} \vec{a}_j \vec{b}_k + \varepsilon_{imn} \vec{c}_m \vec{d}_n$, 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 *all* values 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 $\vec{a}_i \vec{b}_i$ times $\vec{c}_i \vec{d}_i$ should be rewritten $\vec{a}_i \vec{b}_i \vec{c}_j \vec{d}_j$ (or $(\vec{a} \cdot \vec{b})(\vec{c} \cdot \vec{d})$).

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

$$\int_V d^3r \vec{\nabla} \cdot \vec{E} = \int_{\partial V} d^2\vec{\Sigma} \cdot \vec{E}. \quad (0.8)$$

Here d^3r is a volume element.

V is a finite volume and ∂V denotes the closed surface bounding it. Thus, ∂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 ∂V , called $d^2\Sigma$, has two perpendicular directions (sometimes called **normal vectors**). The surface separates space into “inside” and “outside,” so one of the normals is the “outward-pointing normal.” On the right side of Equation 0.8, the vector $d^2\vec{\Sigma}$ denotes the product of an area element $d^2\Sigma$ times the unit outward-pointing perpendicular vector.

0.3.4 Stokes theorem

$$\int_{\Sigma} d^2\vec{\Sigma} \cdot (\vec{\nabla} \times \vec{E}) = \oint_{\partial\Sigma} d\vec{\ell} \cdot \vec{E}. \quad (0.9)$$

Here $d\vec{\ell}$ is a vector line element. Σ 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 (endpoints). An open patch of surface has no “inside/outside” distinction, so we may choose either face as “outward” when defining the sign of $d^2\vec{\Sigma}$. 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.¹⁶

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

Your Turn 0B

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

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

0.3.5 Two useful lemmas

Your Turn 0C

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

- $\vec{\nabla} \times \vec{V} = \hat{r} \times d\vec{V}_i/dr$, and
- $\hat{r} \cdot (\vec{\nabla} \times (\vec{r} \times \vec{V})) = -2\hat{r} \cdot \vec{V}$.

¹⁶Point the thumb of your right hand along the chosen normal, then traverse the boundary in the sense that follows the curve of your fingers.

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 $\frac{1}{2}[\bar{b}e^{-i\omega t} + \text{c.c.}]$, where “c.c.” stands for “complex conjugate.” If we choose $\bar{b} = 1$, then this expression equals $\cos(\omega t)$; if we choose $\bar{b} = i$, then it equals $\sin(\omega t)$. If \bar{b} is complex, we may write it as $|b|e^{i\alpha}$; then

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

which still has frequency ω but is phase shifted relative to sine or cosine.

0.3.7 Angle and solid angle

A short line element $d\vec{\ell}$, seen from a great distance, subtends an angle $d\theta \rightarrow \|d\vec{\ell} \times \hat{r}\|/r$, 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 $d^2\vec{\Sigma}$, seen from a distance, subtends a **solid angle**¹⁷ $d\Omega \rightarrow d^2\vec{\Sigma} \cdot \hat{r}/r^2$. 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 **msr** = 0.001 or **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 \mathbb{L}), ρ (radius, or distance away from the cylinder axis, dimension \mathbb{L}), and φ (**azimuthal angle** in the plane perpendicular to cylinder axis, dimensionless).
- Spherical polar coordinates, consisting of r (radius, or distance away from the origin, dimension \mathbb{L}), θ (**polar angle**, tilt downward from the $+\hat{z}$ axis, dimensionless), and ϕ (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**¹⁸ is not really a 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 .

¹⁷A better name for this quantity might be **angular area**.

¹⁸Sometimes called “Dirac delta function.”

For our purposes, it will usually suffice to regard $\delta(x)$ in a sloppy way as the limit of a bump function, for example $e^{-x^2/(2\sigma^2)}/(\sqrt{2\pi}\sigma)$, which becomes sharply peaked as $\sigma \rightarrow 0$, with the area under the curve fixed to 1. That viewpoint also 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.

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 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.¹⁹

One point that everybody *could* understand,²⁰ 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).

¹⁹A good lesson: We old 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.

²⁰For example, Heaviside had already noted this disconnect in 1885.

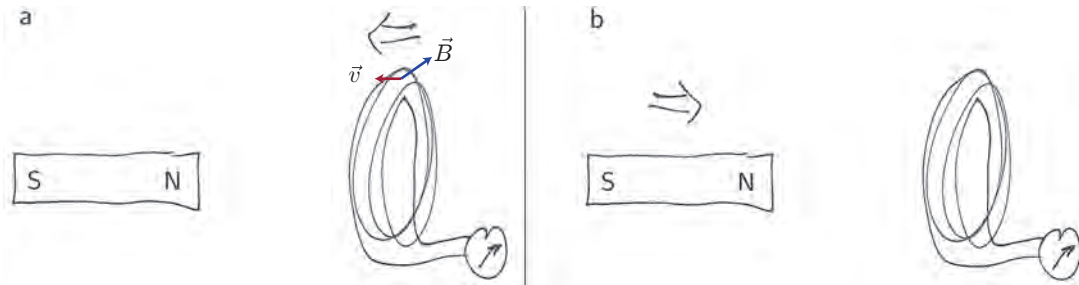


Figure 0.5: Two experiments with the same apparatus. (a) A coil of wire is pushed over a stationary permanent magnet. The turns of the coil lie in a plane coming out of the page, but its velocity \vec{v} as it approaches the magnet lies in the plane. At a point where the wires pass through the plane of the page, the static magnetic field \vec{B} also lies in that plane as shown. (b) A permanent magnet is plunged into a stationary coil of wire.

- 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 \vec{B} , 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, so there is no magnetic force. In this case, however, the \vec{B} field is time-dependent. Faraday’s law (Equation 0.3) then implies an \vec{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.²¹

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.”

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

²¹Actually, 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.

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 abandoned.

This was one of the most amazing examples of (successful) lateral thinking in the history of science,²² 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?

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 $\{\vec{E}_i, \vec{B}_i\}$?

Hanging #E: Our equations are full of cross products, which depend on an arbitrary choice of which is our “right” hand.²³ 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:

Schey, 2005.

²²For another, see Section 21.6 (page 306).

²³Section 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 $\vec{r}(t)$ (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 be pushed by a magnetic field gradient, and so on. [\[Not ready yet.\]](#)

PROBLEMS

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):

$\alpha, \beta, \gamma/\Gamma, \delta/\Delta, \epsilon, \zeta, \eta, \theta/\Theta, \kappa, \lambda/\Lambda, \mu, \nu, \xi/\Xi, \pi/\Pi, \rho, \sigma/\Sigma, \tau, \upsilon/\Upsilon,$

ϕ (sometimes written φ)/ $\Phi, \chi, \psi/\Psi, \omega/\Omega.$

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 φ for phi and ϑ 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 $\Theta\epsilon\delta\varsigma\ \alpha\epsilon\iota\ \gamma\epsilon\omega\mu\epsilon\tau\rho\epsilon\hat{\iota}$.” From the sounds made by each letter, can you guess what Thompson was trying to say? [*Hint*: ς is an alternate form of σ .]

0.2 *By any other name*

Fundamental constants can be expressed in whatever units are convenient for the problem at hand. Express the constant $e^2/(4\pi\epsilon_0)$ in the units MeV fm convenient for nuclear physics.

CHAPTER 1

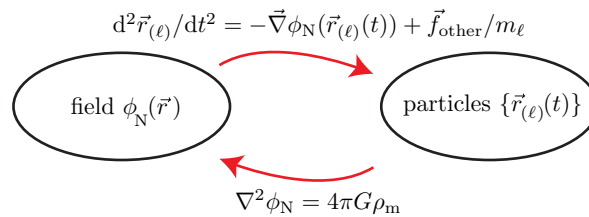
Warmup: Newtonian Gravitation

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 course, but it's useful to introduce some themes with a scalar field theory before we move up to a vector field theory.



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.

Electromagnetic phenomenon: The same hyperbolic trajectories taken by comets can also be relevant when charged particles are fired into matter.

Physical idea: The same equations will have the same solutions in any physical context.

1.2 SPACE CARRIES A PHYSICAL FUNCTION CALLED THE NEWTONIAN POTENTIAL

The **newtonian potential** ϕ_N is a function that obeys

$$\nabla^2\phi_N = 4\pi G_N\rho_m. \quad (1.1)$$

Here G_N is a universal constant of Nature and ρ_m is the mass density of matter.

We can think of matter as a collection of N point masses m_{ℓ} following trajectories $\vec{r}_{(\ell)}(t)$. Here the mass m_{ℓ} is a constant characterizing particle number ℓ . In the notation $\vec{r}_{(\ell)}$, the particle number ℓ appears in parentheses to avoid confusing it with a vector index labeling which component we’re discussing; the vector index has been suppressed. (If we want a particular component we can write $\vec{r}_{(\ell)i}$.)

With that notation understood, then we can finish specifying Equation 1.1 by constructing the mass density distribution as

$$\rho_{\text{m}}(t, \vec{r}) = \sum_{\ell} m_{\ell} \delta^{(3)}(\vec{r} - \vec{r}_{(\ell)}(t)). \quad (1.2)$$

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 ρ_{m} .
- The $3N$ functions of time, $\vec{r}_{(\ell)}(t)$, specify the N particle trajectories as functions of time t .

Often it's a good approximation to blur the many delta functions together. Then ρ_{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:

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

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, ϕ_{N} is potential energy of a test particle *per unit mass*.

The term \vec{f}_{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 *really* 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 vortices, or 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 ϕ_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 $\vec{\nabla}(\frac{1}{r})$, 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 $\vec{\nabla}$, 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

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

a familiar result. Here $\hat{r} = \vec{r}/r$ is the unit vector pointing to \vec{r} .

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

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

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

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

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

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:¹

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

So the integral of $\nabla^2(-G_N M/r)$ over any spherical volume containing the origin is always $4\pi G_N M$, even though $\nabla^2(-G_N M/r) = 0$ 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 familiar newtonian potential (which gives rise to the familiar newtonian force) really does solve Equation 1.1 for a point mass.

¹See 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.

1.4 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):

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

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 ω is the angular frequency. Taking the value of ω that corresponds to a sidereal month, and r to be the Earth–Moon distance, and substituting into Equation 1.3 gives a rough² 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 ϕ_N :

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

Later we'll similarly exploit the linearity of Maxwell's equations, for a similar win. The fundamental solution that must be integrated against the mass density is called the **Green function** for whatever field equation we are studying. 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}_*

²See Problem 1.1. It's reasonable to suppose the Earth stationary, because the Moon's mass is much less than Earth's. Newton did better by using the “reduced mass.”

approaches the field 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 ϕ_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

$$\rho_m \int_0^{\text{small}} (r_*^2 dr_* d(\cos \theta) d\varphi) r_*^{-1}.$$

But the singularity is more than compensated by the r_*^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.³

1.6 PLUS ULTRA

1. Why introduce the potential function? Why not just work directly with the forces? One huge 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 successors⁴ 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.⁵

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 natural laws, not the whims of a supernatural being, then the divine rights of capricious kings looked absurd. But that is another story.

3. 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. Later developments showed that even the very coordinates $\vec{r} = (x, y, z)$ and t require careful interpretation.⁶

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 to a solution of the equations.

³Although we have disposed of the singularity issue in principle, in practice it can reappear when we attempt numerical evaluations; see Problem 1.4.

⁴Notably d'Alembert, Clairaut, Euler and Laplace.

⁵Chapter 57 will follow Einstein's steps 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 handout is still a good way to think about more advanced theories.

⁶Chapter 26 and following chapters will look more closely on coordinate choice.

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 apparatus that we might wheel into the lab,⁷ any initial conditions we may set on that apparatus, 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 , or 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.*

1.7 MORE HANGING QUESTIONS

Hanging #F: 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?

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

⁷Henry Cavendish designed a gravitational experiment that fits in a room.

PROBLEMS

1.1 *The first grand unification*

Repeat Newton's early unification: 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.) Next look up the Earth's radius and again estimate $G_N M_{\text{earth}}$, this time based on the terrestrial acceleration of gravity. 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.⁸ 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 $t \rightarrow -\infty$. Set up polar coordinates centered on M , in which φ is measured clockwise from the $-\hat{x}$ axis. Thus, the incoming body starts with $\varphi = 0$, and φ increases with time. If M were not present, then the trajectory would have $\varphi \rightarrow \pi$ at $t \rightarrow +\infty$.

- 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 $t \rightarrow -\infty$, we have $u \rightarrow 0$. 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 φ at which $u(\varphi)$ stops increasing and turns around. Double this angle to find the total angular 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.

⁸See 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.

Choose axes such that at some moment, the stars are both on the x axis at $x = \pm a$. 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 \hat{z} parallel to the long edge, \hat{x}, \hat{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 ϕ_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.⁹ (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 `plt.contour` function or something similar.]

1.5 Below the surface

Earth's gravitational acceleration decreases as we descend a deep mineshaft.

Section 1.5 (page 19) asserted that the apparent singularity in Equation 1.6 (page 19) 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 ρ_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.

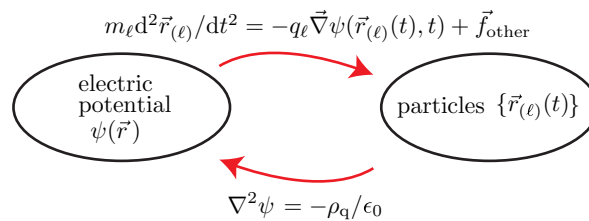
⁹Certainly there are more sophisticated ways to do an integral numerically, and you're free to do that if you prefer.

CHAPTER 2

Electrostatics Introduced

2.1 FRAMING: COEQUAL PARTNERS

Maxwell’s 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:



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 . Each sets up a corresponding density: ρ_m (as before) and ρ_q respectively.

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 **static**, that is, invariant under both time shift and 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 “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.

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

²For the first part, see Problem 2.2.

³In contrast, current flowing steadily through a wire is invariant under time shifts but not under time reversal—that’s called **stationary**, not static.

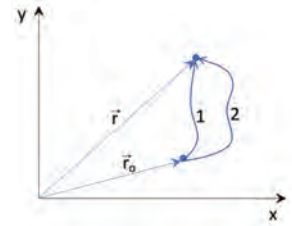


Figure 2.1: Path-independence of Equation 2.2.

- 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 will still be static, however, because the *average* velocity of charges in any region will still be 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 $\vec{j} = 0$. The static condition also implies that $d\vec{E}/dt = 0$, so Equations 0.2 and 0.4 imply that there are no magnetic fields ($\vec{B} = 0$), and all we have left of Maxwell are Equations 0.1 and 0.3:

$$\vec{\nabla} \cdot \vec{E} = \rho_q/\epsilon_0, \quad \vec{\nabla} \times \vec{E} = \vec{0}. \quad (2.1)$$

Here ρ_q is electric charge density, analogous to Equation 1.2 (page 17), and ϵ_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 \vec{E} is force *per charge*, ϵ_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” \vec{r}_0 in space and define the **electrostatic potential** as the scalar function

$$\psi(\vec{r}) = - \int_{\vec{r}_0}^{\vec{r}} d\vec{r}' \cdot \vec{E}(\vec{r}'). \quad (2.2)$$

Here the notation denotes the line integral along any path that starts at the reference point \vec{r}_0 and ends at the **field point** \vec{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 ψ by the integral $-\oint d\vec{r}' \cdot \vec{E}$ around that closed loop (Figure 2.1). By Stokes's theorem, this can be written as a surface integral of $\vec{\nabla} \times \vec{E}$ (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, ψ 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 in a trivial way: Changing \vec{r}_0 just adds a *constant* to ψ . We don't explicitly indicate the dependence on \vec{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 ψ contains the same information as the vector field \vec{E} . To prove that, let's evaluate the gradient of ψ . For example, we'll find $\psi(\vec{r} + \epsilon\hat{x}) - \psi(\vec{r})$. We can evaluate Equation 2.2 using any path we like, so choose any path from the origin to \vec{r} , and another that follows the first one but then moves from \vec{r} parallel to the x axis a distance ϵ . Both integrals are the same and cancel, except for the last bit of the first one, which contributes $\epsilon\vec{E}_1$. We conclude that $-\vec{\nabla}_1\psi(\vec{r}) = \vec{E}(\vec{r})$, a generalization of the Fundamental Theorem of Calculus. More generally,

$$\vec{E} = -\vec{\nabla}\psi. \quad (2.3)$$

Then the first of Equations 2.1 becomes the **Poisson equation**:

$$\nabla^2\psi = -\rho_q/\epsilon_0. \quad (2.4)$$

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 becomes:

$$\frac{d}{dt}\vec{p}_{(\ell)} = -q_\ell\vec{\nabla}\psi(\vec{r}_{(\ell)}(t)). \quad (2.5)$$

Here q_ℓ is the electric charge, a fixed quantity that is attached to particle ℓ .

The electrostatic potential $\psi(\vec{r})$ is the potential energy *per unit charge* of a test body located at \vec{r} . Its units are therefore joules per coulomb, which is the definition of

“volt.” Most authors abbreviate this unit “V,” but that could lead to confusion with volume or something, so we will write volt.

The electric field $-\vec{\nabla}\psi$ 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 ψ (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 DIFFERENCES FROM GRAVITATION

There is an obvious big difference between newtonian gravity and electrostatics: The mass density ρ_m must always be nonnegative (everything attracts everything), but charge density ρ_q need not be nonnegative (some pairs of things attract but others repel). (Placement of the 4π 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 ϵ_0 .)

2.4 BASIC SOLUTIONS

2.4.1 Point charge

One solution of the Poisson equation is the one we found in gravitation: A point charge of strength q located at the origin gives $\psi(\vec{r}) = q/(4\pi\epsilon_0 r)$, or more generally⁴

$$\psi(\vec{r}) = \frac{q}{4\pi\epsilon_0 \|\vec{r} - \vec{r}_*\|}$$

if the charge is located at \vec{r}_* .

Your Turn 2A

Find the negative gradient of this function,⁵ then go back via Equation 2.2 to see how it all fits together.

⁴The 4π had to pop up somewhere! We banished it from the Poisson equation, so it appears here.

⁵H. 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.

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.

2.4.2 Continuous charge distribution

The Poisson equation is linear in ψ , so we can quickly generalize our point-charge solution to the case of a continuous distribution with charge density $\rho_q(\vec{r}_*)$. Simply subdivide charge into small elements $dq = \rho_q(\vec{r}_*)d^3r_*$ and add up their contributions. We'll call \vec{r}_* 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:

$$\psi(\vec{r}) = \int d^3r_* \frac{\rho_q(\vec{r}_*)}{4\pi\epsilon_0\|\vec{r} - \vec{r}_*\|}. \quad (2.6)$$

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

We should address a possible objection to Equation 2.6. Suppose that we wish to know the potential at a field point somewhere inside the distribution, that is, a point where $\rho_q(\vec{r}) \neq 0$. The expression in Equation 2.6 seems to involve $1/0$ when $\vec{r}_* = \vec{r}$! But consider the integrand close to that point. Let $\vec{R} = \vec{r} - \vec{r}_*$. Then the suspicious part of the integral is d^3r/R , times the smooth function $\rho_q(\vec{r} - \vec{R})$. And $d^3r/R = RdRd\varphi d(\cos\theta)$ presents no problems near $R \rightarrow 0$.

2.5 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.⁶

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 ψ 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.

$$\vec{E}_{\parallel} = 0. \quad \text{just outside a conductor, static} \quad (2.7)$$

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

⁶“Eventually” because charges may rearrange slowly, due to friction.

Often we don't even need to know the surface charge distribution. But if we do, we can find it by computing $\epsilon_0 \vec{\nabla}_\perp \psi$ once we have solved the boundary-value problem for the potential.

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).

2.6 UPCOMING

2.6.1 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.

2.6.2 Quasi-static

We'll see in Section 8.6 that many situations of interest are not precisely static, but may nevertheless be regarded as such because charges are moving slowly.

2.6.3 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 version of the potential that applies in this case as well. It won't have any interpretation as potential energy per unit charge, but nevertheless it will still be called a “potential.” Sorry for that misleading, but standard, terminology.

2.2' Falsifiable content of the equations

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. 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:

- α 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_ℓ/m_ℓ characterizing each test body ℓ (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.*
- β If the apparatus consists of charges executing specified motions, then the functions \vec{E} and \vec{B} , measured as described in (α) above, *are not arbitrary, but are solutions to the partial differential Equations 0.1–0.4* with sources determined by the charges.
- γ If the apparatus consists of point charges which are themselves free (other than being influenced by EM fields and known forces \vec{f}_{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 34.8 (page 461)).

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 21 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.

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 Einstein found illuminating, however, was the transformations *between* the presumed good systems. Chapters 28–30 will describe how they were not what everybody had expected.

PROBLEMS

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.

- What is the total charge q_{tot} ?
- 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 .

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

- 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.
- 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} m .) What is the total deflection?
- 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).

Energetic protons lose energy in matter even without direct collisions.

Your formula involved the quantity

$$Y \equiv K/(m_e A v_0^2),$$

where $K = e^2/(4\pi\epsilon_0)$, 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.

- 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.

- When a proton flies through a gas of many electrons, all initially at rest⁷, 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, c_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 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

⁷You may neglect screening in this problem.

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_{\max} at which W equals the ionization energy I of a molecule. Find a formula for A_{\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 , c_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 $K = e^2/(4\pi\epsilon_0) = 1.4 \text{ eV nm}$.
 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 c_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

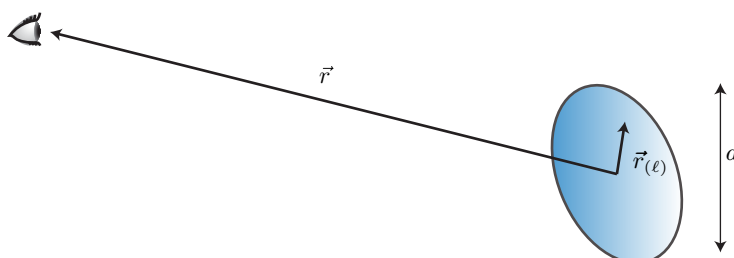
3.1 FRAMING: *DISTILLATION*

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 H_2O or CO_2 . 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.

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 FORMULA

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

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

We'll choose a reference point somewhere inside that region and use it as an origin of coordinates; thus, charge $\# \ell$ sits at a position $\vec{r}(\ell)$ with $r(\ell) \lesssim a \ll r$. The goal is to

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.

show that the electrostatic potential at \vec{r} can be expanded as

$$\psi(\vec{r}) = q_{\text{tot}}\psi^{[0]}(\vec{r}) + \vec{\mathcal{D}}_{\text{E}} \cdot \vec{\psi}^{[1]}(\vec{r}) + \sum_{ij} \left[\vec{\mathcal{Q}}_{\text{E},ij} \vec{\psi}_{ij}^{[2]}(\vec{r}) \right] + \dots \quad (3.2)$$

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

q_{tot} is a scalar constant called **electric monopole moment** or **zeroth moment** of charge. The three constants $\vec{\mathcal{D}}_{\text{E}}$ form a vector called **electric dipole moment** or **first moment** of charge.² The constants $\vec{\mathcal{Q}}_{\text{E},ij}$ 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 $\vec{\mathcal{Q}}_{\text{E}}$ as a 3×3 matrix.

The moments just mentioned are defined by³

$$q_{\text{tot}} = \sum_{\ell} q_{\ell}, \quad \vec{\mathcal{D}}_{\text{E},i} = \sum_{\ell} q_{\ell} \vec{r}_{(\ell)i}, \quad \vec{\mathcal{Q}}_{\text{E},ij} = \sum_{\ell} q_{\ell} (3\vec{r}_{(\ell)i}\vec{r}_{(\ell)j} - \|\vec{r}_{(\ell)}\|^2 \delta_{ij}). \quad (3.3)$$

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 $\vec{\mathcal{Q}}_{\text{E},ij}$, regarded as a matrix, is always symmetric and traceless. More explicitly, the first term of $\vec{\mathcal{Q}}_{\text{E}}$ is the sum over charges of

$$3 \begin{bmatrix} x_{\ell}^2 & x_{\ell}y_{\ell} & x_{\ell}z_{\ell} \\ y_{\ell}x_{\ell} & y_{\ell}^2 & y_{\ell}z_{\ell} \\ z_{\ell}x_{\ell} & z_{\ell}y_{\ell} & z_{\ell}^2 \end{bmatrix},$$

weighted by electric charge. The matrix just given is symmetric. Its trace (sum of diagonal entries) is $3 \sum q_{\ell} (x_{\ell}^2 + y_{\ell}^2 + z_{\ell}^2)$. In the second term of $\vec{\mathcal{Q}}_{\text{E}}$, we get the symmetric matrix δ_{ij} , whose trace is 3, times $-\sum q_{\ell} \|\vec{r}_{\ell}\|^2$. When combined, these terms form a symmetric matrix whose trace equals zero.

For a continuous charge distribution, we have analogously

$$q_{\text{tot}} = \int d^3r_* \rho_{\text{q}}(\vec{r}_*), \quad \vec{\mathcal{D}}_{\text{E}} = \int d^3r_* \rho_{\text{q}}(\vec{r}_*) \vec{r}_*,$$

the **zeroth** and **first moments** of the charge distribution with respect to the chosen reference point, and similarly for $\vec{\mathcal{Q}}_{\text{E}}$.

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

$$\psi^{[0]}(\vec{r}) = \frac{1}{4\pi\epsilon_0 r}; \quad \vec{\psi}_i^{[1]}(\vec{r}) = \frac{1}{4\pi\epsilon_0 r^2} \hat{r}_i; \quad \vec{\psi}_{ij}^{[2]}(\vec{r}) = \frac{1}{8\pi\epsilon_0 r^3} (\hat{r}_i \hat{r}_j - \frac{1}{3} \delta_{ij}). \quad (3.4)$$

²Chapters 13–14 and 32–34

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

⁴The δ_{ij} terms in Equations 3.3 and 3.4 are redundant: Omitting either (but not both) leaves ψ 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. 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 $(1 + \epsilon)^{\pm 1/2}$ near $\epsilon = 0$:

$$\begin{aligned}\sqrt{1 + \epsilon} &= 1 + \frac{1}{2}\epsilon - \frac{1}{8}\epsilon^2 + \dots \\ 1/\sqrt{1 + \epsilon} &= 1 - \frac{1}{2}\epsilon + \frac{3}{8}\epsilon^2 + \dots\end{aligned}\quad (3.5)$$

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 ϵ must be, and so on.

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

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

$$1/\sqrt{1 + \delta + A\delta^2} = 1 - \frac{1}{2}(\delta + A\delta^2) + \frac{3}{8}(\delta + A\delta^2)^2 + \dots \quad (3.6)$$

$$= 1 - \frac{1}{2}\delta + \delta^2\left(-\frac{1}{2}A + \frac{3}{8}\right) + \mathcal{O}(\delta^3). \quad (3.7)$$

Note that:

- We chose to stop the expansion at some fixed order in δ (here second order).
- Part of the term that was first order in ϵ in Equation 3.6 has entered into the term that is second order in δ in Equation 3.7.
- Some but not all of the order ϵ^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 ϵ^3 or higher, because anything contained in such a term would be at least order δ^3 .

3.4 DERIVATION OF THE FORMULA

Now that we have unpacked the 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:

$$\begin{aligned}\psi(\vec{r}) &= \sum_{\ell} \frac{q_{\ell}}{4\pi\epsilon_0} (\|\vec{r} - \vec{r}_{(\ell)}\|^{-1/2}) = \sum_{\ell} \frac{q_{\ell}}{4\pi\epsilon_0} (r^2)^{-1/2} \left(\left(\frac{\vec{r}}{r} - \frac{\vec{r}_{(\ell)}}{r} \right)^2 \right)^{-1/2} \\ &= \frac{1}{4\pi\epsilon_0 r} \sum_{\ell} q_{\ell} \left(\hat{r}^2 - 2 \frac{\vec{r} \cdot \vec{r}_{(\ell)}}{r^2} + \frac{(\vec{r}_{(\ell)})^2}{r^2} \right)^{-1/2} \\ &= \frac{1}{4\pi\epsilon_0 r} \sum_{\ell} q_{\ell} \left(1 - \frac{1}{2} \left(-2 \frac{\vec{r} \cdot \vec{r}_{(\ell)}}{r^2} + \frac{r_{(\ell)}^2}{r^2} \right) + \frac{3}{8} \left(-2 \frac{\vec{r} \cdot \vec{r}_{(\ell)}}{r^2} + \dots \right)^2 + \dots \right).\end{aligned}$$

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 r^{-1} , keeping up through r^{-3} :

$$\begin{aligned}&= \frac{1}{4\pi\epsilon_0 r} \sum_{\ell} q_{\ell} \left(1 + \frac{\vec{r} \cdot \vec{r}_{(\ell)}}{r^2} + \frac{1}{r^2} \left(-\frac{1}{2} r_{(\ell)}^2 + \frac{3}{2r^2} (\vec{r} \cdot \vec{r}_{(\ell)})^2 \right) + \dots \right) \\ &= \frac{1}{4\pi\epsilon_0} \sum_{\ell} \left(\frac{q_{\ell}}{r} + \frac{q_{\ell} \hat{r} \cdot \vec{r}_{(\ell)}}{r^2} + \frac{q_{\ell}}{2r^3} \hat{r}_i \hat{r}_j (3\vec{r}_{(\ell)i} \vec{r}_{(\ell)j} - r_{(\ell)}^2 \delta_{ij}) + \dots \right).\end{aligned}$$

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 $1/(8\pi\epsilon_0 r^3)$ times

$$-\frac{1}{3} \vec{\mathcal{Q}}_{\mathbf{E},ij} \delta_{ij} = -\frac{1}{3} \sum_{\ell} q_{\ell} (3\vec{r}_{(\ell)i} \vec{r}_{(\ell)j} - r_{(\ell)}^2 \delta_{ij}) \delta_{ij} = 0.$$

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 the sum of products of:

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

The “multipole fields” $\psi^{[p]}$ have nothing to do with the source object—they just catalog possible solutions of the Laplace equation. The moments have nothing to do with observer position \vec{r} —they just 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 this approach 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 have found that:

- If a static, localized charge distribution has any term in its potential that falls as $1/r$, that contribution to ψ must be spherically symmetric.
- If ψ has any term of order $1/r^2$, then 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}_E .
- 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}_E , which we called a “tensor.”⁵ Later chapters will generalize this notion.

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 $\vec{\psi}^{[1]}$ and $\vec{\psi}^{[2]}$ appearing in Equation 3.4. Note how the units work out.

3.6.3 Apparent singularity

Every term in the multipole expansion of ψ is singular at $r = 0$. The corresponding singularities in the electric field are worse still. Is that a problem? No: The expansion is a power series in a/r , so it breaks down (becomes inaccurate) at $r \rightarrow 0$. (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.

⁵Chapter 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.”

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

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

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

Your Turn 3C

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

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

Your Turn 3D

- So can we forget about electric dipole fields? Why or why not?
- Can we use a similar argument to eliminate $\vec{\mathcal{Q}}_E$?

3.6.5 Spherical distributions

Any spherically-symmetric distribution of charge trivially has $\vec{\mathcal{D}}_E = 0$, and not so trivially $\vec{\mathcal{Q}}_E = 0$ also.⁷ In fact *all* moments beyond the 0th are zero: $\psi = q_{\text{tot}}/(4\pi\epsilon_0 r)$ outside any such distribution (Birkhoff's theorem).⁸

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

Next, suppose that $+q$ is located at $(0, 0, a)$ and $-q$ is at $(0, 0, -a)$. Then $\vec{\mathcal{D}}_E = (2qa)\hat{z}$. You should compute that $\vec{\mathcal{Q}}_E = 0$ directly from the definition, but here is a more insightful, and generalizable, argument.

⁶There 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."

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

⁸Robert Hooke intuitively understood this result and communicated it to Newton around 1679. Newton proved it in 1685.

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

- Invert all positions, $\vec{r}'_{(\ell)} = -\vec{r}_{(\ell)}$, and also
- Reverse the signs of each charge, $q'_\ell = -q_\ell$. transform T1

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 $\vec{\mathcal{D}}'_E = \vec{\mathcal{D}}_E$.
- The new distribution has three minus signs in the quadrupole moment, so $\vec{\mathcal{Q}}'_E = -\vec{\mathcal{Q}}_E$.

If transformation T1 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_{tot} must equal zero.
- $\vec{\mathcal{D}}_E = \vec{\mathcal{D}}'_E = \vec{\mathcal{D}}_E$, which is a tautology, so there is no restriction on the dipole moment. consequences
- But $\vec{\mathcal{Q}}_E = \vec{\mathcal{Q}}'_E = -\vec{\mathcal{Q}}_E$, so the quadrupole moment equals zero. In fact, every 2^p -pole moment with p an even integer must be zero. of T1 symmetry

(3.9)

Returning to the specific distribution with $+q$ located at $(0, 0, a)$ and $-q$ at $(0, 0, -a)$, we see it is indeed unchanged under T1. 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 (see Problem 3.2).

Your Turn 3E

- a. Think up a charge distribution that becomes *minus* itself under T1. [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).

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 $q = \mathcal{D}_E/(2a) \rightarrow \infty$. That idealized limit is called the **pure dipole** or **point dipole** distribution.

3.7 FORCE AND TORQUE ON A FIXED CHARGE DISTRIBUTION

3.7.1 Potential energy depends both on position and on orientation

Suppose that a localized charge distribution (subsystem 1, for example, a molecule) sits in an externally created, static electric potential ψ^{ext} (from subsystem 2, for example, a macroscopic lab apparatus). Choose a reference point somewhere inside distribution 1 and then describe it by stating the location \vec{r} in space of that point, the constituent charges q_ℓ and their offsets $\vec{r}_{(\ell)}$ 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 is⁹

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

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 is slowly varying over the size of subsystem 1. Mathematically, this means that it is characterized by some length scale L , the n th derivatives of ψ^{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: $L \gg \|\vec{r}_{(\ell)}\|$. Certainly a macroscopic apparatus will be much bigger than any individual molecule that we choose to study. In this circumstance, we may Taylor expand the external potential about the reference point:

$$U = \sum_{\ell} q_{\ell} \psi^{\text{ext}}(\vec{r}) + \sum_{\ell} q_{\ell} \left. \frac{\partial \psi^{\text{ext}}}{\partial \vec{r}} \right|_{\vec{r}} \cdot \vec{r}_{(\ell)} + \sum_{\ell} q_{\ell} \left. \frac{1}{2} \frac{\partial^2 \psi^{\text{ext}}}{\partial \vec{r}_i \partial \vec{r}_j} \right|_{\vec{r}} \vec{r}_{(\ell)i} \vec{r}_{(\ell)j} + \dots \quad (3.10)$$

$$= q_{\text{tot},1} \psi^{\text{ext}}(\vec{r}) + \vec{\mathcal{D}}_{\text{E}(1)} \cdot \vec{\nabla} \psi^{\text{ext}} \Big|_{\vec{r}} + \dots \quad (3.11)$$

In the last formula, the ellipsis denotes terms with second and higher derivatives.

Your Turn 3F

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 $\vec{\mathcal{D}}_{\text{E}(2)}$.

⁹See Section 2.2.1.

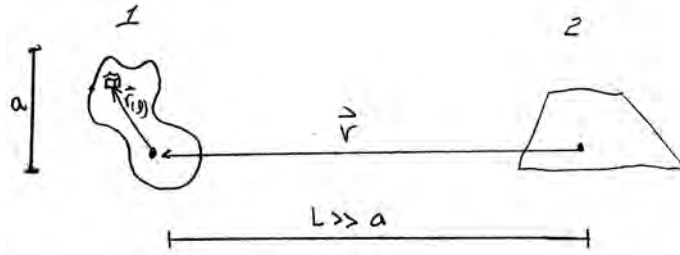


Figure 3.1: Two interacting, distant, rigid charge distributions.

Your Turn 3G

- Find the leading-order contribution to the interaction potential energy, $U(\vec{r})$. How does it depend on the separation distance r ?
- 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 \hat{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 $q_{\text{tot}}\vec{E}_i^{\text{ext}}$ (from the first term of Equation 3.11), the next term is

$$= (-\vec{\nabla}_i \vec{\nabla}_j \psi^{\text{ext}}) \vec{\mathcal{D}}_{E,j} = \vec{\mathcal{D}}_E \cdot \vec{\nabla}_i \vec{E}^{\text{ext}}.$$

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{z} and passing through the reference point we used to define the multipole expansion. Then $-dU/d\theta$ is the z component of torque, $\vec{\tau}_3$. To find it, we displace each constituent charge from $\vec{r}_{(\ell)}$ to $S\vec{r}_{(\ell)}$, where the infinitesimal **rotation matrix** S is defined by

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

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

$$U = \sum_{\ell} q_{\ell} \psi^{\text{ext}}(\vec{r} + \vec{r}_{(\ell)} + d\theta \vec{T} \cdot \vec{r}_{(\ell)}) \\ = \cancel{q_{\text{tot}} \psi^{\text{ext}}(\vec{r})} + \sum_{\ell} q_{\ell} \vec{\nabla} \psi^{\text{ext}}|_{\vec{r}} \cdot (\cancel{\vec{r}_{(\ell)}} + d\theta \vec{T} \cdot \vec{r}_{(\ell)}) + \dots$$

The crossed-out terms are constants; the change as we rotate is

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

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

$$\vec{\tau}_3 = -dU/d\theta = -\varepsilon_{ij3} \vec{E}_i^{\text{ext}} \vec{\mathcal{D}}_{\text{E},j} = (\vec{\mathcal{D}}_{\text{E}} \times \vec{E}^{\text{ext}})_3.$$

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

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 $\vec{\mathcal{D}}_{\text{E}} \parallel \vec{E}^{\text{ext}}$. 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.
- CO₂ 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 ∞).¹⁰ 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 permanent dipole moment but in liquid state. For example, in water at room temperature the dipoles are thermally randomized, so the average $\langle \vec{\mathcal{D}}_{\text{E}} \rangle = 0$. However, the random thermal fluctuations of neighboring molecules will be partially *correlated*, leading to nonzero $\langle \vec{\mathcal{D}}_{\text{E}(1)} \cdot \vec{\mathcal{D}}_{\text{E}(2)} \rangle \neq 0$, and hence decreased energy via Equation 3.11: Each of the dipoles can be thought of as partially *aligning* the other one. So there will be a net attraction after all, in this context sometimes called **Keesom interaction**.

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

Even spherically symmetric atoms acquire weak, long-range electrostatic interactions via fluctuations.

¹⁰See Problem 3.4.

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 quantum fluctuations. And these quantum fluctuations again have an energetic tendency to correlate with those of a neighboring atom. This source of electrostatic attraction is sometimes called **London force** or **dispersion interaction**.

Together, the quantum and the statistical correlation attraction effects are sometimes called the **van der Waals** interaction. Van der Waals interactions play a dominant 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* a dipole moment that does not average to zero. For example, a CO₂ molecule can bend. 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.

Objects that are initially neutral and unpolarized nevertheless feel electrostatic forces in a nonuniform field.

Thus, for example, the resulting **induced dipole** 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. And 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: $\vec{D}_E = \alpha \vec{E}$, where α is a constant called the molecular **polarizability**.¹¹ That induced moment in turn feels a force $\alpha \vec{E}_i \vec{\nabla}(\vec{E}_i) = \frac{1}{2} \alpha \vec{\nabla}(E^2)$ 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 CO₂. In that case, the quadrupole term is the dominant one.

¹¹Later, we'll account for the possibility that the polarizability may not be isotropic (Section 13.3.1 and Chapter 50).

- There is also a multipole expansion for electromagnetic *radiation*, as we'll see. Here, too, if the transition dipole moment is zero, still the atom or molecule can radiate via its 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 a 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 (unless that's zero).

FURTHER READING

Intermediate:

General: Pollack & Stump, 2002, §3.8; Zangwill, 2013, chap. 4.

Optical tweezers: Perkins, 2014.

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

Technical:

Almost all about multipole expansions: Raab & de Lange, 2005.

T₂

3.2'a Counting moments

There's only one kind of monopole field, characterized by only one overall constant of proportionality, q_{tot} . There's *essentially* only one kind of dipole field: You can convert any of the $\psi_i^{[1]}$ 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 interesting. Even if we choose a standardized normalization, the quadrupole tensor $\vec{\vec{Q}}_E$ 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, a symmetric matrix like $\vec{\vec{Q}}_{E,ij}$ has three real eigenvalues, each of which is rotation-invariant. One of these is redundant because $\vec{\vec{Q}}_E$ is traceless, but the other two are invariants characterizing the quadrupole. Qualitatively, we may say that some quadrupoles have more symmetry than others, because there is an invariant distinction between those for which two eigenvalues match (**uniaxial symmetry**) and those for which no two match (**biaxial symmetry**).¹² Try to find concrete examples of each case.

It's an example of the unity of physics that these same concepts arise in liquid crystals.

3.2'b Connection to spherical harmonics

We won't say much about the spherical harmonic functions $Y^{\ell m}$ in this course, but take a moment to examine the quadrupole fields (Equation 3.4), and show that:

- The angular dependences of the dipole potentials $\vec{\psi}_i^{[1]}$ are simple linear combinations of Y^{1m} .
- The angular dependence of $\vec{\psi}_{zz}^{[2]}$ is the same as that of Y^{20} : Both are $-\frac{1}{3} + \cos^2 \theta$.
- The angular dependence of $\vec{\psi}_{xx}^{[2]}$ is the same as the linear combination $Y^{22} + Y^{2,-2} - Y^{20}$.
- The angular dependence of $\vec{\psi}_{yy}^{[2]}$ is the same as the linear combination $Y^{22} + Y^{2,-2} + Y^{20}$.
- The angular dependences of $\vec{\psi}_{zx}^{[2]}$ and $\vec{\psi}_{zy}^{[2]}$ are the same as the linear combinations $Y^{21} \pm Y^{2,-1}$.
- (You think about $\vec{\psi}_{xy}^{[2]}$.)

If you've studied spherical harmonics, you probably 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 result 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).

T₂

3.7.3'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 $\approx (10^{-38} e) \text{ cm}$, whereas in 2018 the experimental bound was $\mathcal{D}_E \lesssim (10^{-29} e) \text{ cm}$. (In contrast,

¹²Why can't all three match?

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.)

- Sketch a charge distribution that could result in such a field.
- If the electric field's magnitude falls with distance as $\vec{E} \sim r^p$, what is the exponent p ?

3.2 *Electrostatic multipole*

- 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.
- 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.

- Such a distribution must have vanishing dipole moment, because no vector other than $\vec{0}$ can be rotationally invariant. But work this out directly from the definition of dipole moment as an integral over the distribution.
- More precisely, the dipole moment *computed about the point of symmetry* must be

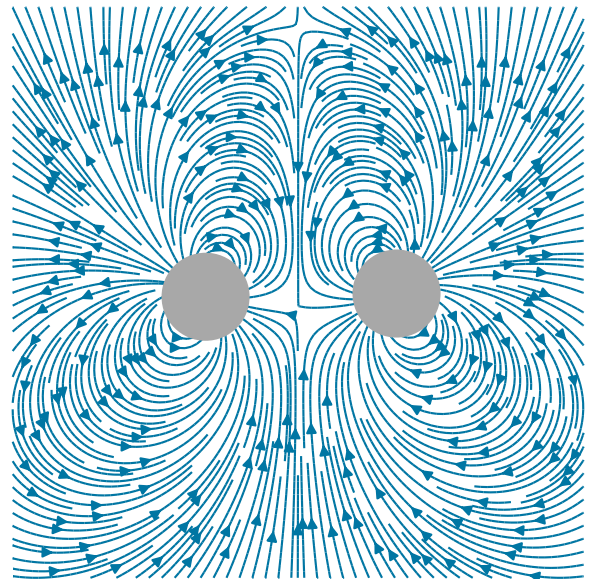


Figure 3.2: See Problem 3.1.

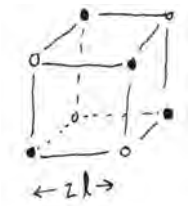


Figure 3.3: A tetrahedron constructed as the solid dots.

- zero. Repeat your calculation but this time suppose that the distribution, while spherically symmetric, is centered about some point \vec{h} other than the origin.
- Repeat (a) but for the distribution's quadrupole moment. This time we can't just say, "It must equal zero because there's no such thing as a rotationally-invariant tensor of rank 2," because that's *not a true statement*. So work it out and then discuss.
 - Repeat (b) for the quadrupole moment.

3.4 Tetrahedron

- Consider four identical point charges q rigidly fixed at the vertices of a tetrahedron (solid dots shown in Figure 3.3), and $-4q$ fixed at its center. The distance from the center to any vertex is 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: A convenient construction of a tetrahedron begins with a cube centered on the origin, that is, with eight vertices $(\pm\ell, \pm\ell, \pm\ell)$ where ℓ is a length scale related to a . You can select four of the cube's eight vertices and use them as the vertices of the desired tetrahedron, as shown in the figure.]
- Does your result appear to be relevant to the behavior of some well known small molecule? Does it explain a big qualitative difference between that molecule's properties and those of, say, water?

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.

- 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"). 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 $\vec{\alpha}$, β , and γ 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 approximation.

3.8 Multipole math

Derive a formula for each of the functions $\nabla^2(r^{-5}\vec{r}_i\vec{r}_j)$ where ∇^2 is the Laplace operator and the indices i and j each can be 1, 2, or 3. 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.

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 described a limiting charge distribution whose potential consists of only the dipole term of Equation 3.2 (page 37). By computing minus the gradient of such a function, you can find the corresponding electric field. In this problem, you are to find

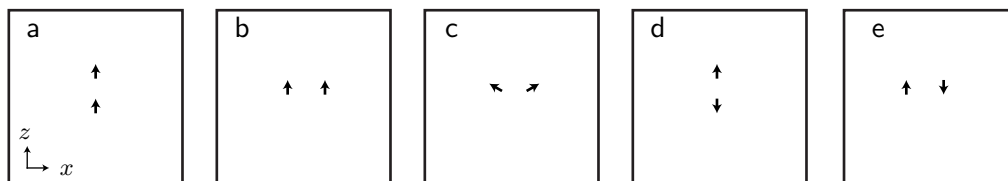


Figure 3.4: See Problem 3.10.

and display exact expressions for the fields outside *pairs* of pure dipoles that are *not* located at the origin.

- 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 $-3a < z < +3a$.¹³ (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 streamlines¹⁴ 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?).
- Repeat for the situation in (b): two dipoles directed along $+\hat{z}$ located at $(\pm a, 0, 0)$.
- Repeat for (c): two dipoles tilted ± 60 deg away from \hat{z} towards the $\pm x$ -axis and located at $(\pm a, 0, 0)$. What familiar molecule might this model?
- Repeat for (d): similar to (a), but the dipoles oppose each other. What familiar molecule might this model?
- Repeat for (e): similar to (b), but the dipoles oppose each other. This might model two familiar molecules electrostatically sticking to each other (like what?)
- 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

Learn how to use a computer to create 3d streamplots, and show them for an electric dipole field. Look at various viewing angles till you find one that is most informative.

3.12

[Not ready yet.]

3.13 Scalar potential

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

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

¹³One 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* directly.

¹⁴See Section 0.3.1. Changing the normalization does not change the streamlines.

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

- Compute $\vec{\nabla} \cdot \vec{E}$ for the field you found in (a) and comment. To what class of potential functions does this one belong?
- Could this function describe the newtonian gravitational potential far from a localized distribution of mass?

3.14 Animate equipotentials

Consider a pure dipole with $\vec{\mathcal{D}}_E = q \begin{bmatrix} 1 \text{ m} \\ 1 \text{ m} \\ 0 \end{bmatrix}$. 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 $f(\vec{r}) = \frac{4\pi\epsilon_0}{(1 \text{ m})^q} \psi(\vec{r})$. 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**, also called a **level set** (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.¹⁵ 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 explained below.

Specifically, think about an equipotential surface of a pure dipole field:

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

where A' is a constant related to A . The right side has a special property: 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 θ, φ . For example, if the dipole moment lies along the \hat{z} direction, then $r^2 = B \cos \theta$, where $B = \|\vec{\mathcal{D}}_E\|/A$. For each value of B , we can set up a grid of θ, φ values, evaluate r , drop the points that have no solution ($r^2 < 0$), and create a 3D surface plot.

- Carry out the steps just mentioned for one interesting nonzero value of $1/B$. Then repeat over an interesting range of $1/B$ values and make the animation described earlier.
- Repeat for an axisymmetric (uniaxial) pure quadrupole field, for example, the one

¹⁵Problem 3.14 may be a useful warmup for this.

with

$$\vec{Q}_{E,ij} = (\text{const}) \begin{bmatrix} 1 & & \\ & 1 & \\ & & -2 \end{bmatrix}_{ij} .$$

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

$$\vec{Q}_{E,ij} = (\text{const}) \begin{bmatrix} 1 & & \\ & -1 & \\ & & 0 \end{bmatrix}_{ij} .$$

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 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 orientations of the donor, the acceptor, and the vector separating them.

Physical idea: Electrostatic dipole-dipole interaction explains these dependences.

4.2 FLUORESCENCE AND AN UNEXPECTED PHENOMENON

Before we address the main question, here is a little history on a phenomenon that may at first seem arcane and in any case distant from classical electrostatics.

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

Some molecules are fluorescent: They can capture a photon, wait a long time (typically a nanosecond), and then emit another photon. Even a single atom can absorb and reemit light, but for medium-size molecules, there is an interesting twist. A fluorescent molecule, or **fluorophore**, has a characteristic **excitation spectrum**, the probability per incoming photon of getting excited as a function of incoming wavelength. Each fluorophore also has a characteristic **emission spectrum**, the probability density function for the wavelengths of emitted photons. 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,¹ 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.

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, with excitation spectrum peaking around $\lambda_{1,\text{ex}}$ and emission spectrum peaking around $\lambda_{1,\text{em}} > \lambda_{1,\text{ex}}$ (a **donor**). 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 $\lambda_{2,\text{em}}$. It is at least possible to imagine illuminating with a wavelength near $\lambda_{1,\text{ex}}$ but completely missing the acceptor’s excitation spectrum (so that only the donor gets directly excited, Figure 4.1), but nevertheless observing emitted fluorescence around $\lambda_{2,\text{em}}$. In such a situation, some light emitted from the donor could excite 2, instead of leaving the sample or getting absorbed.

In fact, this sort of two-stage fluorescence was 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,² and it is a workhorse tool in labs around the world. Other modified versions of FRET have names like

¹One elegant method genetically encodes a protein resembling a natural protein, but with an extra fluorescent group. Cells with this gene will express (manufacture) a fluorescent version of the protein of interest (a **chimera**). Another method fuses the fluorophore to an antibody that attaches specifically just to the objects under study, then introduces that construct into a cell (or uses it in vitro).

²Some authors drop the first word and instead say “RET.”

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 filters that eliminate the exciting light but pass light in one of the emission bands. [Data from Johnson et al., 1993.]

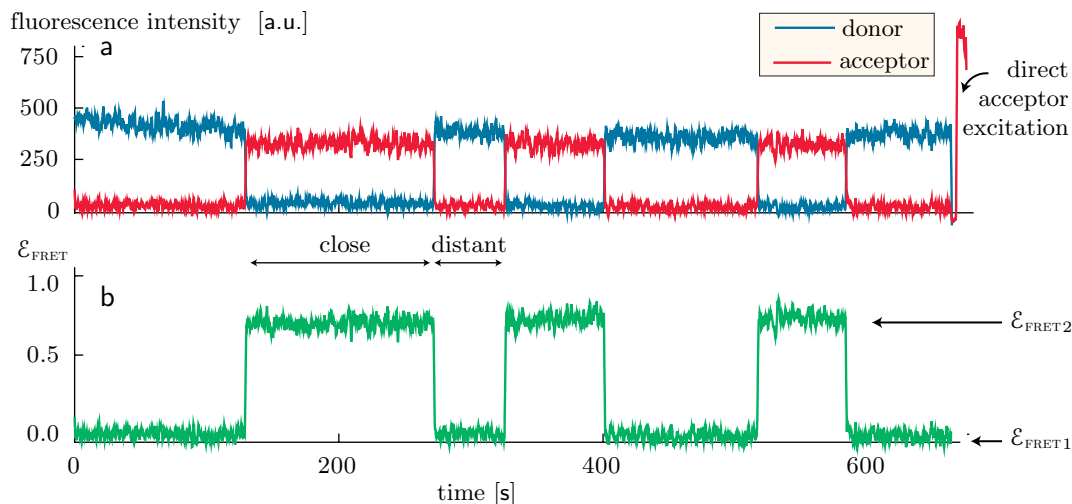
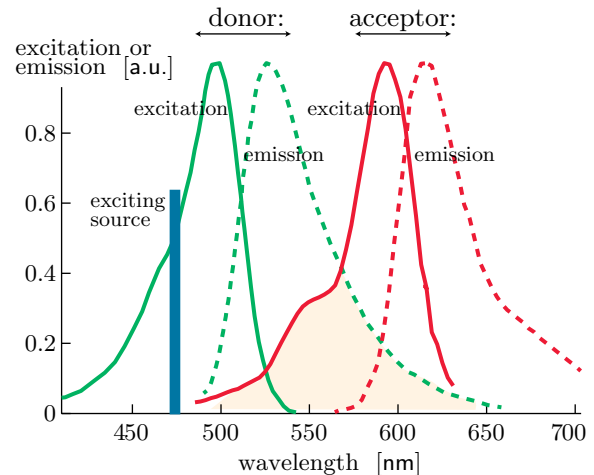


Figure 4.2: [Experimental data.] **Modern application of FRET.** (a) Time series of the fluorescence intensities in the donor and acceptor emission bands, measured from a single DNA molecule. The DNA was labeled with a donor (Cy3) at one end and an acceptor (Cy5) at the other, and illuminated near the donor's excitation peak. There are distinct episodes indicating that the donor and acceptor are either close to, or far from, each other. A constant has been subtracted from each trace to account for stray light, detector noise, and other background. The *arrow* shows a brief interval of illumination at the *acceptor's* excitation peak, to confirm that the acceptor had not yet photobleached. (b) Corresponding FRET efficiency defined in Equation 4.1. [Data courtesy Taekjip J Ha; see also Vafabakhsh & Ha, 2012.]

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

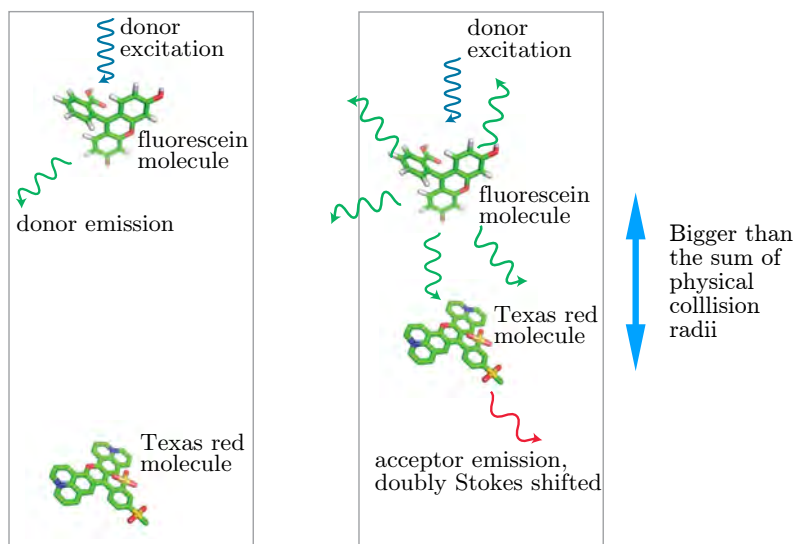


Figure 4.3: [Cartoons; not to scale.] **The puzzle of resonant energy transfer.** *Left:* When the donor (*top*) is far from the acceptor (*bottom*), excitation transfer seemingly should be almost impossible—yet it occurs. *Right:* Even when the donor is close to the acceptor, transfer should not be very probable—yet in Figure 4.4 its probability sometimes exceeds 90%. Section 4.3.1 will argue that the resolution of this puzzle is that for separations up to a few times the physical molecule size, donor and acceptor are coupled mainly by strong oscillating dipole fields—not by transfer of radiation as sketched here. In these simplified cartoons the fluorophores are depicted as isolated molecules, but in applications they are usually attached to something else, for example, the two ends of a DNA molecule.

low concentrations, because nearly every photon emitted from a donor would *miss* hitting any acceptor (Figure 4.3a). Even at high concentration, most emitted photons would be going in the wrong direction to hit an acceptor, so we’d still expect very low probability of two-stage fluorescence (Figure 4.3b). Experimentally, however, it was seen at low concentrations, corresponding to average intermolecular separation of several nanometers. Moreover, at higher but still modest concentrations the probability of transfer can be quite high.³ The experiments were repeated; researchers reluctantly concluded that there was some major missing piece in their understanding, documented the anomaly, and moved on. Excitation transfer certainly did obey the rule that $\lambda_{1,\text{em}}$ must match $\lambda_{2,\text{ex}}$ —but it seemed that it shouldn’t have been happening at all.

We can only admire the tenacity and thoroughness of these early researchers, insisting that something was wrong even before the ink had dried on most of quantum mechanics.

³ $[T_2]$ 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.

4.3 DIPOLE-MEDIATED TRANSFER

4.3.1 Electrostatic 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 solvent 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?

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 like r^{-1} . 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 powers of r^{-1} . 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 pulsating electric fields 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 photon 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”

We 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 generally 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 \hat{a} , 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{a} .

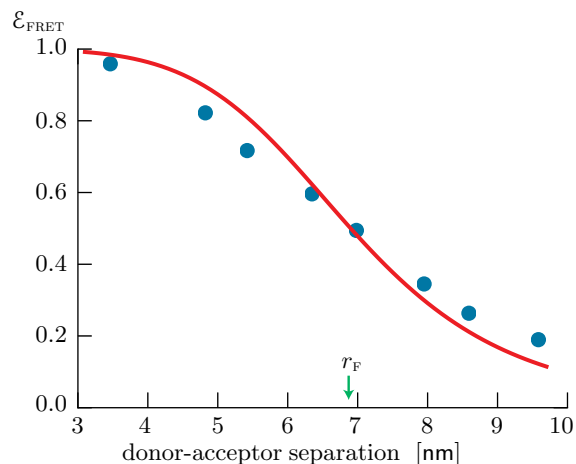
We know from mechanics that energy will be transferred from one oscillator to another at a rate proportional to the amplitude of the force squared.⁵ The amplitude of the electric force on a test particle from a dipole falls with distance as r^{-3} , so we expect this rate to be $\propto r^{-6}$. In addition,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

FRET has a characteristic falloff with distance between donor and acceptor.

⁴The Keesom and dispersion interactions in Section 3.7.3 (page 45) were of dipole-dipole origin.

⁵You’ll recall the details in Problem 4.1.

Figure 4.4: [Experimental data.] **FRET efficiency as a function of the separation between donor and acceptor.** $\mathcal{E}_{\text{FRET}} = 1$ 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 $\mathcal{E}_{\text{FRET}}$. The *curve* shows the value of $\mathcal{E}_{\text{FRET}}$ 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. [Data from Lee et al., 2005.]



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 $\frac{Ar^{-6}}{B+Ar^{-6}}$, where A and B are constants. Rephrasing gives the FRET efficiency as

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

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** $r_{\text{F}} = (A/B)^{1/6}$.

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 will sell you donor/acceptor fluorophores and 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 *time resolution* limited only by our ability to gather photons—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, averaging these angular dependences and leading to a transfer rate with a single effective Förster radius in Equation 4.1. Something similar may also happen even if the fluorophores

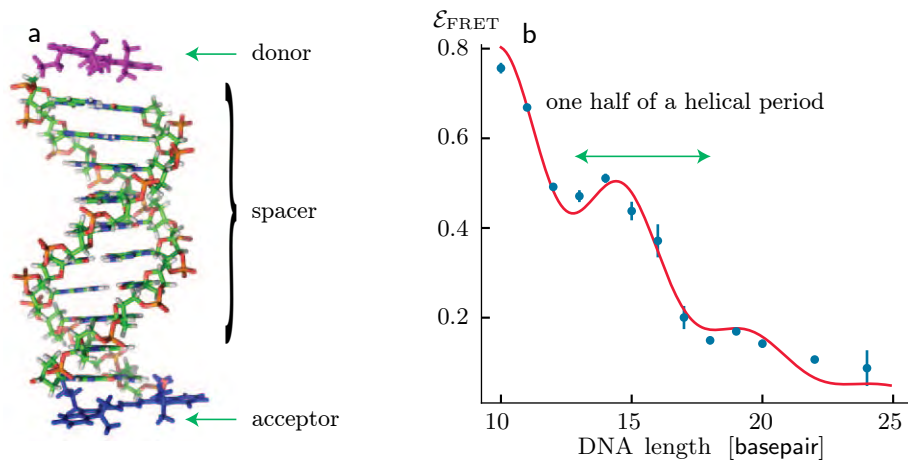


Figure 4.5: [Cartoon; experimental data.] **Orientation dependence of FRET.** (a) See text. (b) FRET efficiency as a function of spacer length. The curve shown is a fit to data. The molecule is not rigid; instead, thermal motion constantly flexes it. To account for this effect, the fitting function was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

FRET has a characteristic dependence on the orientations of the donor, the acceptor, and the vector separating them.

are covalently bound to different points on a single object, as long as the connections are flexible. However, in general the Förster radius does depend on orientation:⁶

Your Turn 4A

Use your result in Your Turn 3B (page 40) 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}_E and the acceptor's polarizability \hat{a} 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 is a relative rotation φ of the basepair at one end relative to the other that depends on length. Specifically, this angle sweeps through a full circle after the addition of 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 when φ increases by π . The data shown in the figure indeed show such modulation, partially washed out by orientational fluctuations (the construct was not perfectly rigid).

⁶You'll work out more aspects in Problem 4.1.

4.4 PLUS ULTRA

A “FRET pair” (donor and acceptor) can be used to report on conformational transformations 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 a “FRET-based calcium reporter.”

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.

Some intracellular reporters utilize FRET.

FURTHER READING

Intermediate:

Many more applications of FRET: Nelson, 2017, chap. 2.
en.wikipedia.org/wiki/Single-molecule_FRET.

Technical:

History of FRET: Clegg, 2006.

Quantum theory of FRET, and why the classical treatment works: Nelson, 2017, chap. 14.

Single molecule FRET: Ha et al., 1996. Review: Hwang et al., 2009.

Spectroscopic ruler: Sindbert 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 ω_D (determined by the donor’s excited state), and amplitude J (determined by the donor’s state and the distance to the acceptor):

$$f_D(t) = J \cos(\omega_D t). \quad (4.2)$$

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, for example by fluorescence. Calling the friction constant η , Newton’s law $f_{\text{tot}} = ma$ states that the donor’s position $x(t)$ obeys

$$m \frac{d^2}{dt^2} x = -kx - \eta \frac{d}{dt} x + f_D. \quad (4.3)$$

To simplify this equation, define new symbols $\omega_A = \sqrt{k/m}$, $\bar{\eta} = \eta/m$, and $L = J/m$, and eliminate k , η , and J by writing them in terms of the new symbols.

- After a short transient, the solution $x(t)$ will oscillate at angular frequency ω_D . So consider the trial solution⁷ $x(t) = A \cos(\omega_D t) + B \sin(\omega_D t)$. Find the constants A and B in terms of L , $\bar{\eta}$, ω_D , and the acceptor's resonant frequency ω_A .
- 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 $\mathcal{P} = \bar{\eta}(dx/dt)^2$. Evaluate this for your solution.
- 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.
- 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 ω_D , ω_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 $\wp_D(\omega_D)$ and $\wp_A(\omega_A)$:

$$\langle \langle \mathcal{P} \rangle \rangle = \int d\omega_D \wp_D(\omega_D) \int d\omega_A \wp_A(\omega_A) \langle \mathcal{P}(\omega_D, \omega_A) \rangle.$$

To simplify this expression, suppose that the damping $\bar{\eta}$ 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 ω_D , ω_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 $\Delta\omega$.

- The donor creates a dipole field, which shakes charges on the acceptor. Imagine the acceptor dipole as having a fixed axis $\hat{\mathcal{D}}_A$ 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 $\hat{\mathcal{D}}_A$. 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.
- 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.

⁷If you prefer, you can use complex exponential notation (Section 18.7, page 266).

CHAPTER 5

Curvilinear Coordinates and Separation of Variables

[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 2.6 (page 30) 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. 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 Maxwell’s 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 “pretty good” because Maxwell’s equations look *almost* the same in them,

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 is separable. We’ll see that separation of variables is useful whenever we use such coordinates.

Although the Laplace operator is separable in ordinary cartesian coordinates x, y, z , nevertheless many problems have boundaries that don’t look simple in those coordinates. So we will find some other coordinate systems, collectively called **curvilinear**, in which the Laplace operator is again separable, but the surfaces with one coordinate constant are not 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 y, z , plus a term not involving x, z , plus a term not involving x, y . Because it separates in this way, we say that the operator is **separable** in these coordinates. 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 $\kappa + \lambda + \nu = 0$.

If our boundaries are planes of constant x, y , or z (rectangular box), then this coordinate choice can be especially useful.

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:

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

Note that the first of these 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. It will soon be convenient that these two vector fields are everywhere perpendicular to each other.

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

$$\vec{\nabla} f = \mathbf{J} \begin{bmatrix} f_r \\ f_\varphi \end{bmatrix}, \quad \text{where } \mathbf{J} = \begin{bmatrix} \partial r / \partial x & \partial \varphi / \partial x \\ \partial r / \partial y & \partial \varphi / \partial y \end{bmatrix}. \quad (5.2)$$

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

$$\int d^2\vec{r} \vec{\nabla} f \cdot \vec{\nabla} g = \int d^2\vec{r} [f_r, f_\varphi] J^t J \begin{bmatrix} g_r \\ g_\varphi \end{bmatrix}. \quad (5.3)$$

However, we also have

$$\int d^2\vec{r} \vec{\nabla} f \cdot \vec{\nabla} g = \int d^2\vec{r} [\vec{\nabla} \cdot (f\vec{\nabla}g) - f\nabla^2g] = - \int d^2\vec{r} f\nabla^2g. \quad (5.4)$$

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 . To derive a formula for ∇^2g , 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 2×2 matrix $J^t J$. It's messy to compute J directly, because once we compute $\partial r / \partial x$ and so on we must then re-express everything as functions of r and φ . 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 $\vec{e}_{(r)}$ and $\vec{e}_{(\varphi)}$ defined by Equation 5.1. Thus, we may write

$$J^{-1}(J^{-1})^t = \begin{bmatrix} \|\vec{e}_{(r)}\|^2 & \vec{e}_{(r)} \cdot \vec{e}_{(\varphi)} \\ \vec{e}_{(r)} \cdot \vec{e}_{(\varphi)} & \|\vec{e}_{(\varphi)}\|^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & r^2 \end{bmatrix}.$$

We want $J^t J$, 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 $\vec{e}_{(i)}$'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 $J^t J$, then integrate by parts:

$$\int (r dr d\varphi) (f_r g_r + f_\varphi r^{-2} g_\varphi) = - \int (dr d\varphi) \left(f \frac{\partial}{\partial r} (r g_r) + f r^{-1} \frac{\partial}{\partial \varphi} g_\varphi \right).$$

We want to rephrase this expression into a form resembling the right side of Equation 5.4, so multiply and divide by r :

$$= - \int (r dr d\varphi) f \left[r^{-1} \frac{\partial}{\partial r} (r g_r) + r^{-2} g_{\varphi\varphi} \right]. \quad (5.5)$$

Equation 5.4 says that the last expression equals $-\int d^2\vec{r} f \nabla^2 g$ 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 *arbitrary* f is if the terms in square brackets of Equation 5.5 are equal to $\nabla^2 g$, and this is a familiar formula:

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

5.3.3 Plane polar payoff

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

$$0 = \frac{B_{\varphi\varphi}}{B} + \frac{r}{A} \frac{\partial}{\partial r}(rA_r).$$

The first term is completely independent of r . The second term is completely independent of φ . 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 $A(R) = 1$, 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 $\partial^2/\partial r^2 + \partial^2/\partial\varphi^2$! 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 θ and φ ; by adding a constant we may take its value to be zero. Far from the sphere, we get the same uniform electric field we’d have had from the charged plates alone (without the sphere), so

$$\psi \rightarrow Cz \quad \text{at } r \gg R, \tag{5.6}$$

where C is a constant related to the surface charge density on the plates. Now we want ψ 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: ψ will be independent of azimuthal angle φ . 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 φ -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 θ to $\mu = \cos \theta$, so $d\mu = -\sin \theta d\theta$. Thus, the second equation becomes the **Legendre equation**:

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

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 αz at $r \rightarrow \infty$, as desired.¹ And we can satisfy the inner boundary condition by choosing $\beta = -R\alpha$. 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: The multipole expansion missed it because it does not drop off with distance.²

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.) Ben's breakthrough was to connect the abstractions of natural philosophy to the urgent practical matter of saving lives.³

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

A charged conductor with a sharp point creates strong fields there.

¹The apparent singularity at $r \rightarrow 0$ is not a problem because this solution is only to be used outside the sphere.

²Physically, the charged plates at infinity violate the multipole expansion's assumption that all charges are confined to a small zone.

³Some people objected—lightning strikes were manifestations of divine will, which humans would defy at their peril. Ben was persistent (see the epigraphs on pages 25 and 65).

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.⁴ 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 set of one of the coordinates are ellipsoids. By following the steps in this chapter, you'll show that the Laplace operator is separable in those coordinates as well, and hence get an exact solution for the lightning-rod problem almost as readily as in Section 5.4.

5.6 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 need to be sure you know how they work.

For example, we will later need a formula for the divergence the vector field

$$\vec{V}(\vec{r}) = \frac{1}{r} \hat{x} e^{ikr}. \quad (5.8)$$

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

$$\vec{\nabla} \cdot \vec{V} = r^{-2} \frac{\partial}{\partial r} (r^2 \vec{V}_r) + \frac{1}{r \sin \theta} \left(\frac{\partial}{\partial \theta} (\sin \theta \vec{V}_\theta) + \frac{\partial \vec{V}_\varphi}{\partial \varphi} \right). \quad (5.9)$$

But it's tricky to apply this formula to Equation 5.8 correctly! Note that $\hat{r} = [\sin \theta \cos \varphi, \sin \theta \sin \varphi, \cos \theta]^t$, so

$$\hat{x}_r = \hat{r} \cdot \hat{x} = \sin \theta \cos \varphi, \quad \hat{x}_\theta = \hat{\theta} \cdot \hat{x} = \cos \theta \cos \varphi, \quad \hat{x}_\varphi = \hat{\varphi} \cdot \hat{x} = -\sin \varphi.$$

After you substitute the first of these into Equation 5.8 and similar formulas into the black-box formula Equation 5.9, you *still* must do a lot of algebra to find

$$\vec{\nabla} \cdot \vec{V} = \sin \theta \cos \varphi (-r^{-2} + ik/r) e^{ikr}.$$

5.6.2 Easy way

Instead of using a black-box formula, let's do it from scratch. Use the product rule, $\vec{\nabla} \cdot (f \hat{x}) = \hat{x} \cdot \vec{\nabla} f + f \vec{\nabla} \cdot \hat{x}$ 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,

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

⁴It'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).

5.7 PLUS ULTRA

There are a total of 11 coordinate systems in which the 3D Laplace operator is separable, plus two more that are almost as good. See the references.

FURTHER READING

Semipopular:

Kite experiment: Tucker, 2003.

Intermediate:

Ellipsoidal coordinates for electrostatics problems: Vanderlinde, 2004, chap. 5.

Pollack & Stump, 2002, §5.2.1.

About the famous list of 13 separable 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 & Feshbach, 1953.

Ben Franklin: Cohen, 1990; Franklin, 1941.

PROBLEMS

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.

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

$$(x/\alpha)^2 + (z/\beta)^2 = 1,$$

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

Consider two points \mathbf{P}_{\pm} on the \hat{z} axis, located at $z = \pm\sigma$. For any other point, let r_{\pm} be the distances from that point to \mathbf{P}_{\pm} . We can specify that point either by its x , y , z values, or its cylindrical polar coordinates ρ , φ , z , or by new coordinates ξ , η , and φ . Here φ is the same as in cylindrical coordinates and

$$\xi = (r_+ + r_-)/(2\sigma) \quad \eta = (r_+ - r_-)/(2\sigma).$$

- a. Show that the surface $\{\xi = \xi_0\}$ is what you get by rotating an ellipse about its axis. Find its major and minor axis lengths in terms of σ and ξ_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 ρ and z in terms of ξ and η . [*Hint:* Express $\xi\eta$ and $(\xi^2 - 1)(1 - \eta^2)$ in terms of ρ and z , then think.]

Near-field scanning optical microscopes generate huge electrostatic fields localized to nanometer regions.

- c. Thus, express x, y, z in terms of $\xi, \eta,$ and φ . 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 very nice 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 ξ - η plane corresponding to the region outside the surface in (a).
- e. Use (c,d) to express the integral $\int d^3r \vec{\nabla}\Upsilon \cdot \vec{\nabla}\psi$ in the coordinates $\xi, \eta,$ and φ . Here ψ is any function independent of φ , while Υ , also independent of φ , is nonzero only in some small region of ξ and η .
- f. Use integration by parts to work out the Laplace operator $\nabla^2\psi$ in these coordinates, for the case where ψ is independent of φ .

You're ready to begin the problem, which is to find the electrostatic potential in the region outside the conductor, subject to the boundary conditions:

$$\psi = 0 \text{ on the surface, } \quad \psi \rightarrow -E_\infty z \text{ far away.}$$

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 η , 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 enforce 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_∞ .
- 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 ψ/E_∞ 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 ψ/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, ξ - η coordinates are still useful in realistic treatments of NSOM probes and their field-focusing properties.]

5.2 Razor's edge

A charged conductor with a sharp edge also creates strong fields there

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 ψ 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, not math, 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 2.5 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 Σ . 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 surface charge density on the left plate.

Use the electric Gauss law to find that between the planes, $\vec{E}_x = \sigma_q/\epsilon_0$. Integrate $-\vec{E}$ along x to find the potential throughout the gap, and its total change $\Delta\psi = \psi(0) - \psi(w) = \sigma_q w/\epsilon_0$. 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.

For this system, $C = \epsilon_0\Sigma/w$, that is, fixed capacitance per plate area of $\mathcal{C} = \epsilon_0/w$. The natural SI unit for capacitance is coulombs per volt, which is called the farad: $1\text{ F} = 1\text{ coul/volt}$.

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

$$\mathcal{E} = \frac{1}{2}q^2/C.$$

Rephrasing using Equation 6.1 gives the stored electrostatic potential energy as

$$\mathcal{E}/(\text{volume}) = \frac{\epsilon_0}{2} \|\vec{E}\|^2. \quad (6.2)$$

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.

Atomic nuclei carry enormous electrostatic self-energy.

Your Turn 6A

- Adapt the preceding argument to find the work that must be done to bring total charge q onto a spherical shell of radius R .
- Then evaluate the expression in Equation 6.2 everywhere outside the shell, integrate it over space, and compare the result to (a).
- A heavy atomic nucleus may contain charge of around $100e$ confined to a sphere of radius ≈ 10 fm. Suppose that nucleus fissions into two fragments each with about half the charge, and with radius smaller by a factor of $2^{1/3}$. Approximate by supposing that all the charge sits on the surfaces of the spheres. Compute the change in electrostatic self-energy and comment.

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 . The electrostatic potential energy of such a shell is $\frac{1}{2}q\psi(R)$, or $q^2/(8\pi\epsilon_0 R)$. 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

$$\left(\frac{q}{e}\right)^2 = \left(\frac{10^3 \text{ kg}}{\text{m}^3} \frac{6 \cdot 10^{23}}{0.018 \text{ kg}} \times \frac{4\pi}{3}(10^{-3} \text{ m})^3 \times 0.01\right)^2 = 1.9 \cdot 10^{36}.$$

Multiplying by $2.3 \cdot 10^{-28}$ J m and dividing by $2R$ yields about $2 \cdot 10^{11}$ 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 very high accuracy. Later, however, we’ll see that things look different in the nanoworld.

6.4 CYLINDRICAL CONDUCTORS IN VACUUM

Consider a long, straight metal cylinder (“wire”) carrying linear charge density $\rho_q^{(1D)}$ (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:² $\psi(r, \varphi, z) = B \ln(r/r_0)$ outside the cylinder (and uniform inside). Here B is a constant related to $\rho_q^{(1D)}$ and to the radius (thickness) of the wire.

Your Turn 6B

Find that relation.

Changing the radius r_0 just adds a constant to the potential.

Next, consider *two* long, parallel cylinders with charge densities $\pm\rho_q^{(1D)}$. 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 $\rho_q^{(1D)}$, 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

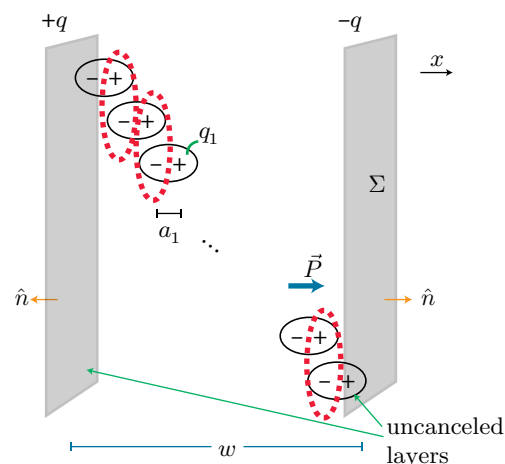
Now imagine filling the gap between conductors with nonpolar atoms or molecules, maybe liquid argon, or more prosaically some kind of oil. What matters is that there be no free charges, so that the material is an insulator. In this context such a material is generically called a **dielectric**.

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.³ Figure 6.1 suggests that the resulting uniform polarization density will lead

²See Your Turn 5A (page 68).

³Recall Section 3.7.4 (page 46).

Figure 6.1: Dielectric medium. Polarizable “molecules” with density ρ_{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 surface charge density can be expressed as $\sigma_b = \hat{n} \cdot \vec{P}$, because the outward-pointing perpendicular is $\hat{n} = -\hat{x}$. Similarly, on the right side there is again a partial cancellation of free and bound charges.



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 ρ_{stuff} . Then the uncanceled net charge forms a thin “bound” layer at the interface, with areal density

$$\sigma_b = \hat{n} \cdot \vec{P}, \quad (6.4)$$

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 away from the medium.⁴ We will refer to σ_b as the **bound surface 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 σ_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 charge we put there.

Most dielectric materials have zero polarization in the absence of an externally applied field. So it’s natural to suppose that it 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.⁵ It is traditionally expressed as $\epsilon_0 \chi_e$, where the dimensionless constant χ_e is called the **dielectric susceptibility**. The relation

$$\vec{P} = \epsilon_0 \chi_e \vec{E} \quad (6.5)$$

is our first example of a **linear response function**. Unlike laws of Nature, it is approximate (for example, we assumed the response was linear in the field strength) and nonuniversal (different materials will have different values of χ_e).⁶

⁴Section 6.6 will look at the general situation, where the polarization density may be nonuniform.

⁵Section 3.7.4 introduced a single-molecule polarizability α . Sections 6.5.2–6.5.3 will relate these quantities.

⁶We also assumed that the induced polarization points parallel to the applied field. Section 13.3.3 will introduce materials that don’t obey that assumption, and Chapter 50 will explore interesting phenomena that arise in that case.

Applying the electric Gauss law to the total charge at the left plate gives

$$\vec{E}_x = (\sigma_f + \hat{n} \cdot (\epsilon_0 \chi_e \vec{E})) / \epsilon_0 \quad \text{where} \quad \hat{n} = -\hat{x}. \quad (6.6)$$

The **electric displacement** is defined as⁷

$$\vec{D} = \epsilon_0 \vec{E} + \vec{P} \quad (\text{generally}). \quad (6.7)$$

Solving Equation 6.6 gives

$$\vec{D}_x = \sigma_f. \quad (6.8)$$

So in the context of our specific model, we have

$$\vec{D} = \epsilon \vec{E}, \quad (\text{linear isotropic medium}) \quad (6.9)$$

where the **permittivity** ϵ of the medium⁸ is

$$\epsilon = (1 + \chi_e) \epsilon_0. \quad (6.10)$$

Thus, the effect of the medium is simply to replace the vacuum permittivity ϵ_0 by a *larger* effective value ϵ . Instead of accounting explicitly for every charge in the medium, we can simplify by *forgetting* it and making this one substitution. Equation 6.9 is called a **constitutive relation** for the material in the capacitor.

The same argument as earlier now gives capacitance as

$$C = \epsilon \Sigma / w, \quad (6.11)$$

which is greater than the vacuum value.

T2 Section 6.5.1' (page 89) mentions some generalizations of the phenomena discussed here.

6.5.2 An energy puzzle

Can we still maintain our idea of energy as stored in the space between the plates? At first it looks bad: Our previous formula gave $\frac{1}{2} \epsilon_0 E^2$. We could minimize this expression by assuming enough polarization to completely neutralize the applied charge, and hence get *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 final polarization must involve optimizing the *total* energy (field plus deformation).

⁷To understand this name, notice that the second term of this expression really involves the movement of charges in the dielectric. Maxwell initially imagined the first term as having a similar origin, a "displacement" of charge in the æther.

⁸Many authors use the notation advocated here. Beware, however, that some older works write the permittivity as $\epsilon \epsilon_0$, so for them the symbol ϵ is what we would call ϵ / ϵ_0 , a *dimensionless* quantity often called the **dielectric constant**. To avoid confusion, we will not introduce any symbol for dielectric constant.

To keep things simple, this section will temporarily make the unjustified assumption⁹ that each dipole responds to the applied electric field \vec{E} . 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, $a_1 = q_1 \vec{E}_x / k_1$. Again suppose that the polarizable objects are distributed with density ρ_{stuff} .

Molecular polarizability leads to bulk susceptibility.

Your Turn 6D

Show that in this model, $\vec{P}_x = q_1^2 \vec{E}_x \rho_{\text{stuff}} / k_1$, and so

$$\epsilon_0 \chi_e = q_1^2 \rho_{\text{stuff}} / k_1$$

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):

$$\begin{aligned} \mathcal{E}/(\text{volume}) &= \frac{1}{2} \epsilon_0 \vec{E}_x^2 + \frac{1}{2} k_1 a_1^2 \rho_{\text{stuff}} = \frac{1}{2} \left(\epsilon_0 + \frac{q_1^2}{k_1} \rho_{\text{stuff}} \right) \vec{E}_x^2 \\ &= \frac{1}{2} (\epsilon_0 + \epsilon_0 \chi_e) \vec{E}_x^2 = \frac{1}{2} \epsilon \vec{E}_x^2 \end{aligned} \quad (6.12)$$

But Equations 6.8–6.9 give $\vec{E}_x = \sigma_f / \epsilon$, and (using Equation 6.10),

$$\mathcal{E}/(\text{volume}) = \frac{1}{2} \frac{\sigma_f^2}{\epsilon}. \quad (6.13)$$

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 $\epsilon > \epsilon_0$). The energy is also lower than it would have been if the material had polarized enough to eliminate the electric field altogether. But it's not *zero*, as suggested at the start of this section!

Another key point about Equation 6.13 is that once again *stored energy is proportional to volume*. Nobody is surprised that the elastic part of the energy has this property—the polarizable objects are spread through space at fixed density—but we already showed that the electric term *also* has that property.

Instead of using the Hooke law, we could have left a_1 arbitrary. Then Equation 6.12 has two terms that are analogous to a mechanical system: two springs in series. We know that that system minimizes 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.

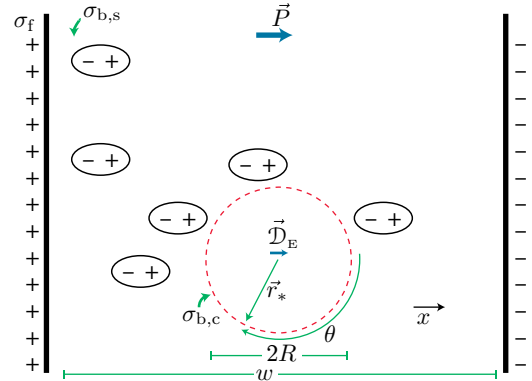
A charged capacitor will pull dielectric material into its gap.

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.¹⁰

⁹Section 6.5.3 will justify this when the dielectric is of low density, and will give an improved derivation for dense matter.

¹⁰See Problem 6.4.

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.



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 $1/\rho_{\text{stuff}}$. After all, surely it is foolish to insist on a continuum distribution below the molecular scale.¹¹

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}_{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 find all of the bound charge densities and the average 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 $\sigma_{b,p} = (-\hat{x}) \cdot \vec{P} = -b$.

Let \hat{r}_* be the unit vector from the dipole 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

¹¹More 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.

surface and “outward” (away from the bulk material) is $-\hat{r}_*$, and the bound surface charge at the cavity is $\sigma_{b,c} = (-\hat{r}_*) \cdot \vec{P} = -b \cos \theta_*$, where θ_* is polar angle measured from \hat{x} . The figure illustrates why the cosine factor is needed: For example, at $\theta_* = \pi/2$ 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:

$$\vec{E}_{\text{plate}} = \frac{\sigma_f + \sigma_{b,p}}{\epsilon_0} \hat{x} = \frac{\sigma_f - b}{\epsilon_0} \hat{x}. \quad (6.14)$$

The other contribution, \vec{E}_{cav} , comes from $\sigma_{b,c}$. To find it, first use the potential formula Equation 2.6 (page 30)

$$\begin{aligned} \psi_{\text{cav}}(\vec{r}) &= \frac{1}{4\pi\epsilon_0} \int_{\text{sphere}} d^2\Sigma \frac{\sigma_{b,c}(\vec{r}_*)}{\|\vec{r} - \vec{r}_*\|} \\ \vec{E}_{\text{cav}}(\vec{r}) &= -\vec{\nabla}\psi_{\text{cav}} = \frac{-1}{4\pi\epsilon_0} \int_{\text{sphere}} R^2 d(\cos \theta_*) d\varphi_* \frac{-b \cos \theta_*}{(-2)\|\vec{r} - \vec{r}_*\|^3} 2(\vec{r} - \vec{r}_*) \\ \vec{E}_{\text{cav}}(\vec{0}) &= \frac{-b}{4\pi\epsilon_0} \frac{R^2}{R^3} \int_{\text{sphere}} d(\cos \theta_*) d\varphi_* \cos \theta_* (-\vec{r}_*). \end{aligned}$$

Only the x component of this vector will survive averaging over φ_* , so

$$= \hat{x} \frac{b2\pi}{4\pi\epsilon_0} \int_{-1}^1 d(\cos \theta_*) \cos^2 \theta_* = \hat{x} b / (3\epsilon_0).$$

The induced dipole moment equals the total field times the molecular polarizability α :

$$\vec{D}_E = \alpha(\vec{E}_{\text{plate}} + \vec{E}_{\text{cav}}(\vec{0})) = \hat{x} \alpha \left(\frac{\sigma_f - b}{\epsilon_0} + \frac{b}{3\epsilon_0} \right).$$

We have now established a connection between the induced moment \vec{D}_E and the strength b of the average polarization \vec{P} . But the same connection applies to every molecule, so we also have

$$\vec{P} = \rho_{\text{stuff}} \vec{D}_E.$$

Combining the last two displayed equations and recalling that $\vec{P} = b\hat{x}$ gives

$$\frac{b}{\rho_{\text{stuff}}} = \frac{\alpha}{\epsilon_0} \left(\sigma_f - \frac{2b}{3} \right).$$

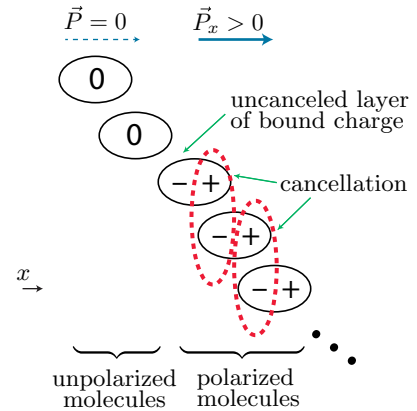
Solving for b gives

$$\vec{P} = \hat{x} \sigma_f \left(\frac{2}{3} + \frac{\epsilon_0}{\alpha \rho_{\text{stuff}}} \right)^{-1}.$$

Now compare the last formula for \vec{P} to Equation 6.14 to find

$$\vec{P} = \vec{E}_{\text{plate}} \frac{\epsilon_0}{\alpha \rho_{\text{stuff}} - \frac{1}{3}}.$$

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 $-\partial\vec{P}_x/\partial x < 0$.



Writing this as $\epsilon_0\chi_e\vec{E}_{\text{plate}}$ at last gives the dielectric susceptibility in terms of molecular polarizability:

$$\chi_e = \frac{\alpha\rho_{\text{stuff}}}{\epsilon_0 - \alpha\rho_{\text{stuff}}/3}. \quad \text{Clausius-Mossotti formula} \quad (6.15)$$

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. Equation 6.15 shows that even in this limit, the susceptibility disagrees with the naïve continuum version in Your Turn 6D. However, in the limit of low density the denominator of Equation 6.15 becomes just ϵ_0 and we recover our earlier, naïve, relation Equation 6.5.

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 now happen.

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 σ_b given by

$$\sigma_b = \hat{n} \cdot \vec{P}, \quad [6.4, \text{page 77}]$$

where \hat{n} is the perpendicular unit vector directed away from the medium. At a sharp interior interface between two media, we can substitute the difference in \vec{P} values on either side, but what should we do for an arbitrary nonuniform \vec{P} ?

If \vec{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** $\rho_{q,b}$ involves spatial gradients of the polarization density. The general formula

$$\rho_{q,b} = -\vec{\nabla} \cdot \vec{P} \quad (6.16)$$

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:

$$\vec{\nabla} \cdot \vec{E} = (\rho_{q,f} - \vec{\nabla} \cdot \vec{P})/\epsilon_0.$$

Combining the divergence terms gives

$$\vec{\nabla} \cdot (\vec{E} + \vec{P}/\epsilon_0) = \rho_{q,f}/\epsilon_0.$$

When phrased in terms of the electric displacement (Equation 6.7), this becomes

$$\vec{\nabla} \cdot \vec{D} = \rho_{q,f}. \quad (6.17)$$

We can now specialize to a the case of a linear, isotropic medium (Equations 6.9–6.10), to find $\vec{\nabla} \cdot \vec{E} = \rho_{q,f}/\epsilon$. This formula is the same as the vacuum case, except that the free charge has been effectively reduced by a factor $\epsilon_0/\epsilon < 1$.

We can now quickly generalize earlier formulas. Outside a spherical charge distribution with total free charge q_f , the solution of Equation 6.17 is $\vec{D}(\vec{r}) = \hat{r}q_f/(4\pi r^2)$, and the corresponding electric field is that function divided by ϵ . We can integrate $dq\vec{E}$ from infinity to R , obtaining the work that must be done to increase q by dq :

$$d\mathcal{E} = dq/(4\pi\epsilon R).$$

Then integrating again, from $q = 0$ to q , gives the total electrostatic energy of a charged sphere in an infinite dielectric medium:

$$\mathcal{E} = q^2/(8\pi\epsilon R). \quad (6.18)$$

This is the same as the vacuum result (Your Turn 6A(a)), but reduced by ϵ_0/ϵ .

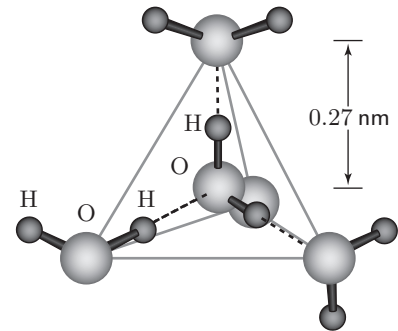
6.7 CHARGE NEUTRALITY BREAKS DOWN ON THE NANOSCALE

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 75 explains why people said that in first-year physics.

Your Turn 6E

- Repeat the calculation for an object of radius $R = 1 \mu\text{m}$ suspended in water. Recall that the static permittivity ϵ of water is about 80 times bigger than the value for air used in the Example.
- Repeat for an $R = 1 \text{ nm}$ object in water.
- Compare both answers to the thermal energy scale $k_B T_r$, where T_r is room temperature, 298 K, and comment.

Figure 6.4: Tetrahedral arrangement of water molecules in an ice crystal. In liquid water, the immediate neighborhood of any one molecule is similar to this, though with some randomness added by thermal motion. The sum of all electric dipole moments in a unit cell is $\vec{0}$, but an applied electric field can bias the distribution of orientations, polarizing the medium. The *gray outline* of a tetrahedron is just to guide the eye. *Dashed lines* are hydrogen bonds that stabilize the structure.



The electrostatic self-energy of an object inside a medium is also called its **Born self-energy** after M. Born.

A dielectric medium reduces the electrostatic self-energy of a point charge, potentially allowing neutral objects to dissociate into ions.

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

A fluid consisting of polar molecules can be enormously polarizable.

So far, we have considered molecules that have no intrinsic dipole moment, 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 simply align to create net polarization.¹²

Polarizability of a polar fluid is limited by entropy, and hence is temperature-dependent.

So once again we face a puzzle: Won't this system always cancel an applied \vec{E} , at least up until the molecules had 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.¹³ Although the net static polarizability is therefore not infinite, for water at room temperature it is quite high: $\epsilon \approx 80\epsilon_0$. 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 at room temperature the effective spring constant for deformation is much stiffer than the one for alignment.

Polarizability can also change dramatically upon phase transition.

The reorientation of molecules in liquid water is accompanied by frictional loss as they rub against their neighbors. When \vec{E} oscillates at microwave frequency, the

¹²Section 3.7.3 (page 45) 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.

¹³You'll work out a quantitative approach in Problem 6.11.

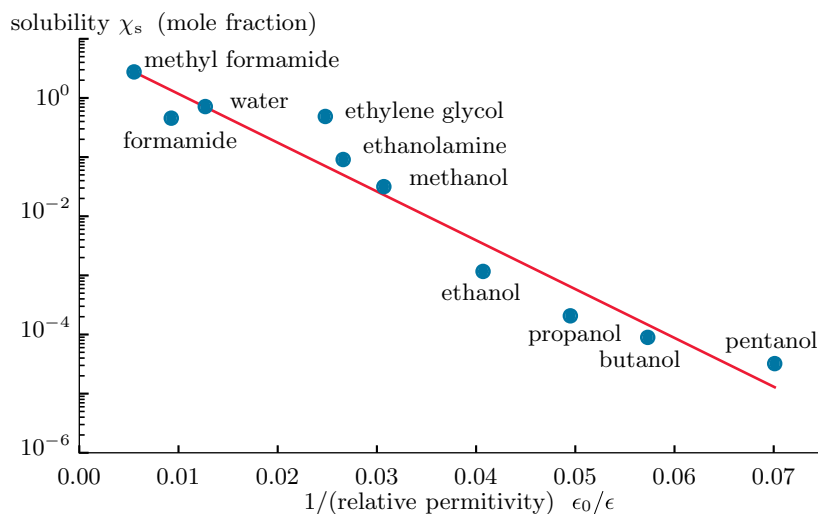


Figure 6.5: [Experimental data.] Semilog plot of the **solubility of sodium chloride** in solvents of different static dielectric constants. Although these chemicals have many specific features, the solubility over a huge range is largely explained in terms of a single characteristic, the solvent permittivity ϵ . The line goes to maximum solubility at $\epsilon \rightarrow \infty$. [Data from Israelachvili, 2011.]

associated heating can be considerable, and indeed, you know that a microwave oven heats liquid water, with its strong and mobile dipoles, much faster than it does glass, plastic, or even ice.¹⁴

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 closer to that of ice in the optical range than at lower frequency. Later, we will find that the polarizability of a medium slows the transmission of light, and indeed, the velocity of light in liquid water is substantially slower than in vacuum (3/4 as fast).

T2 Section 6.8' (page 89) 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 electrostatic attraction (Section 3.7.3, page 45). 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

¹⁴Food contains salty water, which is a conductor. The electric fields in the applied microwaves therefore also induce currents, which give rise to additional heating by the usual resistive mechanism.

the $1/\epsilon$ factor in the Born self-energy (Your Turn 6E). Figure 6.5 indeed shows a strong inverse correlation between the dielectric constant 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 $e^{-\epsilon/k_{\text{B}}T}$. You'll follow up this observation in Problem 6.12.

6.9.2 Partitioning at a fluid interface or cell membrane; permeability

Ions partition unequally across a liquid–liquid interface based on the fluids' polarizabilities.

Next, imagine an oil–water interface. An ion, for example Na^+ , is dissolved in the water ($\epsilon \approx 80\epsilon_0$). Suppose that the ion crosses the interface to the oil side ($\epsilon \approx 2\epsilon_0$). The low 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.

Although they are thin, artificial bilayer membranes are nearly impermeable to ions.

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.¹⁵

Cell membranes have large capacitance per area.

Hence, *cell membranes are electrically insulating*, despite being so thin. Because of that thinness, such membranes also have very high capacitance per unit area (Equation 6.11).¹⁶ The passage of ionic current into or out of a cell can take place only via **ion channels**, water-filled passages embedded in the membrane. Chapters 11–12 will show how the interplay of high capacitance 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. Although water molecules are polar, unlike naked ions they are neutral, and hence have lower Born self-energy in a membrane than small ions.¹⁷

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 2.5): \vec{E}_{\perp} can jump at such an interface, because free charges (in a conductor) or bound charges (in one or both dielectrics) can be localized at the surface.

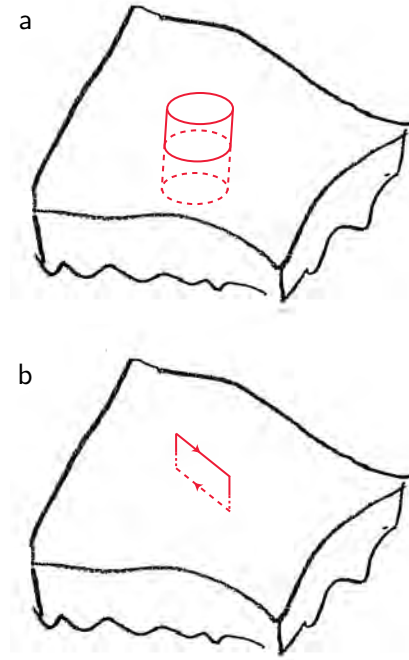
For example, suppose that a dielectric material *1* faces vacuum or air, and let \hat{n} be the perpendicular to a point on the surface that points away from the material. Suppose that there is no free surface charge; for example, the dielectric could have been neutral before an external field was applied. Then Equation 6.4 and the electric

¹⁵ [T2] Screening by salt reduces the self-energy still further on the water side.

¹⁶ Chapter 9 will discuss how this prediction was confirmed experimentally.

¹⁷ The energy of a water's dipole field in the nonpolar membrane (and other nonpolar environments) does contribute to the **hydrophobic effect**, along with other contributions involving hydrogen bonds. But this exclusion is not as strong as that due to the Born energy of a small ion.

Figure 6.6: [Sketches.] **Boundary conditions near a conductor.** (a) The short *red cylinder* has one end cap just outside a conductor and the other just inside. 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 charges at the surface. (b) The *red rectangle* has one of its longer edges just outside a conductor and the other just inside. 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.



Gauss law give that (Figure 6.6a)

$$\hat{n} \cdot (\vec{E}^{[\text{vac}]} - \vec{E}^{[1]}) = \hat{n} \cdot \vec{P}^{[1]}/\epsilon_0. \quad (6.19)$$

If we know the polarization in terms of the electric field, for example via $\vec{P}^{[1]} = \epsilon_0 \chi_e \vec{E}^{[1]}$, then we get a condition for how \vec{E}_\perp jumps. Rephrasing in terms of the displacement (Equation 6.7, page 78) gives the simple form

$$\Delta \vec{D}_\perp = 0. \quad \text{dielectric boundary} \quad (6.20)$$

Turning now to the components of \vec{E} that are tangential to the surface, integrating both sides of $\vec{\nabla} \times \vec{E} = \vec{0}$ 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):

$$\Delta \vec{E}_\parallel = 0. \quad \text{dielectric boundary} \quad (6.21)$$

In particular, these two components must equal zero just outside a conductor, because there is no electric field inside.


FURTHER READING


Semipopular:


Jorgensen, 2021.

Intermediate:

See also Pollack & Stump, 2002, chap. 4 and 6.

 Electrets: en.wikipedia.org/wiki/Electret.

 Piezoelectricity: en.wikipedia.org/wiki/Piezoelectricity.

 Electrostriction: en.wikipedia.org/wiki/Electrostriction.

Bioelectricity, Coulter counter: Grodzinsky, 2011.

Technical:

Piezoelectric dressings to promote wound healing: Long et al., 2018.

T₂

6.5.1' Ferroelectricity, 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 deformation is not imaginary; it 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).

T₂

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 applied this insight to living bacteria, polarizable micrometer-scale objects. 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.

PROBLEMS

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 $\Delta\psi$ is their mutual capacitance (Equation 6.1). Using the electric Gauss law, calculate the capacitance of:

- Two concentric conducting spheres with radii a and b .
- Two concentric conducting cylinders of length L that is large compared to their radii a, b . 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 $0.5 \mu\text{F}/\text{cm}$?

6.2 *Twinlead cable*

Two long, cylindrical conductors of radii a, b are parallel, with centerlines separated by distance w , which is much bigger than either a, b . Long ago, such “twinlead” cables were used as waveguides to bring television signals into the tuner from an antenna.

- Let $c = \sqrt{ab}$ and show that the capacitance per length is approximately proportional to $(\ln(w/c))^{-1}$. Find the constant of proportionality.
- Now suppose that $a = b$. What diameter wire would be necessary to obtain $0.1 \text{ pF}/\text{cm}$ if the separation is $w = 5 \text{ mm}$? (pF means 10^{-12} 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 $R = 17 \text{ cm}$ contains about one mole of gas.

- 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.
- 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 $10^5 \text{ N}/\text{m}^2$.

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 ϵ and mass density ρ_m that we can look up.

Above the fluid there is vacuum. When a fixed potential difference ψ_0 is established between the plates, the fluid between the plates is observed to rise to a higher level Δh above the surrounding fluid.¹⁸ We’d like to know why the fluid rises against the force of gravity, and what determines the equilibrium Δh_* .

¹⁸See Media 2.

- 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 ψ_0 is always fixed (for example, to 1.5 volt 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 Δh 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 Δh_* will depend on the given parameters of the system: the fixed potential ψ_0 , geometry L, W, a , and fluid characteristics, as well as on relevant constants of Nature (ϵ_0 and the acceleration of gravity at Earth's surface). Moreover, argue that Δh_* will *not* depend on W nor on L . Then see how much you can predict about Δh_* just from dimensions.

Now it is time to write some equations and see if your answers to (a,b) are confirmed. For any value of Δh , 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 Δh_* in terms of fixed parameters.
- e. Substitute some rough numbers appropriate to a classroom demonstration and see what value your formula predicts for Δh_* : $\epsilon = 2.5\epsilon_0$ for oil. $a = 5 \text{ mm}$. $\rho_m = 10^3 \text{ kg/m}^3$. $\psi_0 = 10^4 \text{ volt}$. $W = L = 6 \text{ cm}$.

6.5 Biocapacitor

- a. Show that the electric field outside a line of charge in vacuum is $\vec{E} = \hat{r}\rho_q^{(1D)}/(2\pi\epsilon_0 r)$. 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. $\rho_q^{(1D)}$ 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 cylindrical "output line" called the axon. It's got a good conductor inside (**axoplasm**) and outside (salt water), separated by a thin insulating layer (cell membrane). The insulating layer has permittivity ϵ , which may be different from ϵ_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, namely, the parallel-plate capacitor formula $C = \Sigma\epsilon/w$. Here Σ 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 $2\ \text{nm}$.

6.6 Microwave heating

[Not ready yet.]

6.7 Measure ϵ_0

If you know about fringe fields, neglect them in this problem (that is, pretend the plates are infinite).

- 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 $w \ll r$).
- A second such plate is held close to, and parallel to, the first one, and carries total charge $-q$. Find the force $d\vec{f}$ on each surface area element dA of the second plate due to the charge on the first plate.
- Find the electric field strength and the electrostatic potential drop $\Delta\psi$ between the plates as a function of their separation w .
- 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.
- 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 $\Delta\psi \approx 1055\ \text{volt}$. Use this information to determine the approximate numerical value of ϵ_0 (that is, *don't* use the standard value listed in books). [Note: The plates were in air, so really you are finding the dielectric susceptibility ϵ_{air} . But air is similar to vacuum.]
- Compute $1/\sqrt{\epsilon_0\mu_0}$ based on your answer to (d) and the standard value of μ_0 , and comment.

The plates of a charged capacitor feel an attractive mechanical force.

6.8 Mechanical analogy

- 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.
- Follow the analogy introduced in Section 6.5.2 (page 78): The extension of one spring corresponds to minus the bound charge on a dielectric inserted into a capacitor. The extension of the other spring corresponds to total (free plus bound) charge. The sum of those quantities is constrained: Free charge is imposed on the system. Explain why the total stored energy of the spring system corresponds to that of the capacitor system, and why there will again be a linear relation between potential difference and free charge.
- Describe the limits of unpolarizable medium, and of infinitely polarizable medium, in the spring language. Map your answer to (a) over to the capacitor problem and comment.

6.9 Permittivities of various polar liquids

Here are the permittivities of some common substances, all of which have similar mass density, relative to vacuum:

Methanol, $\text{CH}_3\text{-OH}$, has $\epsilon/\epsilon_0 = 33$.

Ethanol, $\text{CH}_3\text{-CH}_2\text{-OH}$, has $\epsilon/\epsilon_0 = 24$.

1-propanol, $\text{CH}_3\text{-(CH}_2)_2\text{-OH}$, has $\epsilon/\epsilon_0 = 20$.

All four of these molecules have about the same dipole moment per molecule, about 1.7 debye, where $1\text{debye} \approx (0.021\text{ nm})e$. Explain qualitatively the differences in permittivities.

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:

Material	Separation, cm	Measured capacitance, nF
Air	6	0.040
Air	1	0.054
Air	0.1	0.34
Paper	3.3	0.058
Liquid water	3.8	2.45
Ice	3	0.1

Make some appropriate quantitative and qualitative comments on these data.

6.11 How to measure molecular dipole moments

Section 6.8 (page 84) 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 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, $\langle \vec{\mathcal{D}}_{\text{E},z} \rangle$, in terms of the field strength E , the fixed magnitude \mathcal{D}_{E} , and temperature.
- Do the integral.
- Simplify your result by considering the limiting case of weak applied field, so that $\langle \vec{\mathcal{D}}_{\text{E},z} \rangle \rightarrow 0$, 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.)
- For a low-density gas, each molecule responds to the imposed field (the effect of other molecules is small). Thus, ϵ differs only slightly from ϵ_0 and the relation between dielectric constant and molecular dipole moment is simple: Combining Equations 6.5 and 6.10 gives

$$\left(\frac{\epsilon}{\epsilon_0} - 1\right) = \vec{P}_z / (\epsilon_0 \vec{E}_z).$$

This formula relates the experimentally determined 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 \mathcal{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 \mathcal{D}_{E} from data. C. Zahn did

A molecular dipole moment can be deduced from the bulk dielectric properties of its gas phase.

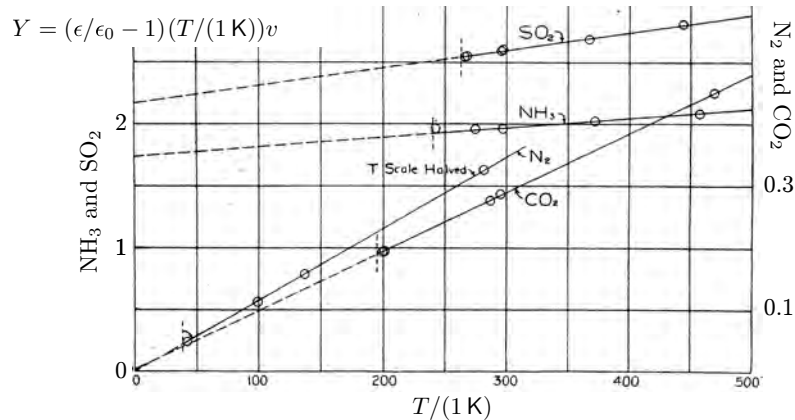


Figure 6.7: [Experimental data.] **Specific polarizabilities of various gases as functions of temperature.** The vertical axis gives permittivity relative to vacuum, multiplied by the temperature and a density factor v described in Problem 6.11. [Data from Zahn, 1926.]

many such experiments. In Figure 6.7, the dimensionless quantity v denotes “specific volume relative to an ideal gas at standard temperature and pressure,” that is,

$$v = \frac{\text{volume}}{\text{molecule}} \bigg/ \frac{k_B T_{\text{STP}}}{p_{\text{STP}}}.$$

Here $p_{\text{STP}} = 1.01 \cdot 10^5 \text{ N m}^{-2}$ is standard pressure and $T_{\text{STP}} = 273 \text{ K}$. In short, v is a constant divided by the number density of gas molecules.

- Find that constant.
- The vertical axis on the graph shows the dimensionless quantity $Y = (\epsilon/\epsilon_0 - 1)(T/(1\text{K}))v$ as a function of temperature. Interestingly, for each gas shown the relation is *linear*. Work out the expected relation between Y and temperature from your earlier result. Can you understand qualitatively the data for ammonia, NH_3 , based on that relation?
- Rather than reading the graph, use these tabulated values from Zahn’s paper to deduce the dipole moment of an ammonia molecule:

	T [K]	Y
first trial	456.9	2.086
second trial	241.7	1.966

- Re-express your answer in the customary unit **debye** $\approx (0.021 \text{ nm})e$, where e is the proton charge.
- 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. John 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 85), 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. Pay more attention to the lower axis, which is linear, than to the upper one.]
- 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, again 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 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 could in principle get a cylinder with closed caps.)

But many other shapes never arise: You never get a cone, or indeed any sort of isolated, sharp point (Figure 7.2e)—neither for open or closed soap films, nor for water droplets.¹ 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 “*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.

¹It is possible to obtain equilibrium soap films with *line* singularities that terminate on corners. We will tacitly exclude these from consideration in this chapter. Also, crystals of frozen water are a completely different matter.

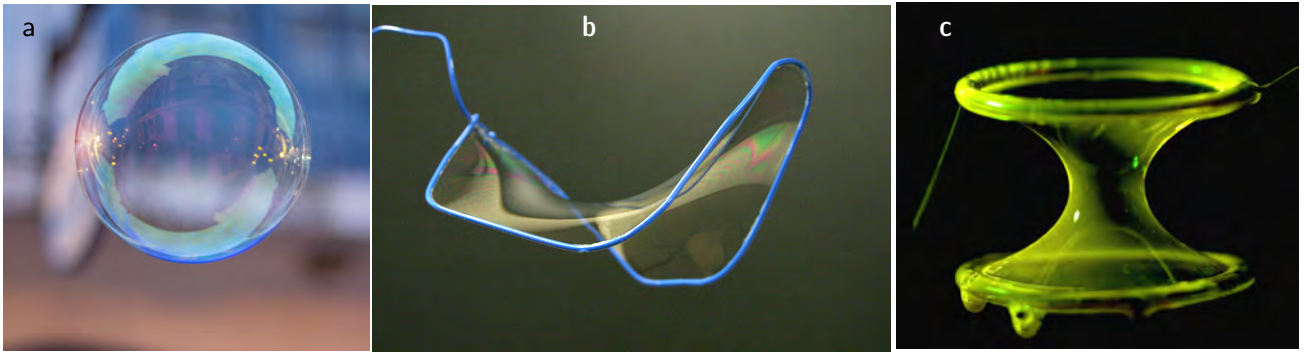


Figure 7.1: Some air–fluid–air interfaces in mechanical equilibrium. (a) Closed (distinct inside and outside regions). (b) Open. (c) Open. [(b) Photo by J. Jacobsen. (c) Photo by R. E. Goldstein, A. Pesci, and K. Moffatt.]

7.2 SOME GEOMETRY OF CURVES AND SURFACES

The next sections introduce many symbols, which are summarized here for reference:

s	arc length coordinate along a curve $\vec{\ell}$
$\vec{\ell}(s)$	curve presented in parametric form
$\Delta(s)$	deviation from tangent line
κ	curvature of a curve in a plane
ξ	small perpendicular displacement of a curve or surface
T	interfacial tension of a fluid–fluid interface or free film
$\Delta T = T_{\text{in}} - T_{\text{out}}$	jump in surface tension across a 1D barrier
F	line tension
L	total arc length of a curve in a plane
u, v	local coordinates centered on a point \mathbf{P}
$\vec{r}(u, v)$	surface presented in parametric form
B_{ij}	description of surface shape near a point; k_i , its eigenvalues when expressed in normal coordinates
$H = (k_1 + k_2)/2$	mean curvature of a surface in 3-space
$G = k_1 k_2$	Gauss curvature of a surface in 3-space
$\Delta p = p_{\text{in}} - p_{\text{out}}$	jump in fluid pressure across a 2D interface
$N(r), M(\theta)$	functions used in separation of variables
θ_0	polar angle for a cone with opening angle $2(\pi - \theta_0)$

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 the curve shown in Figure 7.4a. At the point \mathbf{P} , construct the tangent line as shown. As we walk away from that point, the distance $\Delta(s)$ from the curve to its tangent begins to change as a function of arc length s (unless the line is straight at \mathbf{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 or not at \mathbf{P} . Writing that

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 100). (a–b) A closed soap bubble can reach hydrostatic equilibrium as a surface of constant curvature, possibly confined on 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.

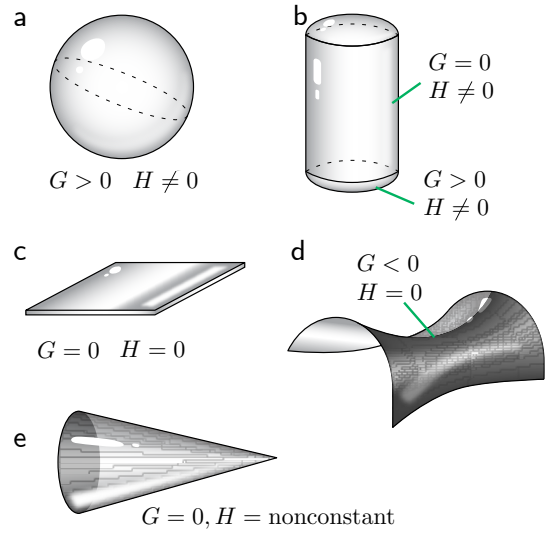
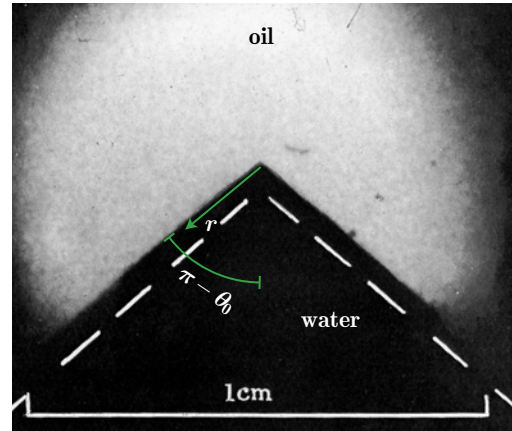


Figure 7.3: [Photo.] **Conical point of an oil–water interface** (side view). The surface with polar angle θ_0 is a cone with half-angle $\pi - \theta_0$. [Adapted from Taylor, 1964.]

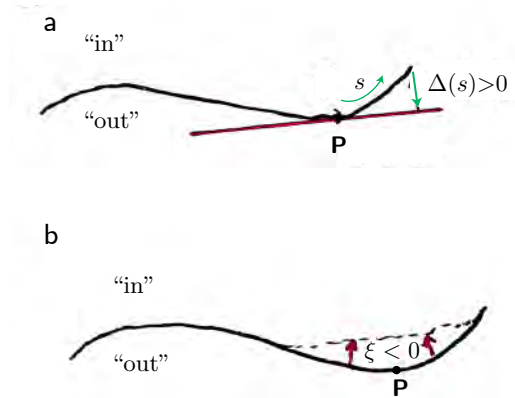


term as $\frac{1}{2}\kappa(\mathbf{P})s^2$, the coefficient $\kappa(\mathbf{P})$ has dimensions \mathbb{L}^{-1} and is called the **curvature** at \mathbf{P} .

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,² because if there’s a bend, “you could instead take a

²In flat euclidean space.

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 \mathbf{P} , after we travel arc length s away from \mathbf{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 perpendicular, and the curvature κ is positive at \mathbf{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 $\xi(s)$. With the same choice of perpendicular as (a), ξ is negative at \mathbf{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.”



shortcut.” To make that more precise, you’ll show in Problem 7.1 that:

- a. To first order in ξ , the total length change is the integral along the curve 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 an amount proportional to the line integral of arc length times ξ (another local formula), but without any factor of curvature. (7.1)

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: It is pulled sideways by 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 not zero, 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 ΔT multiplies $\int ds \xi$, whereas the line tension F multiplies⁴ $\int ds \xi \kappa$.

³Try floating a closed loop of fine thread on water and adding a drop of detergent to the enclosed part of the water surface.

⁴The sign reflects a particular choice of which direction of deviation from the tangent will be called positive (Figure 7.4a,b). [\[T2\]](#)Strictly speaking, interfacial tension involves the *free* energy cost.

In mechanical equilibrium, the net first-order variation of energy must be zero:

$$0 = \int ds (\Delta T + F\kappa)\xi.$$

This relation must hold for any displacement function $\xi(s)$, so:

$$\text{Mechanical equilibrium selects a shape that has constant curvature } \kappa = -(\Delta T)/F. \quad (7.2)$$

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 (κ 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 \rightsquigarrow surface tension T of a soap film, or the interfacial tension of a fluid–fluid interface.
- Interfacial tension difference ΔT between two sides of skimmer \rightsquigarrow pressure difference Δp between sides of our surface.
- Curvature of a curve in a plane \rightsquigarrow ... 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 ξ . Let's proceed as before: At any chosen point \mathbf{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 (for example, latitude and longitude on a sphere). For a sphere, if we measure Δ with respect to the outward-pointing perpendicular, then $\Delta \geq 0$. We'll require that u and v be centered on \mathbf{P} .

- As before, the Taylor series expansion of Δ has no linear terms: That's what tangency means.
- Then to leading nontrivial order, Δ is a quadratic function of the two small excursions. That function equals zero if the surface is flat, so we can use it to describe curvature.
- Write $\Delta(u, v) = \Delta^{[2]}(u, v) + \dots$; quite generally, the quadratic part may be expressed as

$$\Delta^{[2]}(u, v) = \frac{1}{2}(B_{11}u^2 + 2B_{12}uv + B_{22}v^2). \quad (7.3)$$

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 \mathbf{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

$$ds^2 = du^2 + dv^2 + \dots, \quad (7.4)$$

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 \mathbf{P} .

Your Turn 7A

- 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.
- 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

$$\begin{bmatrix} u' \\ v' \end{bmatrix} = \mathbf{S} \begin{bmatrix} u \\ v \end{bmatrix} + \dots,$$

where \mathbf{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 \mathbf{P} , due to the residual coordinate freedom.

Luckily, there is a way out. The quadratic function $\Delta^{[2]}(u, v)$ can be expressed in terms of a matrix $\mathbf{B} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$. Its new form involves a transformed matrix $\mathbf{B}' = (\mathbf{S}^{-1})^t \mathbf{B} \mathbf{S}^{-1}$. But 2D rotations have the special property that $\mathbf{S}^t = \mathbf{S}^{-1}$, so \mathbf{B}' is related to \mathbf{B} via a *similarity transformation*. And any matrix has two famous properties that are invariant under similarity transformations, and hence do not care which local coordinates we chose (as long as they obey Equation 7.4).

In the present context, those invariants are called the **Gauss curvature**, $G = \det \mathbf{B}$ and the **mean curvature**, $H = \frac{1}{2} \text{Tr} \mathbf{B}$. Put differently, the two eigenvalues of \mathbf{B} are called **principal curvatures**. Both are invariant; we just repackage them into $G = k_1 k_2$ and $H = (k_1 + k_2)/2$. 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).

⁵The 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.

Your Turn 7B

Continuing Your Turn 7A, find Earth’s two principal curvatures at your hometown. (What about *my* hometown?)

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 \mathbf{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 \mathbf{P} a variable distance $\xi(\mathbf{P})$ perpendicular to the surface.⁶ We can now state the result we need, analogous to Idea 7.1:

- a. *To first order in perpendicular displacement ξ , the total area change is the integral over the surface 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 integral over the surface of its area element times ξ (another local formula), but without any factor of curvature.*
- (7.5)

We won’t prove Idea 7.5,⁷ 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: Idea 7.5a.

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.⁸ 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

⁶As in Section 7.2.1, our convention is that the displacement is measured along the outward-pointing perpendicular.

⁷[\[T2\]](#) See Section 7.2.4’ (page 108).

⁸A soap film is an air–liquid–air interface. [\[T2\]](#) Again, the interfacial tension is actually the *free* energy cost per area.

⁹[\[T2\]](#) 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

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 $2H = \Delta p/T$ for a closed bubble or fluid–fluid interface.

**Young–Laplace
formula**

(7.6)

Pressure is measured in newtons per meter squared, whereas interfacial tension is in newtons per meter, so Idea 7.6 is at least 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 equals¹⁰ $\epsilon \|\vec{E}\|^2/2$. If we deform the interface, then this 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 to expect a nonconstant electric field in the region near a pointy conductor.¹¹ 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.

Electric field energy leads to a new, nonconstant term in the Young–Laplace formula.

¹⁰See Equation 6.13 (page 79). Detailed derivations must wait for Chapters 35 and 52.

¹¹See 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;
- They make axial symmetry easy to implement; and
- Our boundary condition is simply that the cone with one particular value θ_0 of polar angle must be an equipotential surface (Figure 7.3):

$$\psi(r, \theta_0, \varphi) = 0 \quad \text{for all } r \text{ and } \varphi. \quad (7.7)$$

Following Chapter 5, let us therefore look for potentials of the form

$$\psi(r, \theta, \varphi) = CN(r)M(\cos \theta), \quad \text{where } M(\cos \theta_0) = 0. \quad (7.8)$$

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 obey¹² the radial equation $2rN' + r^2N'' = \lambda N$ for some constant λ . Moreover, we know how N must diverge at $r \rightarrow 0$. The electric field energy involves $\|\vec{\nabla}\psi\|^2$, and our generalized form of the Young–Laplace formula says that its variation must balance the mean curvature, which as mentioned earlier diverges as r^{-1} . So the electric field $-\vec{\nabla}\psi$ must diverge as $r^{-1/2}$, which means that ψ itself, while not infinite, behaves like $r^{1/2}$. Substituting that trial solution into the radial equation shows that the eigenvalue λ 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 69):

$$((1 - \mu^2)M')' = -\lambda M,$$

where now prime indicates $d/d\mu$, and $\mu = \cos \theta$. For integer λ , 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 $r \rightarrow 0$. But we haven't yet enforced the boundary condition Equation 7.7, which also requires $M(\theta_0) = 0$. 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 in Figure 7.3 really is as predicted. This solution is often called the **Taylor cone**.

¹²See Section 5.4 (page 68).

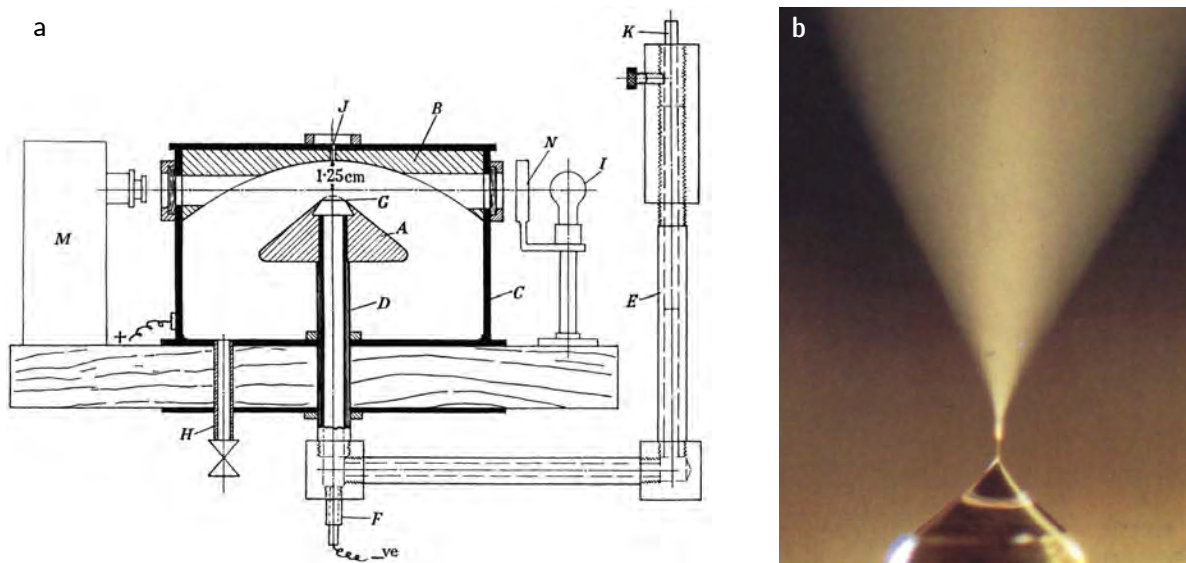


Figure 7.5: Taylor's apparatus. (a) Cross-section, showing the curved electrodes (A, B). A puddle of conducting fluid, G , sits at the top of a truncated metal cone, A . [From Taylor, 1964.] (b) Taylor cone (*bottom*) giving rise to a jet of fluid (methanol with a small amount of hydrochloric acid). [From Pantano et al., 1994.]

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;¹³ 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 . A cross-linked surface, such as the bacterial cell wall, will also resist shear deformation, as well as local changes in area. Other membranes, such as artificial lipid bilayers, have no such shear moduli, but they may 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.

7.5.3 A glimpse of general relativity

Section 7.2.3 took some trouble to characterize a surface using coordinate-invariant local quantities (the scalar fields G and H). Only one of these was needed for the application in this chapter.

But the other curvature G has a remarkable property worth mentioning here. We defined curvature via a procedure involving points *outside* the surface (that is, via the deviation Δ between the surface and its tangent plane). However, the Gauss curvature can be re-expressed solely in terms of distance properties *within* the surface.¹⁴ 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 $\nabla^2 \phi_N$ in the newtonian field equation, and that it also controls the separation of two nearby freely falling bodies.

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.

John Fenn’s Nobel Lecture: www.nobelprize.org/prizes/chemistry/2002/fenn/lecture/.

¹³Chapter 3.

¹⁴Gauss called this fact his “*theorema egregium*” (outstanding theorem).

Intermediate:

Mathematics of curvature: Dubrovin et al., 1992; Spivak, 1999, vol. 2.

Young–Laplace formula: Butt & Kappl, 2018; Safran, 2003; Nelson, 2020, §7.2.2.

Technical:

Taylor cone: Taylor, 1964, p. 392.

T₂

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- $\binom{0}{2}$ tensor field called the metric (also called “first fundamental form”) of the surface.

The quadratic part of the deviation $\Delta^{[2]}$ defines a rank- $\binom{0}{2}$ tensor field, called the “second fundamental form.” We can use the metric to convert it to a rank- $\binom{1}{1}$ 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'a (page 48). 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 $\vec{\mathcal{Q}}_{\mathbb{E}}$. So again its three eigenvalues invariantly characterize different kinds of quadrupole (uniaxial versus biaxial, Section 3.2'a).

T₂

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 \mathbf{P} of interest to us. We can specify the surface near \mathbf{P} by a vector function $\vec{r}(u, v)$, where the parameters range over some region of the uv plane. Let $\hat{n}(u, v)$ be a choice of perpendicular vector at each point of the surface, which we will call “outward” even if the surface is not closed. Then the area of the surface can be written as

$$\Sigma = \int du dv \left\| \frac{\partial \vec{r}}{\partial u} \times \frac{\partial \vec{r}}{\partial v} \right\|. \quad (7.9)$$

Abbreviate $\partial/\partial u$ by ∂_u and so on, and let J denote the square of the integrand above. Thus,

$$J = \|\partial_u \vec{r}\|^2 \|\partial_v \vec{r}\|^2 - (\partial_u \vec{r} \cdot \partial_v \vec{r})^2.$$

A new surface is specified by a perpendicular displacement function ξ via $\vec{r}(u, v) + \hat{n}(u, v)\xi(u, v)$. Suppose that ξ equals zero at the boundary of the chosen region in u, v . The first-order variation of the surface area is then

$$\begin{aligned} \delta\Sigma = \int du dv J^{-1/2} & \left[\partial_u \vec{r} \cdot \partial_u (\xi \hat{n}) \|\partial_v \vec{r}\|^2 \right. \\ & \left. + \partial_v \vec{r} \cdot \partial_v (\xi \hat{n}) \|\partial_u \vec{r}\|^2 - (\partial_u \vec{r} \cdot \partial_v \vec{r})(\partial_u \vec{r} \cdot \partial_v (\xi \hat{n}) + \partial_v \vec{r} \cdot \partial_u (\xi \hat{n})) \right]. \end{aligned} \quad (7.10)$$

Now integrate by parts, using that $\xi = 0$ on the boundary, and write the result as the integral of ξ times some function on the surface. We wish to find a convenient expression for that function at any point \mathbf{P} in terms of the surface shape near that point.

Equation 7.9 is invariant under translations and rotations of \vec{r} , so we may suppose that our 3D coordinates are centered on \mathbf{P} , and moreover, that the tangent to the surface is the xy plane and $\hat{n}(\mathbf{P}) = \hat{z}$. We can also shift the two parameters u, v to center them on \mathbf{P} and make a linear transformation if needed to arrange that the leading terms are

$$\vec{r}(u, v) = u\hat{x} + v\hat{y} - \frac{1}{2} [u, v] \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \hat{z} + \mathcal{O}(3).$$

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

$$\partial_u \vec{r} = \hat{x} - (B_{11}u + B_{12}v)\hat{z} + \mathcal{O}(2), \quad \partial_v \vec{r} = \hat{y} - (B_{12}u + B_{22}v)\hat{z} + \mathcal{O}(2).$$

Hence, u and v are normal coordinates (Equation 7.4, page 101):

$$\|\partial_u \vec{r}\|^2 = 1 + \mathcal{O}(2), \quad \|\partial_v \vec{r}\|^2 = 1 + \mathcal{O}(2), \quad \partial_u \vec{r} \cdot \partial_v \vec{r} = \mathcal{O}(2), \quad \text{and } J = 1 + \mathcal{O}(2).$$

The unit vector perpendicular to the surface is then

$$\hat{n}(u, v) = \frac{\partial_u \vec{r} \times \partial_v \vec{r}}{\|\partial_u \vec{r}\| \|\partial_v \vec{r}\|} = \hat{z} + \hat{x}(B_{11}u + B_{12}v) + \hat{y}(B_{12}u + B_{22}v) + \mathcal{O}(2).$$

Putting it all together, Equation 7.10 becomes

$$\begin{aligned} \delta\Sigma = & - \int du dv \xi \hat{n} \cdot \left[\partial_u (J^{-1/2} \partial_u \vec{r} \|\partial_v \vec{r}\|^2) + \partial_v (J^{-1/2} \partial_v \vec{r} \|\partial_u \vec{r}\|^2) \right. \\ & \left. - \partial_v (J^{-1/2} \partial_u \vec{r} (\partial_u \vec{r} \cdot \partial_v \vec{r})) - \partial_u (J^{-1/2} \partial_v \vec{r} (\partial_u \vec{r} \cdot \partial_v \vec{r})) \right]. \end{aligned}$$

Evaluating the integrand at \mathbf{P} gives

$$-\xi [\partial_u (-B_{11}u + \dots) + \partial_v (-B_{22}v + \dots)] = 2\xi H.$$

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.

PROBLEMS

7.1 *Variation of arc length and area*

- a. A curve in a plane is specified by a vector function $\vec{\ell}(s)$, where s is arc length, $0 \leq s \leq L$. Let $\hat{n}(s)$ be a field of perpendicular vectors all along the curve. A new curve is specified by a function $\xi(s)$ via $\vec{\tilde{\ell}}(s) = \vec{\ell}(s) + \hat{n}(s)\xi(s)$. The displacement function ξ equals zero at $s = 0$ and L .

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 . Establish the formula in Idea 7.1a (page 99). [Hint: Use integration by parts.]

- 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 α is $H(r, \varphi) = (\cot \alpha)/(2r)$. Here r is distance from the cone's apex to the point of interest, and φ is angular position on each "latitude" line. [Hint: If you have difficulty, first draw a very wide cone, with α just slightly less than $\pi/2$. It's nearly a plane, so its curvature must be smaller for 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 98). 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 Δp is everywhere zero and the interfacial tension is fixed to some nonzero constant value T .

- 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 θ_0 and hence $\pi - \theta_0$.
- 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.
- Numerically evaluate the derivative $dM(\mu)/d\mu$ at θ_0 , where $\mu = \cos \theta$.
- Using your results in (a–b), write a formula for the electric field squared just outside the surface ($\theta \lesssim \theta_0$).
- Generalize the Young–Laplace formula (Idea 7.6, page 103) 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).
- Obtain a prediction for the constant C in terms of T and the relative permittivity ϵ/ϵ_0 of oil. Once we look up those values, for example for an oil–water interface, then we learn ψ_{cone} , the potential drop we'd need in an apparatus before we can expect to see a conical singularity.

- g. Evaluate ψ_{cone} at a distance $r = 1 \text{ cm}$ from the point of the cone, using estimated values appropriate for Figure 7.3: $T \approx 3.7 \cdot 10^{-2} \text{ N/m}$ and $\epsilon/\epsilon_0 \approx 2.2$. Is such a potential achievable in the lab?

CHAPTER 8

Charge Flux, Continuity Equation, and Ohmic Conductors

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 quasistatic 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 CONTINUITY EQUATION

Imaging a long, thin pipe with some conserved “stuff” inside. Maybe it’s air, and the “stuff” is mass. Define a 1D density $\rho_m^{(1D)}(t, z)$ (units kilograms per meter). At any z_0 , also define the **1D flux** $j_m^{(1D)}$ as the net rate at which mass crosses the point $z = z_0$, moving from smaller to larger z . Thus, a positively-charged 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 Δz at the rate $j_m^{(1D)}(z_0) - j_m^{(1D)}(z_0 + \Delta z)$. But the rate of pileup is also $\frac{\partial}{\partial t}(\rho_m^{(1D)} \Delta z)$. Dividing through by Δz yields

$$\frac{\partial j_m^{(1D)}}{\partial z} + \frac{\partial \rho_m^{(1D)}}{\partial t} = 0. \quad \text{continuity, 1D} \quad (8.1)$$

Here is a pictorial way to understand Equation 8.1: Imagine a small range of space and time near (t, z) (dashed box in Figure 8.1a). Then conservation of mass¹ implies that the overall mass entering this box is zero:

$$0 = \Delta t \left(\Delta z \frac{\partial j_m^{(1D)}}{\partial z} \right) + \Delta z \left(\Delta t \frac{\partial \rho_m^{(1D)}}{\partial t} \right). \quad (8.2)$$

For example, Figure 8.1a shows three particle trajectories: Point masses *1* and *3'* contribute to the first term of Equation 8.2, whereas *2*, *3*, *1'*, and *2'* contribute to the second term. Because every trajectory that enters the dashed box must also leave it, these terms must sum to zero. Dividing Equation 8.2 by $\Delta t \Delta z$ recovers Equation 8.1.

¹Later, 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.

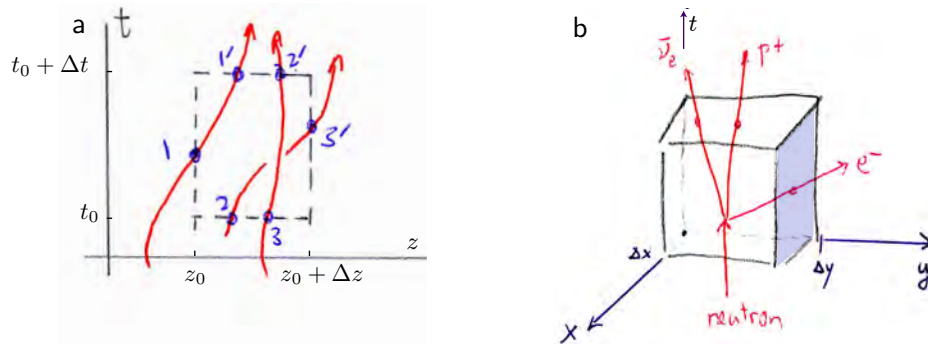


Figure 8.1: Graphical understanding of the continuity equation. (a) Three trajectories in one spatial dimension. (b) A charge-conserving interaction in two spatial dimensions. The *colored face* of the cube shown is transverse to \hat{y} , so any charge traversing it (in this case, the electron) contributes to \vec{j}_2 .

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 to **charge flux**. Also, Figure 8.1b shows a world with two spatial dimensions. We generalize to allow particles to exchange charge, merge, or even explode as shown in the figure. 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 decay² shown, the incoming line has charge zero, whereas the outgoing lines have charges 0, e , and $-e$.

The overall charge entering any fixed region of spacetime, like those shown in the figure, is therefore once again zero. In the neutron decay example, we have:

- 0 (neutron trajectory enters via bottom face of the box);
- $-e$ (proton and neutrino, 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. For trajectories that don’t branch inside the box, it’s even simpler: Everything that enters the box must also exit, carrying its charge.³

Often, it’s reasonable to think of charge as a “river” of many particles, defining an essentially continuous flow. Charge density $\rho_q^{(2D)}(t, \vec{r})$ in 2D has units coul m^{-2} . **Charge flux** $\vec{j}_2^{(2D)}(t_0, x_0, y_0)$ is defined as the net charge per length per time crossing a short line segment of constant $y = y_0$ near position (x_0, y_0) and time t_0 . 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

² [\[T2\]](#) 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.

³ And trajectories that *never* enter the box also never exit it.

of flux, $\vec{j}_1^{(2D)}$, when we consider charge crossing a short line segment with constant x . Overall, $\vec{j}^{(2D)}(t, \vec{r})$ in 2D is a vector field with units $\text{coul m}^{-1}\text{s}^{-1}$.

The overall charge entering the infinitesimal spacetime box in Figure 8.1b is:

- $+\rho_q^{(2D)}(0, 0, 0)\Delta x\Delta y$ from the $t = 0$ (lower) face (plus terms of higher order in Δx and Δy);
- $-\rho_q^{(2D)}(\Delta t, 0, 0)\Delta x\Delta y$ from the $t = \Delta t$ (upper) face;
- $+\vec{j}_2^{(2D)}(0, 0, 0)\Delta x\Delta t$ from the $y = 0$ (left) face;
- $-\vec{j}_2^{(2D)}(0, 0, \Delta y)\Delta x\Delta t$ from the $y = \Delta y$ (right) face;
- $+\vec{j}_1^{(2D)}(0, 0, 0)\Delta y\Delta t$ from the $x = 0$ (rear) face;
- $-\vec{j}_1^{(2D)}(0, \Delta x, 0)\Delta y\Delta t$ from the $x = \Delta x$ (front) face.

As in 1D, these contributions must again sum to zero. Grouping them in pairs and using a Taylor expansion gives

$$0 = \left(-\frac{\partial}{\partial t}\rho_q^{(2D)} - \frac{\partial}{\partial y}\vec{j}_2^{(2D)} - \frac{\partial}{\partial x}\vec{j}_1^{(2D)} \right) \Delta x\Delta y\Delta t. \quad (8.3)$$

(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

$$0 = \frac{\partial}{\partial t}\rho_q^{(2D)} + \vec{\nabla} \cdot \vec{j}^{(2D)}. \quad \text{continuity equation} \quad (8.4)$$

We can do the whole derivation again, with any number of spatial dimensions (for example, three). This time, the relevant definition says that

\vec{j}_1 is the function that, when integrated over $\Delta y\Delta z\Delta t$ 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. (8.5)

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:

$$\frac{dq_{\text{tot}}}{dt} = \frac{d}{dt} \int d^3r \rho_q = \int d^3r (\partial\rho_q/\partial t) = \int d^3r \vec{\nabla} \cdot \vec{j} = 0. \quad (8.6)$$

Not surprisingly, the local conservation of charge that led to the continuity equation implies global charge conservation.

8.4 REMARKS

- Charge flux is sometimes called “current density,” but we will reserve the word “density” to mean only “conserved stuff per unit volume.” In contrast, **flux** will

always mean “conserved stuff per transverse dimensions per time.”⁴ 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.⁵ In 3D, there are two.

- The continuity equation is a purely kinematic identity. It is valid regardless of whether the particle trajectories obey any equation of motion. 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 (perhaps zero), the continuity equation guarantees that \vec{j} is divergence-free.

8.5 NONSTATIC SITUATIONS

8.5.1 Conductivity, resistivity, conductance, resistance

Many materials are insulators:⁶ $\vec{j} = 0$. Some others are approximately **ohmic**: they develop currents via a dissipative law

$$\vec{j} = \kappa \vec{E}. \quad \text{ohmic material} \quad (8.7)$$

The constant κ 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”⁷ as it appeared in first-year physics. To make the connection, consider a thin wire of length h with cross-sectional area Σ . 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,

$$\Delta\psi = IR \quad \text{where} \quad R = h/(\Sigma\kappa). \quad (8.8)$$

The **resistance** R depends on the geometry of the wire (via h and Σ) as well as on the material (via κ). The SI unit for resistance is called ohm and abbreviated Ω . The SI unit for conductivity is then $\Omega^{-1}\text{m}^{-1}$. Another name for Ω^{-1} is the siemens,⁸ abbreviated S.

Other quantities appearing in scientific literature include **resistivity**, defined⁹ as $1/\kappa$, and **conductance**, defined as $1/R$.

⁴Unfortunately, some books use “magnetic flux” to mean something quite different; we will not use that term.

⁵In Figure 8.1b, the colored cube face is transverse to \hat{y} and has one spatial dimension with extent Δx . (In panel (a), the left and right edges have no spatial extent.)

⁶More 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 604).

⁷Discovered 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.” But certain materials, in certain conditions, do have approximately ohmic behavior.

⁸Don’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 \mathcal{U} .

⁹The 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.

Equation 8.7 is called “dissipative” because it relates \vec{j} , a quantity that changes sign under time reversal, to \vec{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

$$\mathcal{P} = (\Delta\psi)I = I^2R = (\Delta\psi)^2/R. \quad (8.9)$$

Indeed, that power ends up as heat, an effect called **Joule heating** or **ohmic heating**. If you plug an appliance with an internal short circuit ($R \lesssim 1\ \Omega$) into the wall ($\Delta\psi$ 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

Electrolysis of salt water releases chlorine, in addition to hydrogen and oxygen.

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 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 chlorine gas.¹⁰ The overall effect is thus that electrons leave the solution into the anode, even though free electrons did not literally pass through the solution.

Ex. A typical diffusion constant for an ion or small molecule in water at room temperature is $D \approx 1\ \mu\text{m}^2/\text{ms}$. What’s the mean velocity of a chloride ion in solution, in an applied electric field of 1 volt/cm?

Solution: The viscous drag coefficient is given by Einstein’s relation, $k_{\text{B}}T/D$. It has the units of force per velocity, so take the force qE on one ion and divide by the drag coefficient:

$$\begin{aligned} qED/k_{\text{B}}T &= (1.6 \times 10^{-19}\ \text{coul})(100\ \text{volt/m})(1\ \mu\text{m}^2/\text{ms})(4\ \text{pN nm})^{-1} \\ &= 1.6 \cdot 10^{-19} \times 100 \times \frac{10^{-12}}{10^{-3}} \frac{1}{4 \cdot 10^{-12} 10^{-9}} \text{coul} \frac{\text{J}}{\text{coul m}} \frac{\text{m}^2}{\text{s N}} \\ &\approx 10^{-5}\ \text{m/s}. \end{aligned}$$

The drift (mean) 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.

¹⁰Another option is for chloride to remain an ion and instead to assist in pulling an electron away from a water molecule, ultimately leading to H^+ ions and liberating neutral oxygen. The situation at the cathode is more complicated: Sodium is too reactive with water to electroplate out. Instead, it remains an ion and assists in adding an electron to a water molecule, ultimately leading to OH^- ions and liberating neutral hydrogen.

8.6 QUASI-STATIC SITUATIONS

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 2.5 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 too fast incurs too much frictional resistance. Combining the continuity equation, the ohmic hypothesis, and the Gauss law yields

$$\frac{\partial}{\partial t} \rho_q = -\vec{\nabla} \cdot (\kappa \vec{E}) = -\kappa \rho_q / \epsilon. \quad \text{spatially uniform, ohmic material}$$

We see that

In a spatially uniform ohmic material, any initial nonuniformity of net charge density gets exponentially suppressed over time scales longer than ϵ/κ . (8.10)

Many nonstatic situations resemble static ones in some respects.

Your Turn 8A

Check that ϵ/κ does have dimensions of time.

Here is another clue that an ohmic material breaks time-reversal invariance: 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 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; indeed, $\epsilon \approx 80\epsilon_0$ at low frequency.¹¹ So for frequencies below about 100 MHz, we can assume that salt water is everywhere locally neutral, and hence also that $\vec{\nabla} \cdot \vec{E} = 0$, just as in electrostatics! This simplification will help us in Section 8.7 and in later chapters.

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 8.2) 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 only study noninvasive experimental ways to detect them.

¹¹See Section 6.8 (page 84).

Figure 8.2: 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 cell. Other tubes represent “input lines” (**dendrites**), each of which communicate with other neurons (including sensory receptors). [Art by D. S. Goodsell.]

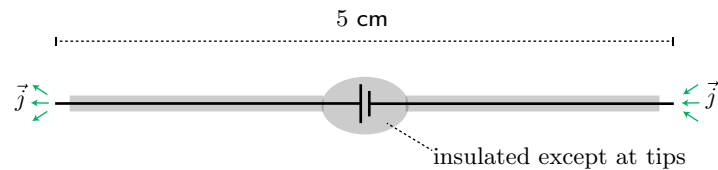
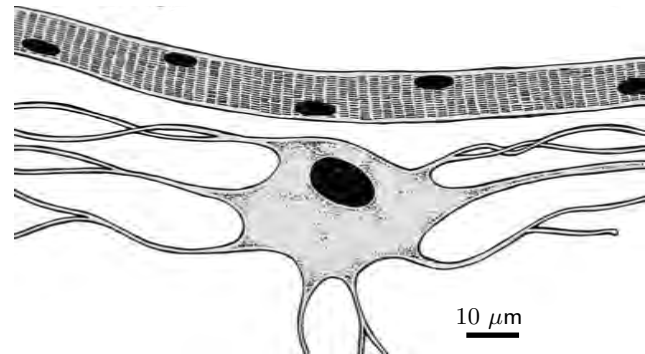


Figure 8.3: A current dipole. When immersed in an ohmic conductor such as salt water, this source sets up a distributed current, and hence an electric field.

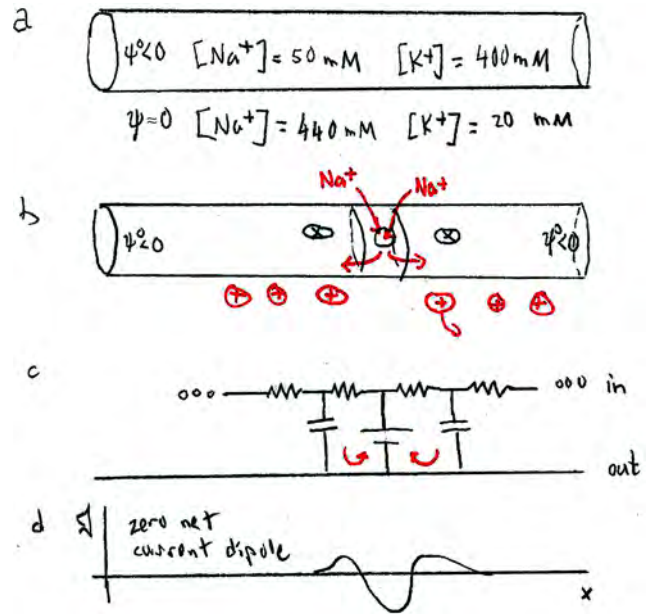
To begin, think about the simple system in Figure 8.3: A battery is connected to two thin wires, each insulated except at their tips; everything is immersed in a salt solution. Each tip is a pointlike boundary between a good conductor and an ohmic medium, so Idea 8.10 does not apply and net charge can build up at the tips. Elsewhere, however, the electric field does obey $\vec{\nabla} \cdot \vec{E} = 0$, 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.

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 8.4a 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

Figure 8.4: 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 insulating 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 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 ψ^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 has net current dipole moment zero.



frustrated by the barrier membrane.¹²

- 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.¹³ 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 an interior electric field from those two layers of charge.¹⁴
- 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.

A traveling nerve impulse or muscle contraction leads to a multipolar electrical disturbance that can be measured from far away.

These nonequilibrium concentrations, enforced by the cell membrane, form a continuously distributed source of free energy, constantly maintained by active transport of sodium out of, and potassium into, the cell.

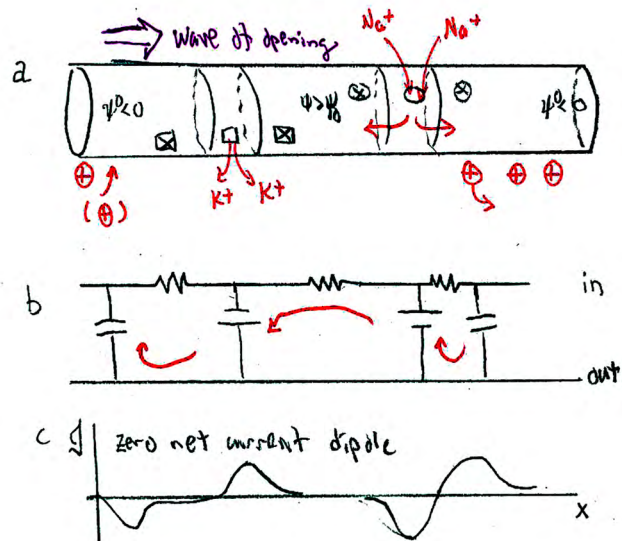
Section 6.9 (page 85) mentioned that cell membranes are studded with doorways, ion channels that, while normally closed, can open upon command, permitting the transport of specific ion types across the membrane. A nerve impulse begins with the opening of ion channels specific for sodium in a small patch of membrane. As sodium

¹²Recall Section 6.9 (page 85).

¹³See Section 10.3.3 (page 139).

¹⁴See Section 6.9 (page 85).

Figure 8.5: Exterior effect of a nerve impulse. A more realistic version of Figure 8.4. (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 then leads to their expulsion in that zone. Later still, all channels close and the membrane repolarizes. So the interior electric potential has its resting (negative) value 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) The linear density of current source seen by the outside world is more complicated than in Figure 8.4, but again has net dipole zero.



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 8.4b).

A real nerve impulse is more complicated than the situation just outlined (Figure 8.5).

- First, 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 8.5a).
- The overall effect is to create a traveling wave of 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 8.5b), adjacent to a *sink* as sodium ions enter. Still farther to the right in the figure, there is a second source (potassium ion outflow), and finally a sink as the membrane capacitance recharges to its pre-impulse value. Together, these changes amount to a complex line source of current \mathbf{J} sketched in Figure 8.5c. 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. This current spreads into the surrounding fluid following the quasi-static rule, Idea 8.10. At any instant, it obeys $\vec{\nabla} \cdot \vec{j} = 0$ with boundary conditions at the axon determined by the form of the nerve impulse. But this equation implies $\vec{\nabla} \cdot \vec{E} = 0$, 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 distribution of point current sources along the axon, but the math is the same as before.¹⁵

8.7.3 Electroencephalogram

Figure 8.5 looks complicated, but we can get its main qualitative feature by remembering charge neutrality. The axon's cross-sectional area Σ and its conductivity κ determine the internal axial current I_x created by the varying interior potential ψ_{in} via the ohmic property of the interior salt solution:

$$I_x = -(\kappa\Sigma) \frac{\partial}{\partial x} \psi_{\text{in}}. \quad (8.11)$$

In an insulated wire, the continuity relation 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, via ion channels, or effectively, by discharging the membrane capacitance (both mechanisms are shown in Figure 8.5). 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 is

$$\mathcal{J} = -\frac{\partial I_x}{\partial x} = +\frac{\partial}{\partial x} \left(\frac{\partial \psi_{\text{in}}}{\partial x} \kappa\Sigma \right). \quad (8.12)$$

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 monopole moment of the current source, $\int dx \mathcal{J}$, equals zero. Moreover, the quantity $x\mathcal{J}$ can be written as

$$(\kappa\Sigma) \frac{\partial}{\partial x} \left(x \frac{\partial \psi_{\text{in}}}{\partial x} - \psi_{\text{in}} \right),$$

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:

$$\int dx x\mathcal{J} = (\kappa\Sigma) \left(x \frac{\partial \psi_{\text{in}}}{\partial x} - \psi_{\text{in}} \right) \Big|_{-\infty}^{\infty} = 0. \quad (8.13)$$

Hence, the leading-order electric field far from the axon is in general of *quadrupole* form (Figure 8.5d).¹⁶ Any one nerve impulse will create extremely small distant currents and fields. However, the concerted firing of impulses on many parallel axons

¹⁵Our picture strictly applies only to an isolated axon in solution. Corrections can be made for the inhomogeneity in the tissues of a complete animal.


¹⁶ Other parts of a neuron, for example its dendrite, may also depolarize, potentially giving rise to a dipole contribution.

Figure 8.6: 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.

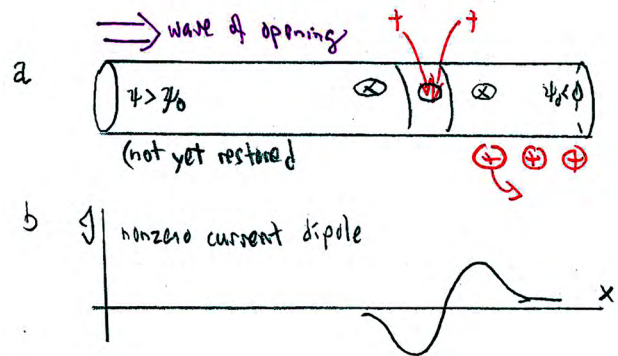
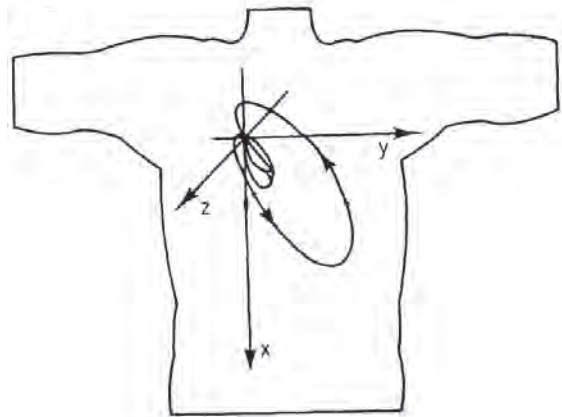


Figure 8.7: Electrocardiogram. The total current dipole vector moves periodically, rotating and stretching with each heartbeat.



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 also triggers the muscle cell to contract; and
- A single depolarization wave spreads over the entire cell for the duration of a contraction (Figure 8.6). 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**, or EKG) but also its spatial direction (vector electrocardiogram, Figure 8.7). 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:

www.youtube.com/watch?v=zG4tMchDTV8&list=PL8sA-GV63nB6AYDS1xLQA8bxom3wgcMD.

Intermediate:

General: Pollack & Stump, 2002, chap. 7.

Quasi-static approximation: Pollack & Stump, 2002, §7.6.

Neuroelectricity, EEG, EKG: Hobbie & Roth, 2015; Benedek & Villars, 2000; P. L. Nunez, R. Srinivasan, *Electric Fields of the Brain: The Neurophysics of EEG*. 2nd ed. Oxford U. Press, New York, 2006.

Einstein relation: Nelson, 2020.

Technical:

Malmivuo & Plonsey, 1995; Gratiy et al., 2017.

PROBLEMS

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 transmembrane potential $\psi(t)$.

- 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 $I = \bar{I} \cos(\omega t)$. 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 electrodes immersed in an infinite bath of poor conductor, 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, κ . The ends are separated by a distance $R \gg R_0$. Find the total DC resistance between the two electrodes as a function of R and comment on the (possibly surprising) form of your answer.

[*Hints:* Start by noticing that the units of conductivity are not the same as those of $1/(\text{resistance})$. Think about the possible forms of the desired formula for resistance as a

function of κ , R , and R_0 , in the stated limit. Next begin the problem by guessing a form for the electrostatic potential in the medium that solves the relevant equations and is approximately constant over each electrode in the stated limit. 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 around the perimeter of the electrode. Consider a thin circular metal disk electrode of radius a and potential ψ_0 surrounded by a bulk medium of conductivity κ . 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, \vec{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 minor axes of the ellipses are much smaller than the major axes.

To define the ellipsoid, let σ be some positive constant (the distance from the center to one focus of an ellipse). Set up cylindrical coordinates ρ, φ, z centered on the center, with \hat{z} the axis of symmetry. Now define¹⁷

$$r_\pm = \sqrt{z^2 + (\rho \mp \sigma)^2} \quad (8.14)$$

and

$$\xi = (r_+ + r_-)/(2\sigma), \quad \eta = (r_- - r_+)/ (2\sigma).$$

The surfaces of constant ξ 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 ξ_0), given that the potential drop between the surface and infinity is ψ_0 . Then the case ξ_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

$$\nabla^2\psi = 0 \text{ for } \xi > \xi_0; \quad \psi \rightarrow 0 \text{ at infinity,} \quad \text{and } \psi(\xi_0, \eta, \varphi) = \psi_0.$$

Thus, the boundary conditions are simple in ellipsoidal coordinates. Let's show that the laplacian is separable in these coordinates.

- a. Invert the preceding formulas to solve for ρ and z in terms of ξ and η . [*Hint:* Express $\xi\eta$ and $(\xi^2 - 1)(1 - \eta^2)$ in terms of ρ and z , then think.] The intersection of a surface of constant ξ 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 η to see the coordinate grid created by ξ and η .

From this point on, all work will be analytic (not numerical). First, promote everything to 3D by expressing x, y, z in terms of ξ, η , and φ .

¹⁷Equation 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").

- 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 very nice property similar to the one found in Section 5.3.2 (page 66) 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 ξ , η , and φ .
- d. Use (b,c) to express the integral $\int d^3r \vec{\nabla} f \cdot \vec{\nabla} g$ in the coordinates ξ , η , and φ . Here f and g are any two functions, both independent of the azimuthal angle φ 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 φ .
- f. What nice property 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.
- 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 ψ . Ensure that your answer has the required behavior at infinity.
- i. Find the charge flux perpendicular to the electrode at its surface (potentially a function of η and φ at fixed ξ_0).
- 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 η).
- [*Note*: One might have worried that the sharp edge of the disk could generate a singularity that gives a pathological answer, such as 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.]

A good conductor with a sharp edge, immersed in a resistive medium, emits a charge flux that is singular on the edge.

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 seawater, and the current source supplies a steady total current $I = 1 \text{ mA}$ (Figure 8.3, page 118).

- What equation governs the steady electric potential throughout the seawater?
- Write down a solution to that equation appropriate to the problem by superposing two simpler solutions.
- The conductivity of seawater is $\kappa \approx 0.1 \Omega^{-1} \text{ m}^{-1}$. Use that fact, and the form of your answer to (b) up close to one electrode tip, to get the overall constant in front of your solution, and hence finish explicitly evaluating the steady electric potential throughout the seawater.

8.5 [Not ready yet.]

CHAPTER 9

Vista: Cell Membrane Capacitance

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* technique.

Knowing that the lipid molecules constituting the cell’s bilayer membrane are similar to other oils then let him predict that the capacitance per unit area would be $\mathcal{C} = \epsilon/\delta$, where ϵ is the permittivity of oil and δ is the membrane thickness.² Thus, knowing the permittivity ϵ and measuring \mathcal{C} would allow a determination of δ .

Although Fricke used egg 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 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.

¹Strutt, 1890.

²See Equation 6.11 (page 78).

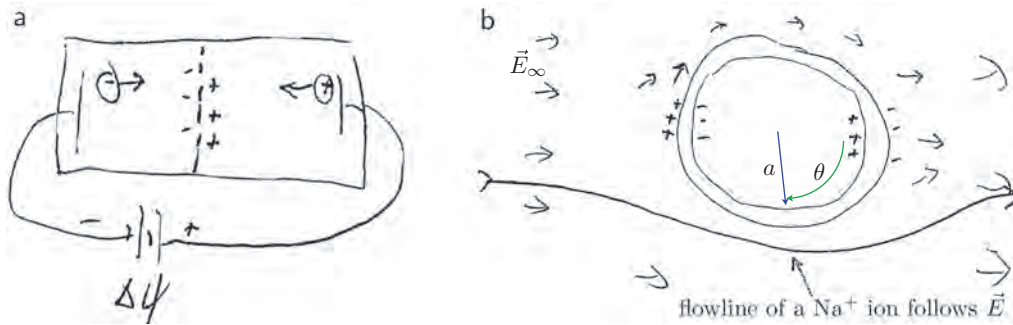


Figure 9.1: Experiments to measure membrane capacitance. (a) Imagined experiment where a bilayer membrane separates two electrodes. (b) In Fricke's experiment, both electrodes were on the same side of the membrane (the cell exterior). Nevertheless, this insulating object disrupted what would otherwise be a uniform flow of current (ions in solution).

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”). Such an approach is possible today 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 clever 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 circular 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.)

In a conducting medium,⁴ $\vec{j} = \kappa \vec{E}$. Because we assume that no current may cross the membrane, we must have $\vec{j}_{\perp} = 0$ at the inner and outer surfaces, and hence $\vec{E}_{\perp} = 0$ 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,⁵ so $\vec{\nabla} \cdot \vec{E} = 0$. Thus, we may write $\vec{E} = -\vec{\nabla} \psi$ as usual, but with a jump in ψ 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 electrostatics problems:

³Section 8.6 (page 117).

⁴See Equation 8.7 (page 115).

⁵Idea 8.10 (page 117).

Inside the cell

$\nabla^2\psi = 0$, subject to $\vec{j}_\perp = 0$ on the boundary; that is,

$$\frac{\partial\psi}{\partial r} = 0 \quad \text{on the spherical surface } r = a.$$

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

$\nabla^2\psi = 0$, subject to $\vec{j}_\perp = 0$ on the boundary; that is,

$$\begin{aligned} \frac{\partial\psi}{\partial r} &= 0 \quad \text{on the spherical surface } r = a, \text{ and} \\ \psi &\rightarrow -E_\infty z = -E_\infty r \cos\theta \quad \text{far away.} \end{aligned} \tag{9.1}$$

We can easily guess one solution to the Laplace equation with the required behavior at infinity, that is,

$$-E_\infty r \cos\theta \tag{9.2}$$

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 multiplying the second solution:⁶

$$\begin{aligned} 0 &= \left. \frac{\partial}{\partial r} \right|_a (-E_\infty r \cos\theta + Ar^{-2} \cos\theta) \\ 0 &= -E_\infty - 2Aa^{-3} \\ \psi_{\text{out}} &= -E_\infty \cos\theta \left(r + \frac{1}{2} \frac{a^3}{r^2} \right) + \text{const.} \end{aligned} \tag{9.3}$$

Match the solutions

By symmetry, no charge piles up near the membrane at the equator, $\theta = \pi/2$. Hence, ψ must not jump as we cross the membrane there. We already found that ψ_{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.

⁶Section 5.4 (page 68) found this solution in a different context (no charge flow).

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 \mathcal{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

$$d\mathcal{E}_{\text{memb}} = \frac{1}{2}(\Delta_{\text{memb}}\psi)^2 dC \quad \text{where} \quad dC = \mathcal{C} d^2\Sigma.$$

We can now find the total stored energy by using Equation 9.3:

$$\begin{aligned} \mathcal{E}_{\text{memb}} &= \int d\mathcal{E}_{\text{memb}} = \int (a^2 d(\cos \theta) d\varphi) \frac{1}{2} (\Delta_{\text{memb}}\psi)^2 \mathcal{C} \\ &= \frac{1}{2} \mathcal{C} \int (2\pi a^2 d(\cos \theta)) (E_{\infty} \cos \theta (a + \frac{1}{2}a))^2 = \frac{1}{2} \mathcal{C} (E_{\infty})^2 2\pi a^4 (3/2)^2 \int_{-1}^1 d\mu \mu^2 \\ &= \frac{3\pi}{2} a^4 E_{\infty}^2 \mathcal{C}. \end{aligned} \quad (9.4)$$

For N well-separated cells in suspension, the total is N times this formula.

Fricke applied alternating voltage $\bar{\psi} \cos(\omega t)$ across his chamber, and measured the resulting current. The current had the same angular frequency ω , so its form was $\bar{I} \cos(\omega t - \phi)$; Fricke therefore measured the dependence of peak current \bar{I} and its phase shift ϕ on $\bar{\psi}$ and ω 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 goal⁷ of extracting the numerical value of the only unknown parameter: the areal density of membrane capacitance, \mathcal{C} .

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 $e\psi(t)$. Thus, the net electric power entering the experimental chamber is⁸

$$\begin{aligned} \mathcal{P} &= \psi I = \bar{\psi} \bar{I} \cos \omega t \cos(\omega t - \phi) \\ &= \bar{\psi} \bar{I} (\cos^2 \omega t \cos \phi + \cos \omega t \sin \omega t \sin \phi). \end{aligned} \quad (9.5)$$

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:

$$-\frac{d}{dt}(N\mathcal{E}_{\text{memb}}) = -N \frac{3\pi}{2} a^4 \mathcal{C} \left(\frac{\bar{\psi}}{L}\right)^2 \frac{d}{dt} \cos^2 \omega t,$$

⁷Problem 8.1 explores Fricke's strategy.

⁸The following derivation is easier with complex exponential notation (Section 18.7, page 266).

where L is the length of the chamber (distance between electrodes).

$$= N3L^{-2}\pi a^4 \mathcal{C}\omega\bar{\psi}^2 \cos \omega t \sin \omega t.$$

Compare that result to the second term of Equation 9.5 to find

$$\bar{\psi}\bar{I} \sin \phi = N3\pi a^4 \mathcal{C}\omega \left(\frac{\bar{\psi}}{L}\right)^2.$$

Rearranging gives the desired result

$$c = \frac{\bar{I}}{\bar{\psi}} \frac{L^2 \sin \phi}{N3\pi a^4 \omega}. \quad (9.6)$$

The formula gives us membrane capacitance in terms of the known cell radius a and count N , the imposed $\bar{\psi}$ and ω , and the resulting \bar{I} and ϕ .

Fricke found⁹ $C \approx 1 \mu\text{F}/\text{cm}^2$. 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.

Historical: Cole, 1972.

Technical:

Historic: H. Fricke. The Electric Capacity of Suspensions of Red Corpuscles of a Dog. Phys. Rev. (1925) vol. 26 (5) pp. 682-687; A Mathematical Treatment of the Electric Conductivity and Capacity of Disperse Systems ii. The Capacity of a Suspension of Conducting Spheroids Surrounded by a Non-Conducting Membrane for a Current of Low Frequency. Hugo Fricke. Phys. Rev. (1925) vol. 26 (5) pp. 678-681;

Cole, KS, 1928. Electric impedance of suspensions of spheres. The Journal of general physiology 12:29–36; Cole, KS, 1928. Electric impedance of suspensions of Arbacia eggs. The Journal of general physiology 12:37–54. Curtis, HJ, and KS Cole, 1937. Transverse electric impedance of Nitella. The Journal of General Physiology 21:189–201.

Much later, similar ideas entered biophysics as “electrical cell-substrate impedance sensing” (ECIS) technology: Wegener, J., C. R. Keese, and I. Giaever, 2000. Electric cell-substrate impedance sensing (ECIS) as a noninvasive means to monitor the kinetics of cell spreading to artificial surfaces. Experimental cell research 259:158–166; Ivar Giaever, et al., 2012. Electric Cell-Substrate Impedance Sensing

⁹See Problem 9.1.

and Cancer Metastasis. Cancer Metastasis - Biology and Treatment 17. Springer Netherlands, 1 edition; Tirupathi, C., A. B. Malik, P. J. Del Vecchio, C. R. Keese, and I. Giaever, 1992. Electrical method for detection of endothelial cell shape change in real time: assessment of endothelial barrier function. Proceedings of the National Academy of Sciences 89:7919–7923; Giaever, I., and C. R. Keese, 1993. A morphological biosensor for mammalian cells. Nature 366:591.

PROBLEMS

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 (also a salt solution). 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 $z = L$.

The goal of the experiment was to measure \mathcal{C} , the membrane capacitance per unit area. But when designing the experiment, we often turn things around and use an *estimate* for \mathcal{C} , in order to predict whether the observed phase lag ϕ between voltage and current will be large enough to measure (for example, with an oscilloscope). Thus, assume $\mathcal{C} \approx 1 \mu\text{F cm}^{-2}$. Here are some other typical numbers extracted from Fricke's original paper:

- Cells of the sort studied by Fricke and Cole have radius $a \approx 3 \cdot 10^{-4} \text{ cm}$.
- The applied current had a circular frequency of 87 000 Hz, or angular frequency $\omega = 2\pi \times 87\,000 \text{ rad/s}$.
- The overall resistance of the seawater in the chamber was $\bar{\psi}/\bar{I} \approx 300 \Omega$.
- The number density of cells in the chamber is such that they occupy about 20% of the chamber volume.
- The chamber dimensions are: cross-section $\Sigma \approx 15 \text{ cm}^2$, length $L \approx 7 \text{ cm}$.

Use these numbers and the analysis in the chapter to find the predicted phase lag angle ϕ 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 spherical cell 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,¹ Cl^- , and leaving the sodium as a positive ion, 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 a force $q\vec{E}$. Initially, then, each charged particle drifts with the net speed $v_{\text{drift}} = qE/\eta$, where η is a constant describing viscous friction. In salt solution there are two ionic species with opposite charge, and hence opposite drift velocities, but for now we only consider one of the species.

¹Negative ions are also called **anions**, because they'd be attracted to an **anode**; similarly, a positive ion is called **cationic**. The terms "cathode," "anode," "ion," "cation," "anion," "electrode," and "electrolyte" were all coined by Michael Faraday.

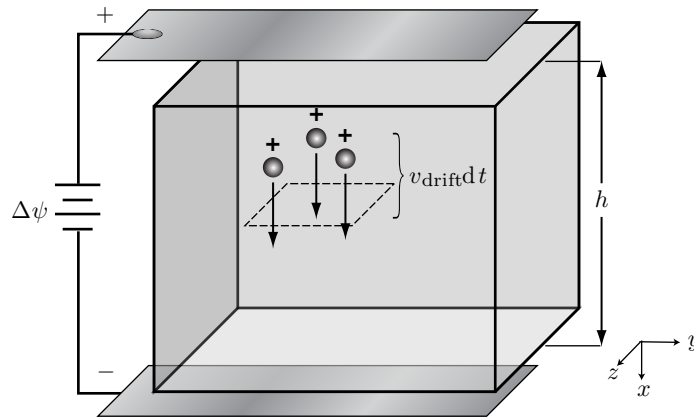


Figure 10.1: [Sketch.] **Origin of the Nernst relation** (Equation 10.3). An electric field pointing downward drives positively charged ions down. Initially after connecting the battery, the number flux \vec{j}_{ion} for the ion species shown points downward with magnitude equal to the number density c_{ion} times v_{drift} . The corresponding contribution to charge flux is $q\vec{j}_{\text{ion}}$. Eventually the system comes to equilibrium with a downward density gradient of positive ions (and an upward gradient of negative ions, not shown.)

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 $v_{\text{drift}}dtd\Sigma$ pass the surface. The number we seek equals this volume times the ion density c_{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 make sure this formula has the proper units.) Substituting the drift velocity gives $\vec{j}_{\text{ion}} = q\vec{E}c_{\text{ion}}/\eta$, 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 (Fick's law) flux, obtaining

$$\vec{j}_{\text{ion},x}(x) = \frac{q\vec{E}_x(x)c_{\text{ion}}(x)}{\eta} - D_{\text{ion}} \frac{dc_{\text{ion}}}{dx},$$

where D_{ion} is the diffusion constant for the ion species in question. We next rewrite the viscous friction coefficient in terms of D_{ion} , using the Einstein relation $\eta D_{\text{ion}} = k_{\text{B}}T$ to get²

$$\vec{j}_{\text{ion},x} = D_{\text{ion}} \left(-\frac{dc_{\text{ion}}}{dx} + \frac{q}{k_{\text{B}}T} \vec{E}_x c_{\text{ion}} \right). \quad \text{Nernst-Planck formula} \quad (10.1)$$

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

In solution, electrophoretic motion is impeded by viscosity and can also be opposed by diffusion.

²Recall Section 8.5.2. More generally, in non-planar geometry the Nernst-Planck formula becomes $\vec{j}_{\text{ion}} = D_{\text{ion}}(-\vec{\nabla}c_{\text{ion}} + (q/k_{\text{B}}T)\vec{E}c_{\text{ion}})$.

tendency to erase 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

$$\frac{1}{c_{\text{ion}}} \frac{dc_{\text{ion}}}{dx} = \frac{q}{k_{\text{B}}T} \vec{E}_x. \quad (\text{thermal equilibrium}) \quad (10.2)$$

The left side of this formula can be written 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.³ To understand the right side, start with $\Delta\psi = -\vec{E}_x h$. Thus, the condition for thermal equilibrium is

$$\Delta(\ln c_{\text{ion}}) = -q(\Delta\psi_{\text{eq}})/k_{\text{B}}T. \quad \text{Nernst relation} \quad (10.3)$$

The subscript on $\Delta\psi_{\text{eq}}$ reminds us that this is the voltage 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 real numbers into Equation 10.3 yields a suggestive result. Consider a singly charged ion like Na^+ , for which $q = e$. Suppose that we have a moderately big concentration jump, $c_{\text{bot}}/c_{\text{top}} = 10$. Using the fact that $(k_{\text{B}}T_r/e) \approx \frac{1}{40}$ volt, we find $\Delta\psi \approx +58$ 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_{ion} controls how *fast* things move in 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 $e^{-q\psi(x)/k_{\text{B}}T}$. 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 $k_{\text{B}}T$. 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.

A living cell maintains an electric potential drop across its membrane.

³Normally 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.⁴ 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_{ion} , the electric field is $E = (k_{\text{B}}T/(D_{\text{ion}}qc_{\text{ion}}))j_{\text{ion}}$. Thus, our solution is ohmic (Equation 8.7, page 115) with conductivity

$$\kappa = \frac{D_{\text{ion}}q^2c_{\text{ion}}}{k_{\text{B}}T}. \quad (10.4)$$

Indeed, saltier water conducts better. To use Equation 10.4, remember that each type of ions 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 115):

$$\Delta\psi = IR \quad \text{where} \quad R = h/(\Sigma\kappa). \quad [8.8, \text{page 115}]$$

T2 Section 10.2.2' (page 148) mentions other points about electrical conduction.

Conductivity of a solution is quantitatively linked to diffusion of its mobile charges.

10.3 A REPULSIVE INTERLUDE

10.3.1 Electrostatic interactions are crucial for proper functioning of living cells

Section 6.7 (page 83) pointed out that when we put an acidic 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 maximize entropy. This chapter will show that for a large flat macroion, the compromise chosen by the counterions is to remain hanging in

⁴See Section 8.5.2 (page 116).

a cloud near the macroion's surface.⁵ After working Your Turn 6E (page 83), 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 in solution are short-range.

*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 diffuse layer 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 diffuse charge layer 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.⁶ 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, 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 repel one another.

Specific binding

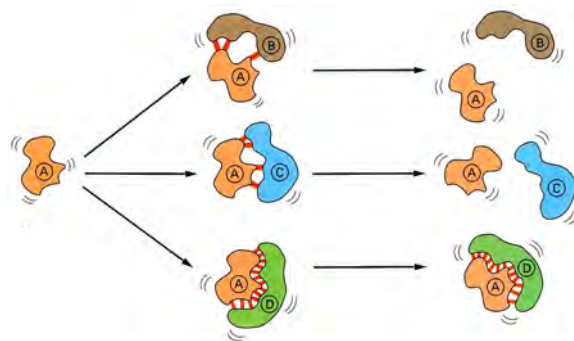
Idea 10.5 says that electrostatic forces are effectively of short range in solution, and moreover that this range is smaller than a typical macromolecule. That observation matters crucially for cells, because it means that two macromolecules will not feel one another until they're nearby. Even when they 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 around any particular location in the cell, typically only those with precisely matching shapes and charge distributions will bind together.

Biomacromolecules recognize one another and interact stereospecifically.

⁵Similar methods can be applied to a long, thin line of charge, such as a DNA molecule; see Problem 10.2.

⁶See Section 3.7.3.

Figure 10.2: [Cartoon.] **One source of binding specificity.** The bottom pair of molecules have complementary shapes and charge distributions, creating multiple attractive contributions to their mutual electrostatic energy. The molecules in the other two examples may have matching net charges, but due to the short-range character of electrostatic interactions in solution they are not as strongly bound.



One reason for this amazing specificity is that

Even though each individual electrostatic interaction between matching charges is rather weak (relative to $k_B T_r$), 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 75); a similar result holds in salt solution.⁷ 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.⁸

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 separates far enough to get past the resulting activation barrier, then it can greatly reduce its overall energy by separating completely into two fragments each with about half the original charge.⁹

A highly charged molecule reduces its electrostatic self-energy when it breaks up.

⁷See Problem 10.5.

⁸A quantum-mechanical effect (resonance) also reduces the bond energies of the fragments more than it does the original ATP.

⁹Recall Your Turn 6A (page 75).

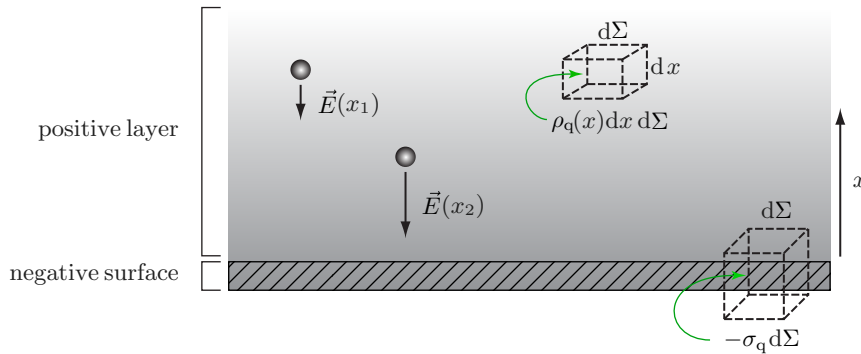


Figure 10.3: [Schematic.] **A planar distribution of charges.** A thin sheet of negative charge (*hatched, bottom*) lies next to a neutralizing positive layer of free counterions (*shaded, top*). The individual counterions are not shown; the shading represents their average density. The lower box encloses 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 surface 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 $\vec{E}(x)$ 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 test charge downward than between x_2 and infinity.

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 8.4 (page 119)–8.5.

10.3.2 The Gauss law

It is time to get quantitative. Figure 10.3 shows a thin, negatively charged sheet with uniform surface charge density $-\sigma_q$, next to a spread-out layer of positive charge with volume charge density $\rho_q(x)$. Thus, σ_q is a positive constant with units coul m^{-2} , whereas $\rho_q(x)$ is a positive function with units 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 surface charge density: Applying the Gauss law for a flat, charged surface gives

$$E|_{\text{surface}} = -\sigma_q/\epsilon. \quad (10.7)$$

Away from the surface, the Gauss law gives (see Figure 10.3)

$$\frac{dE}{dx} = \frac{\rho_q}{\epsilon}. \quad (10.8)$$

The following section will use this relation to find the electric field everywhere outside the surface.

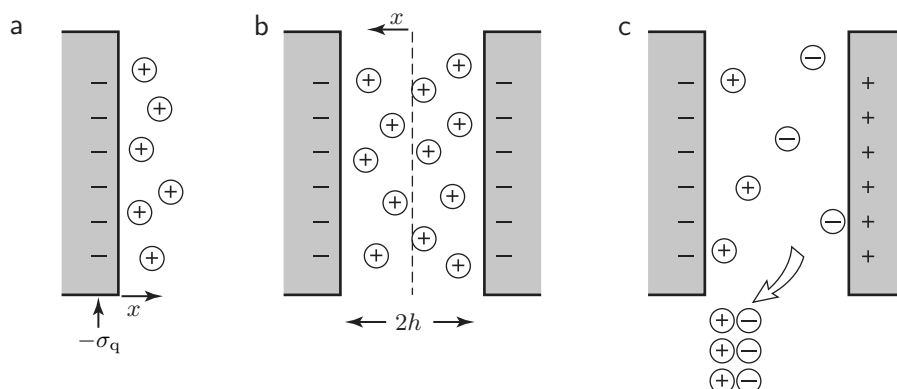


Figure 10.4: [Schematics.] **Behavior of counterions near surfaces.** (a) Counterion cloud outside a charged surface with surface charge density $-\sigma_q$. (b) When two similarly charged surfaces approach, their counterion clouds begin to get squeezed. (c) When two oppositely charged surfaces approach, their neutralizing counterion clouds are partly liberated and entropy increases.

10.3.3 Detailed form of the neutralizing ion cloud outside a charged surface in pure water

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 surface 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, Na^+). That is, each carries a single charge: $q_+ = e \approx 1.6 \cdot 10^{-19}$ 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 148).

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, ρ_q refers to the average density.

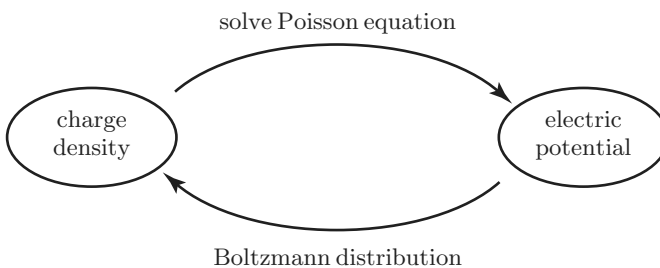


Figure 10.5: [Diagram.] **Strategy to find the mean-field solution.** Neither the Poisson equation nor the Boltzmann distribution alone can determine the charge distribution, but solving these two equations in two unknowns simultaneously does the job.

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, $c_+ \rightarrow 0$. 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 $c_+(0) = c_0$; so 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 ρ_q equal to the number density of counterions times e . The potential obeys the Poisson equation: $d^2\psi/dx^2 = -\rho_q/\epsilon$. 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 $ec_+(x) = ec_0 e^{-e\psi(x)/k_B T}$.

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 ρ_q to solve the Poisson equation for the potential ψ . But we need ψ to find ρ_q from the Boltzmann distribution! Luckily, each of the arrows in Figure 10.5 represents an equation in two unknowns, namely, ρ_q and ψ . 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 dimensionless rescaled potential $\bar{\psi}$:

$$\bar{\psi}(x) \equiv e\psi(x)/k_B T. \quad (10.9)$$

That change simplifies the exponential:

$$\frac{d^2\bar{\psi}}{dx^2} = -\frac{e^2 c_0}{k_B T \epsilon} e^{-\bar{\psi}}.$$

We can simplify still further by changing variables from x to a dimensionless rescaled variable:

Your Turn 10A

Let $\bar{x} = x/A$, where A is a constant.

- Confirm that the choice $A = \sqrt{\epsilon k_B T / (e^2 c_0)}$ has dimensions of length.
- Confirm that this choice simplifies our equation to the dimensionless form

$$\frac{d^2 \bar{\psi}}{d\bar{x}^2} = -e^{-\bar{\psi}}. \quad \text{Poisson–Boltzmann equation} \quad (10.10)$$

The payoff for introducing the abbreviations $\bar{\psi}$ and \bar{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 \bar{x} has the property that both its derivative and its exponential are powers of \bar{x} . We don't want $\bar{\psi}(\bar{x}) = \ln \bar{x}$, because that's divergent (equal to minus infinity) at the surface. Nevertheless, a slight modification gives something promising:

$$\bar{\psi}(\bar{x}) \stackrel{?}{=} \beta \ln(1 + (\gamma \bar{x})), \quad \text{trial solution of Equation 10.10} \quad (10.11)$$

where β and γ are two constants that we must find.

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:

- Our convention that $\bar{\psi}(0) = 0$. The trial solution Equation 10.11 always has that feature, regardless of what values we choose for β and γ .
- Our expectation that there will be no electric field at infinity because no charge is located there: $d\bar{\psi}/d\bar{x} \rightarrow 0$. Our trial solution also automatically satisfies this condition.

We now check whether we can choose values for the constants β and γ in such a way that the trial solution also solves the Poisson–Boltzmann equation. Substituting $\beta \ln(1 + (\gamma \bar{x}))$ 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 surface charge density, so we are not yet done. The surface form of the Gauss law (Equation 10.7) gives $-d\psi/dx|_{\text{surface}} = -\sigma_q/\epsilon$, or

$$\left. \frac{d\bar{\psi}}{d\bar{x}} \right|_{\text{surface}} = \frac{eA\sigma_q}{k_B T \epsilon}. \quad (10.12)$$

When using this formula, remember that σ_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 ψ 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 ψ 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 c_0 . We are given the surface 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

$$\frac{k_B T}{e} \left(\frac{\epsilon k_B T}{e^2 c_0} \right)^{-1/2} 2^{1/2} = \frac{\sigma_q}{\epsilon},$$

which we can solve for the unknown c_0 .

Your Turn 10B

- Show that $c_0 = \sigma_q^2 / (2\epsilon k_B T)$.
- Hence show that in the original variables the electrostatic potential is

$$\psi(x) = \frac{k_B T}{e} 2 \ln(1 + x/x_0), \quad (10.13)$$

where $x_0 = 2\epsilon k_B T / (e\sigma_q)$. Check the units.

Notice that *increasing the surface 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.¹⁰ Let's extract some physical conclusions from the math.

¹⁰ **[T2]** Or more realistically, a highly charged surface in a salt solution whose concentration is low enough; see Section 10.3.4' (page 148).

First, your answer to Your Turn 10C shows that indeed, a diffuse layer 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 think 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 Σ , stores electrostatic energy $\mathcal{E} = q_{\text{tot}}^2/(2C)$. Here q_{tot} is the total charge separated; for our case, it's $\sigma_q \Sigma$. The capacitance of a parallel-plate capacitor is given by $C = \epsilon \Sigma / x_0$ (Equation 6.11, page 78). Combining the preceding formulas gives an estimate for the density of stored electrostatic energy per unit area for an isolated surface in pure water:

$$\mathcal{E}/(\text{area}) \approx k_{\text{B}}T(\sigma_q/e). \quad (\text{electrostatic self-energy, no added salt}) \quad (10.14)$$

That makes sense: The environment is willing to invest about $k_{\text{B}}T$ per counterion in electric field energy. This energy gets stored in forming the diffuse layer.

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 spherical vesicle of radius $10 \mu\text{m}$ then carries stored free energy $\approx 4\pi(10 \mu\text{m})^2 \times (0.7/\text{nm}^2)k_{\text{B}}T_r \approx 10^9 k_{\text{B}}T_r$. 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 escaping. We may then expect that the diffuse charge layer will hug the surface more tightly than it does in Equation 10.13. That is,

$$\text{Increasing salt in the solution shrinks the diffuse layer.} \quad (10.15)$$

You'll make this expectation quantitative in Problem 10.4.

T2 Section 10.3.4' (page 148) 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,¹¹ 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 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 surface charge density $-\sigma_q$. We again choose the constant in ψ so that $\psi(0) = 0$; hence, the parameter $c_0 = c_+(0)$ is the unknown concentration of counterions at the midplane. $\psi(\bar{x})$ 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 $\bar{\psi}(\bar{x}) \stackrel{?}{=} \beta \ln \cos(\gamma\bar{x})$, where β and γ are again unknown constants. Certainly this trial solution is symmetrical and equals zero at the midplane, where $\bar{x} = 0$.

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 $\xi = (c_0 e^2 / (2\epsilon k_B T))^{1/2}$ gives

$$\tan(h\xi) = \frac{1}{\xi} \frac{\sigma_q e}{2\epsilon k_B T}. \quad (10.16)$$

Given the surface charge density $-\sigma_q$, we solve Equation 10.16 for ξ as a function of the spacing $2h$; then the desired solution is

$$\bar{\psi}(x) = 2 \ln \cos(\xi x), \quad c_+(x) = c_0 (\cos \xi x)^{-2}. \quad (10.17)$$

As expected, the charge density is greatest near the plates; the potential is maximum in the center.


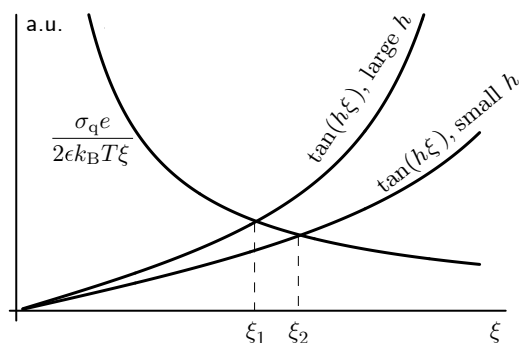
¹¹  This is not as restrictive as it sounds. Even in the presence of salt, our result will be accurate if the surfaces are highly charged because in this case, the Gouy–Chapman length is less than the Debye screening length (see Section 10.3.4', page 148).

Figure 10.6: [Mathematical functions.] **Graphical solution of Equation 10.16.** The sketch shows the dimensionless function $\sigma_q e / (2\epsilon k_B T \xi)$, as well as $\tan h\xi$ for two values of the plate separation $2h$. The value of ξ at the intersection of the rising and falling curves gives the desired solution. The figure shows that smaller plate separation gives a larger solution ξ_2 than does large separation (yielding ξ_1). Larger ξ in turn implies a larger ion concentration at the midplane and larger repulsive pressure.



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, partially rectifying its Brownian motion and creating a high-pressure zone in the gap. 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:

$$f/(\text{area}) = c_0 k_B T. \quad \text{repulsion of like-charged surfaces, no added salt} \quad (10.18)$$

In this formula, $c_0 = 2\xi^2 \epsilon k_B T / e^2$ and $\xi(h, \sigma_q)$ 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 ξ increases as the plate separation decreases (Figure 10.6). Thus, the repulsive pressure increases, too, as expected.

Your Turn 10D

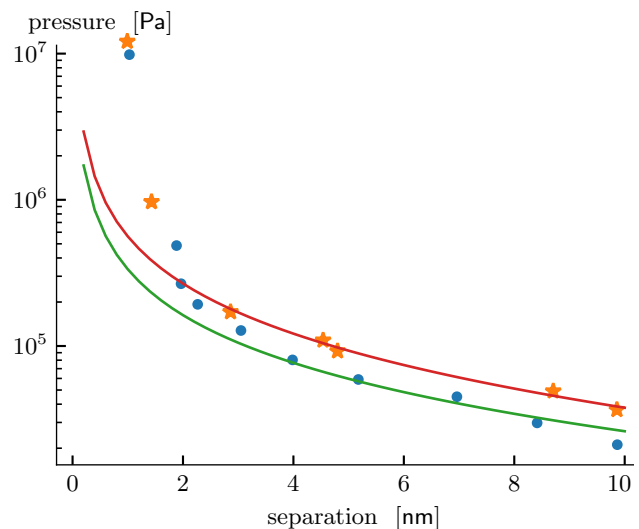
Make a similar graphical argument to find qualitatively what happens to ξ if we change the surface charge density, holding h fixed.

Note that the force just found is not simply proportional to the absolute temperature, because ξ 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 at the end of Section 10.3.4, the qualitative effect of adding salt to the solution is to tip this balance away from entropy, thereby shrinking the diffuse layers 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.

[T₂] Section 10.3.5' (page 155) derives the electrostatic force directly as a derivative of the free energy.

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 surface charge density σ_q . The *dashed line* shows the solution with one proton charge per 24 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. Beyond 2 nm, the purely electrostatic theory fits the data well, and 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.]



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.

When a complex of macromolecules is bound by many weak physical interactions, its stability can hinge on the surrounding salt concentration.

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 strands 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 139). 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 force, unlike the constant attractive force between two such planar surfaces in air. 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 143 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” has some mobile electrons not involved in covalent bonding; for example a crystal of silicon doped with a small impurity of antimony. When a slab of such material is placed next to a slab of “p-type” semiconductor (for example silicon doped with indium), some of the mobile electrons cross over to the other side, leaving the n-side positively charged and the p-side negatively charged. This charge separation costs energy, but still it happens at nonzero temperature; the details are mathematically similar to the diffuse layer studied here.

A semiconductor p–n junction creates a temperature-dependent potential drop.

FURTHER READING

Semipopular:

Electroosmotic pump: www.youtube.com/watch?v=zzVa_tX10iI.

Intermediate:

Smith et al., 2020; Safran, 2003

Biophysical applications, general: Schwarz, 2021, chap. 3; Benedek & Villars, 2000; Nelson, 2020; Grodzinsky, 2011; Dill & Bromberg, 2011.

Electrostatic model of protein stability: Bahar et al., 2017, §§3A and 9C.

T2 Voltaic cells: Saslow, 2021; Schmidt-Rohr, 2018.

Technical:

Bagotskii, 2006.

T₂

10.2.2' Electric currents in metals

Conductivity of many metals is far larger at low temperature than at room temperature, even without superconductivity.

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 the conductivity of copper should be, and is, greatly increased at cryogenic temperatures.

T₂

10.3.4'a 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 142), 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_∞ of salt in the surrounding water is never exactly zero.

Rather than introducing the unknown parameter c_0 and then going back to set it, this time we'll choose the constant in $\psi(x)$ so that $\psi \rightarrow 0$ far from the surface; then the Boltzmann distribution says

$$c_+(x) = c_\infty e^{-e\psi(x)/k_B T} \quad \text{and} \quad c_-(x) = c_\infty e^{-(-e)\psi(x)/k_B T} \quad (10.19)$$

for the monovalent counterions and coions, respectively. The corresponding Poisson–Boltzmann equation is

$$\frac{d^2 \bar{\psi}}{dx^2} = -\frac{1}{2} \lambda_D^{-2} [e^{-\bar{\psi}} - e^{\bar{\psi}}], \quad (10.20)$$

where again $\bar{\psi} = e\psi/k_B T$ and λ_D is defined as

$$\lambda_D \equiv (\epsilon k_B T / (2e^2 c_\infty))^{1/2}. \quad \text{Debye screening length} \quad (10.21)$$

In a solution of table salt, with $c = 0.1 \text{ M}$, 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

$$\bar{\psi}(x) = -2 \ln \frac{1 + e^{-(x+x_*)/\lambda_D}}{1 - e^{-(x+x_*)/\lambda_D}} \quad (10.22)$$

solves the equation. In this formula, x_* is any constant. [*Hint*: It saves some writing to define a new variable, $\zeta \equiv e^{-(x+x_*)/\lambda_D}$, and rephrase the Poisson–Boltzmann equation in terms of ζ , not x .]

Before we can use Equation 10.22, we must fix the value of x_* by imposing the surface boundary condition. Equation 10.12 (page 141) gives

$$e^{x_*/\lambda_D} = \frac{2\epsilon k_B T}{e\lambda_D \sigma_q} \left(1 + \sqrt{1 + (e\lambda_D \sigma_q / (2\epsilon k_B T))^2} \right). \quad (10.23)$$

10.3.4'b Low-salt limit

Let's examine the low-salt limit ($\lambda_D \rightarrow \infty$ for fixed σ_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 142).

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 142) displays its odd behavior. For $x \gg \lambda_D$, Equation 10.22 reduces to

$$\bar{\psi} \rightarrow -(4e^{-x_*/\lambda_D})e^{-x/\lambda_D}. \quad (10.24)$$

That is,

The electric fields far outside a charged surface in salt solution are exponentially screened at distances greater than the Debye screening length λ_D . (10.25)

Idea 10.25 and Equation 10.21 confirm an earlier expectation: Increasing c_∞ decreases the screening length, shrinking the diffuse charge layer 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 76) 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 (σ_q is small), Equation 10.23 gives $e^{-x_*/\lambda_D} \approx e\lambda_D \sigma_q / (4\epsilon k_B T)$, and so the potential simplifies to

$$\psi(x) \approx -\frac{\sigma_q \lambda_D}{\epsilon} e^{-x/\lambda_D}. \quad \text{potential outside a weakly charged surface} \quad (10.26)$$

There is a shortcut to this result. If a surface is weakly charged, then $\bar{\psi}$ will never deviate much from zero, and we may work with a simplified form of Equation 10.20:¹²

$$\frac{d^2 \bar{\psi}}{dx^2} \approx \lambda_D^{-2} \bar{\psi}. \quad (10.27)$$

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 solution of the general form Equation 10.26.

¹²See Problem 10.4.

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 contribute such **charge renormalization** effects.

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 λ_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

$$\mathcal{E}/(\text{area}) \approx \frac{\sigma_q^2 \lambda_D}{2\epsilon}. \quad \begin{array}{l} \text{(electrostatic energy with added salt,} \\ \text{weakly charged surface)} \end{array} \quad (10.28)$$

10.3.4'f How voltaic cells push electrons

We are now in a position to understand the action of a simple voltaic cell.¹³ 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:

Voltaic cells display several puzzling, but understandable, behaviors.

- What agency pushes electrons down a gradient of ψ , 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 rate of depletion, and the total degree of depletion? To what extent are those idealized statements 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.

¹³A “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 volt battery.” The former is actually a single cell, whereas the latter is six cells in series.

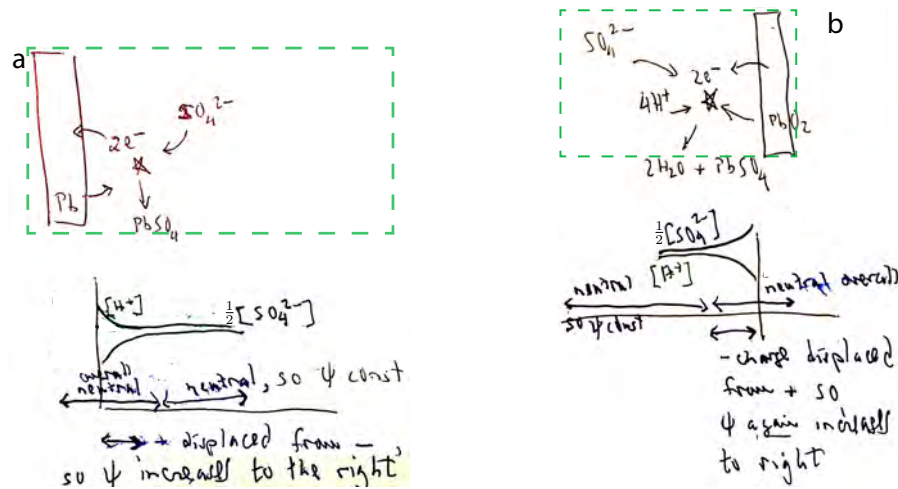


Figure 10.8: [Schematics; sketches of concentration profiles.] **Half-reactions of a lead-acid storage cell.** (a) “Left” reaction. In the lower panel, square brackets around an ion species denote its concentration. The acid solution is neutral in bulk, deviating from neutrality in a thin layer of width roughly the Debye screening length. (b) “Right” reaction.

- 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_2SO_4) between the plates. We will assume that this solution is dilute, and hence fully dissociated into H^+ and SO_4^{2-} ions.

[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 parts of this track-2 section, 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 realistic reactions at each electrode (“half reactions”):

- On the left, lead atoms can join with SO_4^{2-} to form neutral lead sulfate PbSO_4 , liberating its two excess electrons into the conducting lead electrode. PbSO_4 is insoluble, so it remains confined to the left electrode.
- On the right, molecules of lead oxide can join with SO_4^{2-} and four H^+ ions to form neutral species: two H_2O and one PbSO_4 . Two 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_2O 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 133), 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 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 PbSO_4 , as well as liberating some electrons (confined to the electrode) and capturing some SO_4^{2-} 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 10.3.4’d (page 149).¹⁴ 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 neutral solution beyond that).¹⁵ 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 ψ 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 less of its free energy change as heat than it did initially.

In each case, a wall (top of the chamber or electrode) 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.

¹⁴Substitute a solution to Equation 10.27 into Equation 10.19 and remember that the potential $\bar{\psi}$ will be small.

¹⁵The argument is similar to Section 10.3.4’a (page 148).

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 PbSO_4 , 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 diffuse layer 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.¹⁶

In contrast, suppose that we plunge two 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 connecting to a very large resistor, for example, a 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:¹⁷ 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 of the left, and H^+ to replace those lost on the right. (Some sulfate ions also move to the right

¹⁶Faraday correctly identified the drive as being chemical reactions localized at the electrodes.

¹⁷Equation 10.1 (page 133).

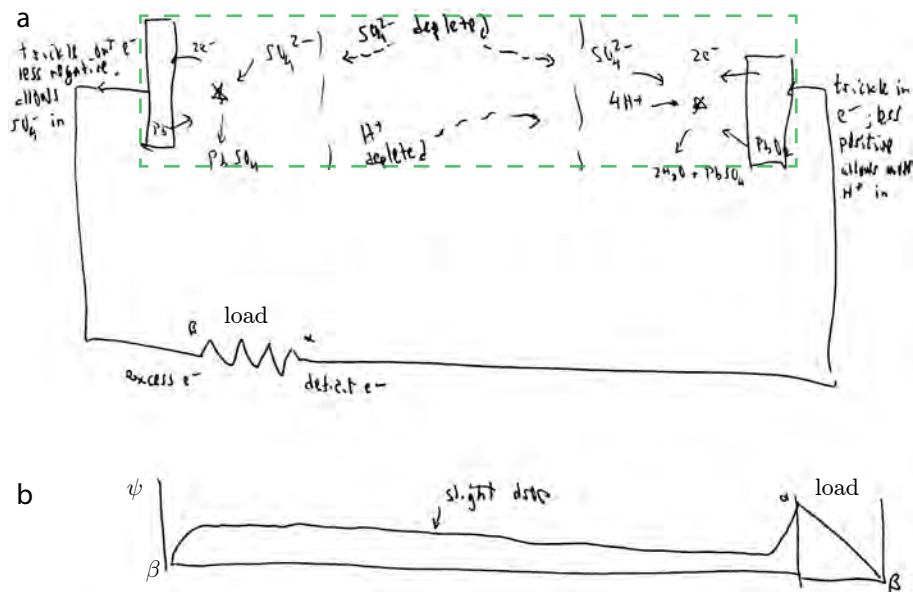


Figure 10.9: [Schematic; sketch graph.] **Closed circuit with load.** (a) Both half-reactions, plus continuous depletion of the bulk. (b) Electric potential profile.

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.

The net effect of all these processes is to set up a profile for ψ 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 PbO₂ being replaced by PbSO₄).

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 volt.

- 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.
- After a cell is depleted, we can connect it to a fresh battery (a “jumpstart 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 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.¹⁸

T₂

10.3.5’ Alternative derivation of force

The crucial last step leading to Equation 10.18 may seem too slick. Can’t we work out the force the same way we calculate any entropic force, by taking a derivative of the free energy? Absolutely. Let’s compute the Helmholtz free energy \mathcal{F} of the system of counterions+surfaces, holding fixed the charge density $-\sigma_q$ on each surface but varying the separation $2h$ between the surfaces (see Figure 10.4b, page 139). Then the force between the surfaces will be $p\Sigma = -d\mathcal{F}/d(2h)$, where Σ is the surface area.

As in the main text, suppose singly-charged counterions (charge $+e$). Define a convenient length scale, the **Bjerrum length**:

$$\ell_B \equiv \frac{e^2}{4\pi\epsilon k_B T}. \quad (10.29)$$

First we notice an important property of the Poisson–Boltzmann equation (Equation 10.10, page 141). Multiplying both sides by $d\bar{\psi}/dx$, we can rewrite the equation as

$$\frac{d}{dx} \left[\left(\frac{d\bar{\psi}}{dx} \right)^2 \right] = 8\pi\ell_B \frac{dc_+}{dx}.$$

Integrating this equation gives a first-order equation:

$$\left(\frac{d\bar{\psi}}{dx} \right)^2 = 8\pi\ell_B (c_+ - c_0). \quad (10.30)$$

¹⁸Our simplified discussion considered only the limit of very slow depletion, so that diffusion could keep pace with it.

To fix the constant of integration, we noted that the electric field is zero at the midplane, and $c_+(0) = c_0$ there.

The free energy density of an inhomogeneous ideal gas (or dilute solution) is

$$c_+(\vec{r}) (q\psi(\vec{r}) + k_B T \ln(c_+(\vec{r})/c_*)).$$

Here c_* is a constant whose value will drop out of our final answer because the integral $\int c_+ dx = 2\sigma_q/e$ is a constant, by charge neutrality. The free energy for our problem is the integral of this quantity, plus the electrostatic energy¹⁹ of the two negatively charged plates at $x = \pm h$:

$$\mathcal{F}/(k_B T \times \text{area}) = -\frac{1}{2} \frac{\sigma_q}{e} (\bar{\psi}(h) + \bar{\psi}(-h)) + \int_{-h}^h dx \left[c_+ \ln \frac{c_+}{c_*} + \frac{1}{2} c_+ \bar{\psi} \right].$$

We simplify our expression by first noting that $\ln(c_+/c_*) = \ln(c_0/c_*) - \bar{\psi}$, so the terms in square brackets are $c_+ \ln(c_0/c_*) - \frac{1}{2} c_+ \bar{\psi}$. The first of these terms is a constant times c_+ , so its integral is $2(\sigma_q/e) \ln(c_0/c_*)$. To simplify the second term, use the Poisson–Boltzmann equation to write $c_+ = -(4\pi\ell_B)^{-1} (d^2\bar{\psi}/dx^2)$. Next integrate by parts, obtaining

$$\mathcal{F}/(k_B T \times \text{area}) = 2 \frac{\sigma_q}{e} \left[\ln \frac{c_0}{c_*} - \frac{1}{2} \bar{\psi}(h) \right] + \frac{1}{8\pi\ell_B} \frac{d\bar{\psi}}{dx} \bar{\psi} \Big|_{-h}^h - \frac{1}{8\pi\ell_B} \int_{-h}^h dx \left(\frac{d\bar{\psi}}{dx} \right)^2.$$

We evaluate the boundary terms by using Equation 10.12 (page 141) at $x = -h$ and its analog on the other surface; they equal $-(\sigma_q/e) \bar{\psi}(h)$.

To do the remaining integral, recall Equation 10.30: it's $-\int_{-h}^h dx (c_+ - c_0)$, or $2(hc_0 - (\sigma_q/e))$. Combining these results gives

$$\mathcal{F}/(k_B T \times \text{area}) = 2hc_0 + 2 \frac{\sigma_q}{e} \left(\ln \frac{c_0}{c_*} - \bar{\psi}(h) - 1 \right) = \text{const} + 2hc_0 + 2 \frac{\sigma_q}{e} \ln \frac{c_+(h)}{c_*}.$$

The concentration at the wall can again be found from Equations 10.30 and 10.12: $c_+(h) = c_0 + (8\pi\ell_B)^{-1} (d\bar{\psi}/dx)^2 = c_0 + 2\pi\ell_B (\sigma_q/e)^2$.

A few abbreviations will make for shorter formulas. Let $\eta = 2\pi\ell_B \sigma_q/e$ and $u = \xi h$, where $\xi = \sqrt{2\pi\ell_B c_0}$ as in Section 10.3.5 (page 144). Then u and ξ depend on the gap spacing, whereas η does not. With these abbreviations,

$$\mathcal{F}/(k_B T \times \text{area}) = 2hc_0 + \frac{\eta}{\pi\ell_B} \ln \frac{c_0 + \eta^2/(2\pi\ell_B)}{c_*}.$$

We want to compute the derivative of this expression with respect to the gap spacing, holding σ_q (and hence η) fixed. We find

$$\frac{p}{k_B T} = -\frac{1}{k_B T} \frac{d(\mathcal{F}/(k_B T \times \text{area}))}{d(2h)} = -c_0 - \left(h + \frac{\eta}{2\pi\ell_B c_0 + \eta^2} \right) \frac{dc_0}{dh}.$$

In the last term, we need

$$\frac{dc_0}{dh} = \frac{d}{dh} \left(\frac{u^2}{h^2 2\pi\ell_B} \right) = \frac{u}{\pi\ell_B h^3} \left(h \frac{du}{dh} - u \right).$$

¹⁹Notice that adding any constant to $\bar{\psi}$ leaves this formula unchanged. To understand the reason for the factor $\frac{1}{2}$ in the first and last terms, think about two point charges q_1 and q_2 . Their potential energy at separation r is $q_1 q_2 / (4\pi\epsilon r)$ (plus a constant). This is *one half* of the sum $q_1 \psi_2(r_1) + q_2 \psi_1(r_2)$. (The same factor of $\frac{1}{2}$ also appeared in the electrostatic self-energy Example on page 75.)

To find du/dh , we write the boundary condition (Equation 10.16 (page 144)) as $\eta h = u \tan u$ and differentiate to find

$$\frac{du}{dh} = \frac{\eta}{\tan u + u \sec^2 u} = \frac{\eta u}{h\eta + u^2 + (h\eta)^2}.$$

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.

PROBLEMS

10.1 *Charged surfaces*

- Use some numerical software to solve Equation 10.16 for ξ as a function of plate separation $2h$ for fixed charge density σ_q . For concreteness, take σ_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.
- Obtain Dataset 1. Repeat (a) with other values of σ_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 charge per unit area $\sigma_q < 0$. 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 $e(R-a)\sigma_q/\epsilon$. The change in free energy is thus approximately $\Delta\mathcal{F} \approx e(R-a)\sigma_q/\epsilon - k_B T \ln(R/a)$, 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.

- 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 κ .
- Apply your result to the case of DNA, which has two ionized phosphate groups (charge $-2e$) for every basepair.

10.3 [Not ready yet.]

10.4 $\boxed{\mathcal{I}_2}$ *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 148) 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 $\bar{\psi}$ 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 $\bar{\psi}$. The resulting approximate equation is easy to solve. Solve it, and give an interpretation to the quantity λ_D defined in Equation 10.21.

10.5 **T₂** 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 is²⁰

$$\frac{1}{r} \frac{d^2(r\psi(r))}{dr^2} = \frac{1}{\lambda_D^2} \psi(r),$$

where λ_D is the Debye length.

a. Justify the following boundary conditions:

$$\psi(r) \rightarrow 0 \text{ as } r \rightarrow \infty, \quad - \left. \frac{d\psi}{dr} \right|_{r=a} = \vec{E}_r(\text{surface}) = \frac{q}{4\pi\epsilon a^2}.$$

b. Find $\psi(r)$ in terms of λ_D , a , and q .

c. The net charge density from salt ions is given by

$$\rho_q(r) = ec_\infty (e^{-\bar{\psi}} - e^{+\bar{\psi}}) \approx -\frac{\epsilon k_B T}{e\lambda_D^2} \bar{\psi}.$$

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.

The solubility of proteins generally increases with increasing ionic strength.

10.6 **T₂** Salt I

a. Calculate the Debye screening length for a 100 mM solution of sodium chloride. That is, the concentration of Na^+ ions is 0.1 mole per liter.

b. But magnesium chloride, for example, dissociates into Mg^{2+} ions (and twice as many 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 Mg^{2+} ions is 0.1 mole per liter.

10.7 **T₂** 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

²⁰See Your Turn 5A (page 68).

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 surface 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

$$\left. \frac{\partial \psi}{\partial z} \right|_{z=0} = A + B \sin(kx) \sin(ky).$$

Suppose that the surface is immersed in a salt solution with Debye screening length λ . 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 \mathcal{T}_2 *Direct calculation of a surface force*

Finish the derivation of Section 10.3.5' (page 155). The goal is to establish Equation 10.18.

10.9 \mathcal{T}_2 [Not ready yet.]

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 (gutta-percha and tarred hemp), 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 thin 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 congratulation 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. A parliamentary inquiry was mounted to see who should be blamed. Eventually a rumor spread that the entire project had been a massive hoax. Not until 1866 (after another snapped-cable fiasco), did a successful *cable*, following Thomson's original advice, come into operation.

Electromagnetic phenomenon: Small electrical disturbances on a nerve axon spread diffusively.

Physical idea: The cable equation is closely related to the diffusion equation.

11.2 COAXIAL CABLE

This chapter introduces a lot of notation. For reference, Table 11.1 lists some symbols introduced below.

Table 11.1: Symbols used in this chapter. See also Appendix B.

x	distance along cable axis
a	cable radius
κ	conductivity of interior
g or g_{tot}	conductance per area of insulating sheath; g_ℓ , membrane conductance per area for ion species ℓ
$\Delta\Sigma$	area of a segment of insulating sheath
C	capacitance of a segment of insulating sheath; \mathcal{C} , per area
R_x, R_r	axial and radial (“leak”) resistances, respectively, for a segment of length Δx and surface area $\Delta\Sigma$
$R_\ell = (g_\ell \Sigma)^{-1}$	membrane resistances for individual species
ψ_{out}	exterior electric potential, = 0 in our simplified model so $\Delta\psi = \psi_{\text{in}} - \psi_{\text{out}} = \psi_{\text{in}}$
$\psi_{\text{in}}(x, t)$	interior electric potential
I_x	axial (rightward) electric current inside cable
I_r	radial (outward) electric current through a segment (leak plus capacitive)
$\lambda_{\text{cable}}, \tau_{\text{cable}}$	space constant and time constant of cable (Equation 11.6, page 165)
w	modified potential (Your Turn 11B, page 166)
ϑ	speed of a traveling wave (Your Turn 11E, page 167)
$\psi_\ell^{\text{Nernst}}$	Nernst potential for ion species ℓ (Section 10.2.1, page 132)
$c_{\ell, \text{in}, \text{out}}$	concentration of ion species ℓ inside (respectively outside) a cell
g_ℓ^0, g_ℓ'	specifically the resting and excited conductances per area, respectively.
j_r	radial charge flux (current per area) actually passing through axon membrane; $j_{r, \ell}$, contribution from ion species ℓ (Equation 11.9, page 169)
ψ^0	combination of Nernst potentials giving the resting potential (Your Turn 11F, page 171)
$v(x, t)$	depolarization ($\Delta\psi$ shifted by ψ^0) (Equation 11.10, page 171)

11.2.1 A mathematical hyperlink to heat conduction

Thomson had understood both the loss and the spread of signals before the first 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

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 κ its conductivity. Let g be the leak

¹Actually, an undersea cable is surrounded by an infinite bath of salt water, so it’s not so unreasonable to neglect exterior resistance.

²See Section 8.6. This approximation breaks down at high frequency; see Chapter 18 for a more general discussion.

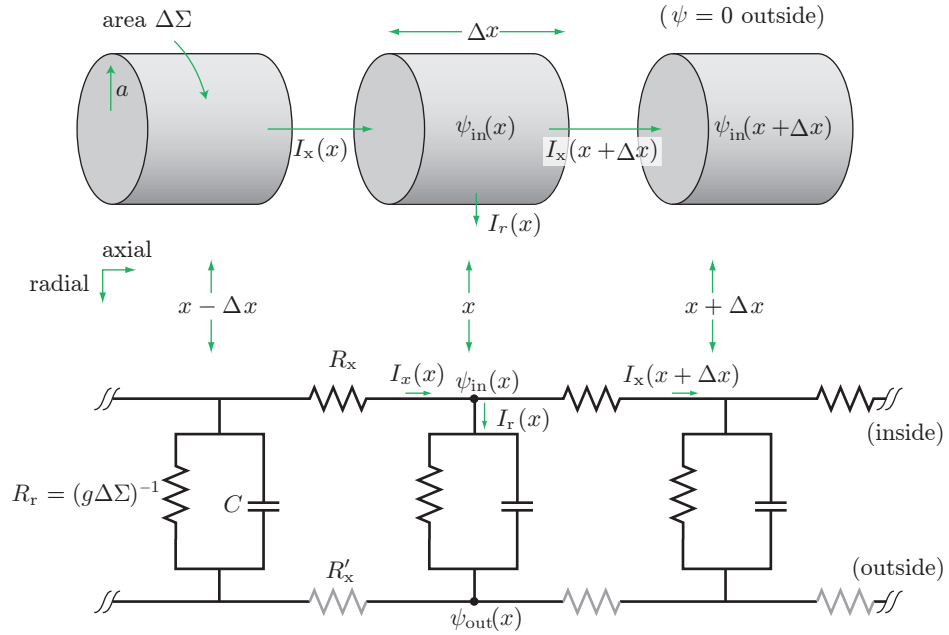


Figure 11.1: [Circuit diagram.] **Lumped-element model of a cable.** We will make the simplification of setting $R'_x = 0$, so that the exterior region is everywhere at potential zero.

conductance per unit area of the insulating sheath. It's positive and has units³ $\Omega^{-1}\text{m}^{-2}$. Also let \mathcal{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 Δx 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 $\Delta x \rightarrow 0$.

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 insulating 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. We are pretending that the corresponding axial resistance for the outer material is $R'_x = 0$, so 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 ψ_{in} .

³See Section 8.5.1 (page 115). Note that *conductance* per area has units different from those of the *conductivity*, κ , of a bulk material: The latter has units $\text{m}^{-1}\Omega^{-1}$.

⁴**[T2]** At ultra-high frequencies, a “skin effect” confines current to just the outermost part of a wire, invalidating this assumption.

Another resistance, $R_r = (g\Delta\Sigma)^{-1}$, 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 $C = \mathcal{C}\Delta\Sigma$ 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_r 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:

$$I_x(t, x) - I_x(t, x + \Delta x) = I_r(x) = \psi_{\text{in}}(t, x)/R_r + C \frac{\partial \psi_{\text{in}}}{\partial t}. \quad (11.1)$$

(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

$$\psi_{\text{in}}(x - \Delta x) - \psi_{\text{in}}(x) = I_x(x)R_x. \quad (11.2)$$

To summarize, we have expressed the discrete element properties in terms of material characteristics and geometry parameters:

$$C = \mathcal{C}\Delta\Sigma \quad \text{and} \quad R_r = 1/(g\Delta\Sigma), \quad \text{where} \quad \Delta\Sigma = 2\pi a\Delta x. \quad (11.3)$$

Also we have (Section 8.5.1, page 115) that

$$R_x = \Delta x/(\kappa\pi a^2). \quad (11.4)$$

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

$$\kappa\pi a^2 \frac{\partial^2 \psi_{\text{in}}}{\partial x^2} = 2\pi a \left(g\psi_{\text{in}} + \mathcal{C} \frac{\partial \psi_{\text{in}}}{\partial t} \right). \quad (11.5)$$

Define the **space constant** and **time constant** as

$$\lambda_{\text{cable}} \equiv \sqrt{a\kappa/(2g)}; \quad \tau_{\text{cable}} \equiv \mathcal{C}/g. \quad (11.6)$$

(Check that these expressions have the units of length and of time, respectively.) These abbreviations yield

$$(\lambda_{\text{cable}})^2 \frac{\partial^2 \psi_{\text{in}}}{\partial x^2} - \tau_{\text{cable}} \frac{\partial \psi_{\text{in}}}{\partial t} = \psi_{\text{in}}. \quad \text{linear cable equation} \quad (11.7)$$

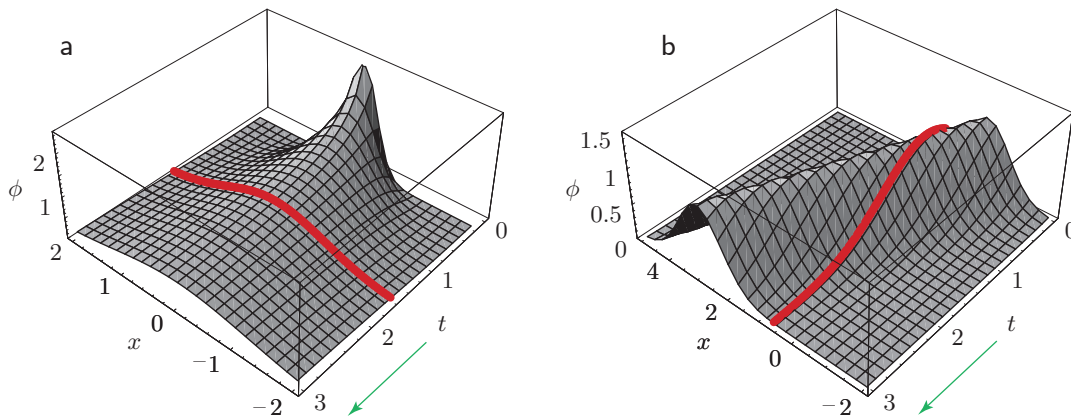


Figure 11.2: [Mathematical functions.] **Functions of two variables.** (a) A function $\phi(x, t)$, describing diffusion as a concentrated lump of solute begins to spread. Notice that time is drawn as increasing as we move diagonally downward in the page (*arrow*). The *heavy line* is the concentration profile at one particular time, $t = 1.6$. (b) This surface represents a function $\phi(t, x)$, describing a traveling wave. The *heavy line* shows the time course as seen by an observer fixed at $x = 0.7$.

Your Turn 11B

Change variables from ψ_{in} to $w(x, t) \equiv e^{t/\tau_{\text{cable}}}\psi_{\text{in}}(x, t)$ 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 $(\lambda_{\text{cable}})^2/\tau_{\text{cable}} = \kappa a/(2\mathcal{C})$, 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):

$$\psi_{\text{in}}(t, x) = \text{const} \times e^{-t/\tau_{\text{cable}}} t^{-1/2} e^{-x^2/(4Dt)}, \quad \text{passive-spread solution} \quad (11.8)$$

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. 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.

⁵This is the result that led Thomson to propose redesigning the cable with thicker insulation (smaller \mathcal{C}) and thicker central conductor (bigger a). But the Suits declared it was too late and too expensive to change the design.

Your Turn 11D

Imagine sitting at a fixed location x_* and observing the time course of the potential disturbance. At what time does the disturbance reach its peak? How does the peak strength vary as a function of x_* ? Maybe also use a computer to draw $\psi_{\text{in}}(t, x_*)$ for various x_* .

In fact, *the linear cable equation has no traveling wave solutions*:

Your Turn 11E

Substitute a trial solution of the form $\psi_{\text{in}}(t, x) = f(x - \vartheta t)$, into Equation 11.7, where ϑ is a constant, the speed of the proposed traveling wave (Figure 11.2b). Is there any value of ϑ that yields a physical solution?

Even if there is no leak conductance ($g \rightarrow 0$), our passive cable still suffers from dispersion. (Indeed, g dropped out altogether in the expression for the diffusion constant.)

11.3 NEURONS**11.3.1 Nerve impulses 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!⁶ 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 (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 over a few micrometers diffusive spreading is not a problem. 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 impulse, called the **action potential**, that moves *unchanged* in form, at constant speed. Your result in Your Turn 11E seemed to show that that is impossible, so we have work to do.

It is true that axons are filled with lots of other machinery, including microtubules. *Amazingly*, experiments have been done in which all those contents are emptied out of the axon and it is refilled with just a salt solution with concentrations of sodium

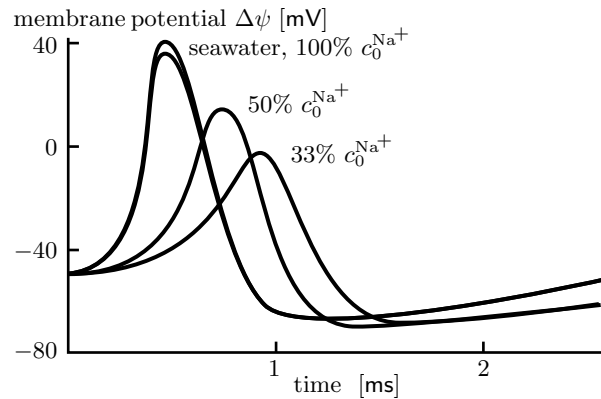
Small electrical disturbances on a nerve axon spread diffusively.

An axon carries signals that preserve their form and amplitude as they travel long distances.

The interior contents of an axon may be replaced by a simple salt solution without altering its transmission of signals.

⁶Hermann von Helmholtz measured this speed in 1850.

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 that 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 134), 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.]



The transmission of nerve impulses depends crucially on interior and exterior ionic conditions.

and potassium similar to the interior of a living cell (and hence different from the exterior, Figure 8.4, page 119). All the phenomena we will discuss (passive spread and the action potential) behave 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, counteracting 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. How these elements conspire to allow a nonlinear traveling wave solution is a remarkable story.

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 ℓ has its own concentration c_ℓ .

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.⁷

⁷A similar remark applies in plasma physics, and indeed there are some phenomena in common between that situation and aqueous solution. For example, both exhibit charge screening.

The membrane leakage conductances per area for each ion species, g_ℓ , 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.⁸ Far from being featureless tubes, each class of channels is sculpted in a way that selects for a particular ion (or type of ions).⁹

Thus, the net charge flow (current) through a channel due to ion species ℓ 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 ℓ .
- 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 concentrations requires a potential drop called the Nernst potential for species ℓ :

$$\psi_\ell^{\text{Nernst}} = -\frac{k_B T}{q_\ell} \ln(c_{\ell,\text{in}}/c_{\ell,\text{out}}). \quad [10.3, \text{page } 134]$$

But beware: The Nernst potential may not be equal to the *actual* potential drop. If they disagree, that just means that species ℓ 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 ℓ , 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_ℓ involves the permeability of a channel, the density of channels in the membrane, and the square of the charge carried by species ℓ ;¹⁰ it is therefore always a positive quantity.

Equation 11.9 makes precise a claim made 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.

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 squid¹¹):

ion	charge q_ℓ	interior		exterior		Nernst potential
		$c_{\ell,\text{in}}$, mM	relation	$c_{\ell,\text{out}}$, mM	$\psi_\ell^{\text{Nernst}}$, mV	
K ⁺	+e	400	>	20	-75	
Na⁺	+e	50	<	440	+54	
Cl ⁻	-e	52	<	560	-59	

⁸See Section 6.9 (page 85) and Section 8.7.2 (page 118).

⁹However, we will make the approximation that ions of each species all have the same mobility in bulk solution, leading to an overall conductivity κ that doesn’t care which species is moving.

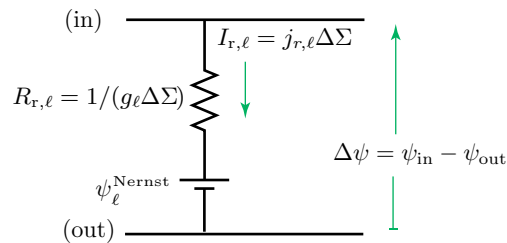
¹⁰The reasoning is similar to Section 8.5.2.

¹¹Nor from a superconducting quantum interference device!

The salient feature of this table is the last column: 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 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.

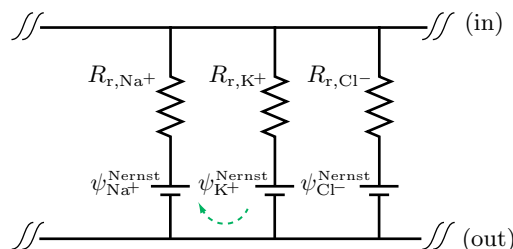
We can summarize the preceding discussion with a little circuit diagram representing the contribution of one species to the current through a patch of membrane:



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



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.

¹²The 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 .

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 ψ^0 and radial resistance $R_{r,\text{tot}}$. Explain the sense in which “the ion species with the biggest conductance gets the biggest vote when 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 $g_{\text{tot}} = \sum_{\ell} g_{\ell} \approx 5 \text{ m}^{-2} \Omega^{-1}$.

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 ψ^0 :

$$\kappa \pi a^2 \frac{\partial^2 \psi_{\text{in}}}{\partial x^2} = 2\pi a \left(g_{\text{tot}} (\psi_{\text{in}} - \psi^0) + \mathcal{C} \frac{\partial \psi_{\text{in}}}{\partial t} \right).$$

We can then eliminate the battery term altogether by changing variables to $v = \psi_{\text{in}} - \psi^0$:

$$(\lambda_{\text{cable}})^2 \frac{\partial^2 v}{\partial x^2} - \tau_{\text{cable}} \frac{\partial v}{\partial t} = v. \quad (11.10)$$

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 165).

Some illustrative numerical values are revealing:¹³ Taking $a = 0.5 \text{ mm}$, $g_{\text{tot}}^0 \approx 5 \text{ m}^{-2} \Omega^{-1}$, $\mathcal{C} \approx 1 \mu\text{F cm}^{-2}$, and $\kappa \approx 3 \Omega^{-1} \text{ m}^{-1}$ yields

$$\lambda_{\text{cable}} \approx 12 \text{ mm} , \quad \tau_{\text{cable}} \approx 2 \text{ ms}. \quad (11.11)$$

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 nerve impulses—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*.

¹³Chapter 9 discussed the early measurement of \mathcal{C} .

The resting membrane potential in squid axon was found to be $\psi^0 \approx -50$ 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 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. (11.12)

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

$$g_{K^+}^0 \approx 2g_{Cl^-}^0 \approx 3.2 \Omega^{-1} m^{-2} \text{ but } g_{Na^+}^0 \approx 0.08g_{Cl^-}^0. \quad (\text{resting}) \quad (11.13)$$

A modern estimate of the momentary values is

$$g_{K^+}' \text{ and } g_{Cl^-}' \text{ unchanged but } g_{Na^+}' \approx 160g_{Cl^-}^0. \quad (\text{at the action potential peak}) \quad (11.14)$$

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-thickness 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 itself 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 ψ . 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 can indeed 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/Transatlantic_telegraph_cable = perma.cc/QU4Y-YF6J

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

History of research on neurons: Raman & Ferster, 2021; Hodgkin, 1992.

Intermediate:

Nelson, 2020, chap. 12.

Technical:

Gutted axon experiment: Baker et al., 1962.

PROBLEMS

11.1 *Fate of a wave*

[Not ready yet.]

CHAPTER 12

Vista: Nerve Impulses

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 conductor, much like a coaxial cable, somehow the axon transmits signals over distances much longer than its diameter without amplitude loss nor waveform degradation—unlike the early undersea 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 *causes* 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 171), thus regenerating the action potential as 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 had the sought behavior.

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 regard it as an empirical fact.) For such a function, the entire

¹Equation 11.9 (page 169).

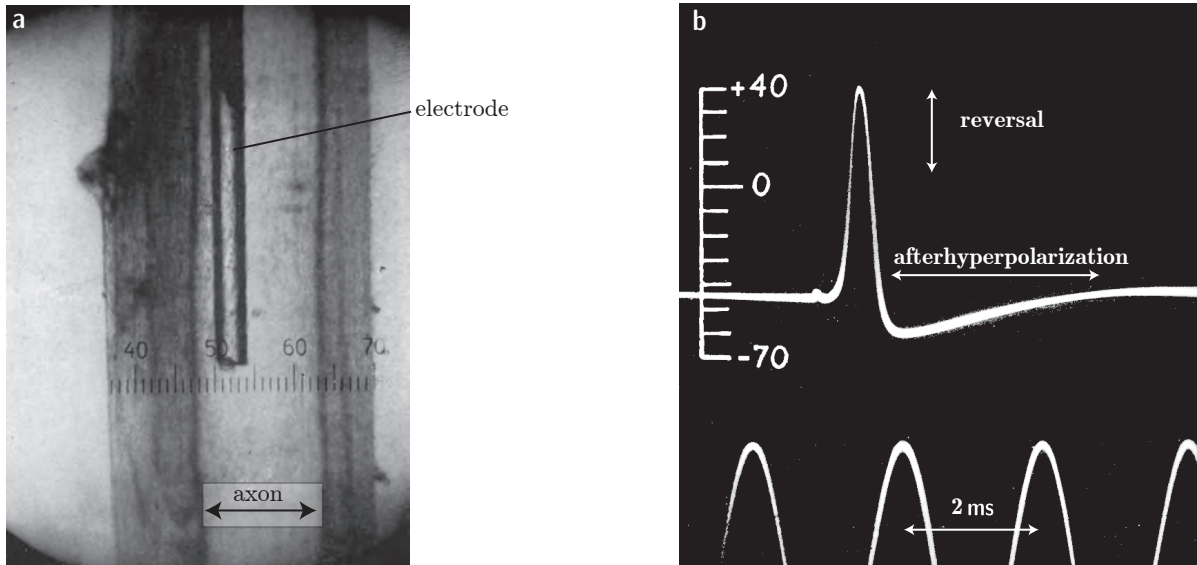


Figure 12.1: [Photomicrograph; oscilloscope trace.] **Hodgkin and Huxley's historic 1939 result.** (a) A recording electrode (a glass capillary tube) inside a giant axon, which shows as a clear space between divisions marked 47 and 63 on the scale. (The axon, in turn, is contained in a larger glass tube.) One division of the horizontal scale equals $33\ \mu\text{m}$. (b) Action potential and resting potential recorded between the inside and outside of the axon. Below the trace appears a time marker, showing reference pulses every 2 ms. The vertical scale indicates the potential of the internal electrode in millivolts, the seawater outside being taken as zero potential. Note that the membrane potential actually changes sign for a couple hundred microseconds; note also the overshoot, or afterhyperpolarization, before the potential settles back to its resting value.

[Reprinted by permission from Springer Nature: Nature, **144**, 710–711 (1939), Action Potentials Recorded from Inside a Nerve Fibre, Hodgkin, AL & Huxley, AF. ©1939.]

history $\psi_{\text{in}}(x, t)$ is completely known once we measure its time course at *one* point (Figure 12.1).² We then have

$$\psi_{\text{in}}(x, t) = \tilde{\psi}(t - (x/\vartheta)), \quad (12.1)$$

where the waveform $\tilde{\psi}(t) \equiv \psi_{\text{in}}(0, t)$ is shown in Figure 12.2a. Hence,

$$\frac{\partial \psi_{\text{in}}}{\partial x} = -\frac{1}{\vartheta} \frac{d\tilde{\psi}}{dt} \Big|_{t-(x/\vartheta)}, \quad (12.2)$$

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 165) to find the conduction charge flux, j_r , from the measured membrane potential $\tilde{\psi}(t)$:

$$j_r = \frac{I_r - C \partial \psi / \partial t}{2\pi a \Delta x} = -\frac{1}{2\pi a} \left(-\frac{\partial}{\partial x} \frac{\partial \psi_{\text{in}}}{\partial x} \Delta x \frac{\kappa \pi a^2}{\Delta x} \right) - c \frac{\partial \psi_{\text{in}}}{\partial t}.$$

²As in Chapter 11, we are considering a simplified model where the potential is everywhere zero outside the cable.

Table 12.1: Symbols used in this chapter. See also Appendix B.

$\psi_{\text{in}}(x, t)$	interior electric potential
ψ_{out}	exterior electric potential, = 0 in our simplified model so $\Delta\psi = \psi_{\text{in}} - \psi_{\text{out}} = \psi_{\text{in}}$
$\tilde{\psi}(t)$	waveform of a traveling wave (Equation 12.1)
ϑ	speed of a traveling wave (Your Turn 11E, page 167)
j_r	radial charge flux (current per area) actually passing through axon membrane; $j_{r,\ell}$, its component from ion species ℓ
a	axon radius
κ	conductivity of interior fluid
\mathcal{C}	capacitance per area of membrane
ψ^0	combination of Nernst potentials giving the resting potential (Your Turn 11F, page 171)
$v(x, t)$	depolarization ($\Delta\psi$ shifted by ψ^0); v_1 and v_2 , special fixed-point values (Figure 12.4)
$\tilde{v}(t)$	depolarization waveform of a traveling wave; \tilde{v} , dimensionless rescaled form
g_ℓ	membrane conductance per area for ion species ℓ ; g_{tot} , total
$\psi_\ell^{\text{Nernst}}$	Nernst potential for ion species ℓ (Section 10.2.1, page 132)
$\lambda_{\text{cable}}, \tau_{\text{cable}}$	space constant and time constant of axon (Equation 11.6, page 165)

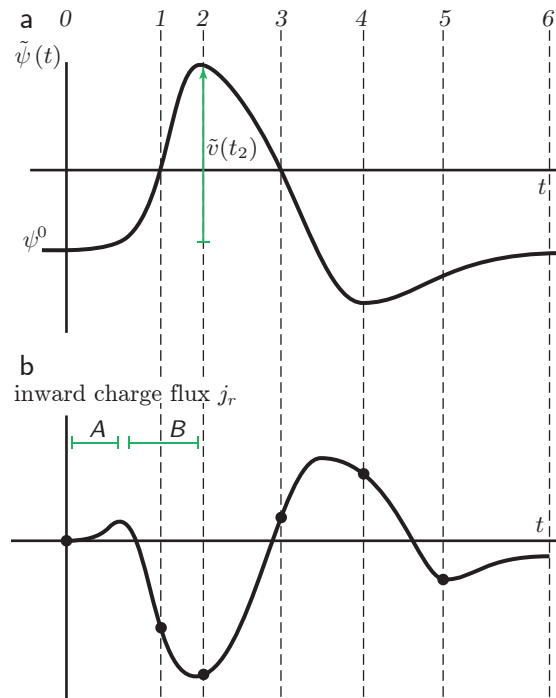


Figure 12.2: [Sketch graphs.] **Membrane current inferred from action potential.** (a) The sketch shows the membrane potential $\tilde{\psi}(t)$, measured at a fixed location $x = 0$. $\tilde{v}(t)$ refers to the difference between the membrane potential and its resting value ψ^0 . The *dashed lines* 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.]

For a traveling wave, Equation 12.2 lets us rephrase in terms of the measured time course at fixed position:

$$j_r = \frac{a\kappa}{2\vartheta^2} \frac{d^2\tilde{\psi}}{dt^2} - \mathcal{C} \frac{d\tilde{\psi}}{dt}. \quad (12.3)$$

The parameters a , κ , ϑ , and \mathcal{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, 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 $\tilde{\psi}(t)$. 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 $\tilde{\psi}(t)$, 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)).

Electrical measurements at the whole-membrane level already disclose non-ohmic behavior without requiring single-channel recording.

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 171). 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 ψ_{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 ψ_{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 172). 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*), so they proposed that

Membrane depolarization itself is the trigger that causes the sodium conductance to increase. (12.4)

That is, they suggested that some collection of unknown molecular devices in the membrane allow the passage of sodium ions, with a conductance depending on the membrane potential. Idea 12.4 introduces an element of **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 169) by allowing each of the membrane’s conductances to depend

³See Your Turn 11F (page 171).

on ψ_{in} :

$$j_r = \sum_{\text{species } \ell} (\psi_{\text{in}} - \psi_{\ell}^{\text{Nernst}}) g_{\ell}(\psi_{\text{in}}). \quad \text{prompt voltage-gating hypothesis} \quad (12.5)$$

In this formula, the transmembrane potential drop $\Delta\psi$ equals ψ_{in} because we still neglect any exterior resistance (Figure 10.1, page 133).

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 $\ln(c_{\text{out}}/c_{\text{in}})$ if we hold ψ_{in} fixed with an external source but change the concentrations. However, the membrane current is now a *nonlinear* function of ψ_{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**.⁴ Now we must translate our ideas into the context of axons, and do the math.

⁴Some authors restrict the phrase “excitable medium” to a more elaborate form that “resets” itself after a transient excursion. **T2** Real nerve axons have that property (Section 12.4'a, page 186), unlike the simplified models here and in Section 12.3.2.

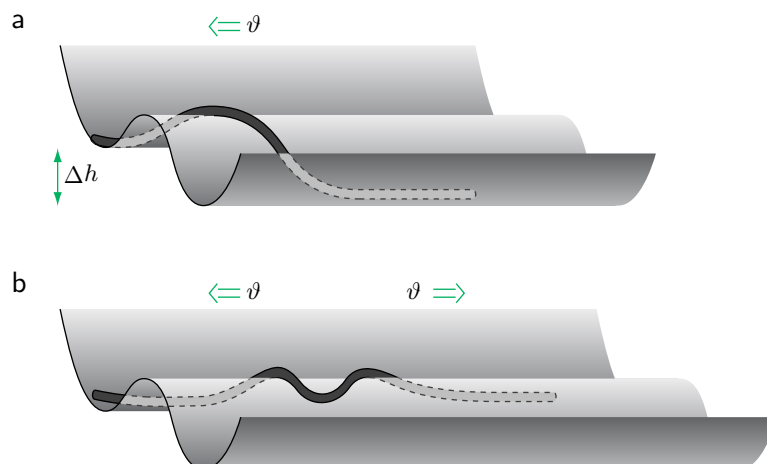


Figure 12.3: [Schematic.] **Mechanical analog of the action potential.** A flexible chain lies in a tilted channel, with two troughs at heights differing by Δh . In the axon context, the upper trough represents the steady or quasisteady state prior to an action potential. (a) An isolated kink will move steadily to the left at a constant speed v : 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 $\pm v$. Media 3 shows a physical realization of this system.

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 174) for the axon, which said that even though the resting axon is in a stable steady state of the membrane,

- Once one segment depolarizes, its depolarization spreads passively to the neighboring segment;
 - Once the neighboring segment depolarizes by more than a threshold value, the positive feedback phenomenon described in the previous section sets in, triggering more depolarization; and
 - The process repeats, spreading the depolarized region.
- (12.6)

We begin by thinking only about the initial sodium influx. Our working hypothesis is that the membrane's conductance per area for this ion, $g_{\text{Na}^+}(v)$, depends on the value⁵ of the depolarization $v \equiv \psi_{\text{in}} - \psi^0$.

A detailed model would use an experimentally measured form of the function $g_{\text{Na}^+}(v)$, 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

$$g_{\text{Na}^+}(v) = g_{\text{Na}^+}^0 + Bv^2. \quad (12.7)$$

Here $g_{\text{Na}^+}^0$ represents the resting conductance per area and B is a positive constant.

⁵Your Turn 11F (page 171) introduced the resting potential ψ^0 ; Equation 11.10 (page 171) 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 the References for more realistic models.

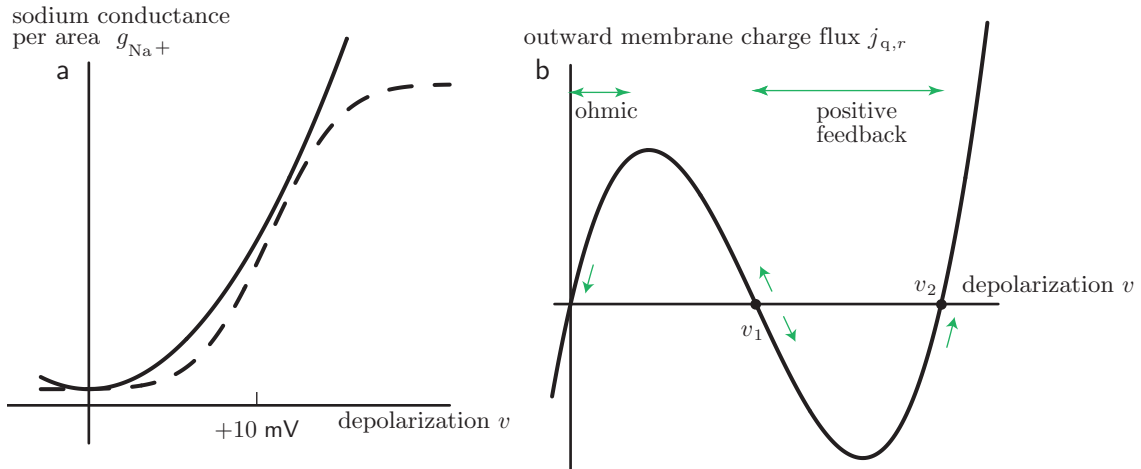


Figure 12.4: [Sketch graphs.] **Voltage-gating hypothesis.** (a) *Dashed curve:* The conductance g_{Na^+} of an axon membrane to sodium ions, showing an increase as the membrane potential increases from its resting value ($v = 0$). *Solid curve:* Simplified form for membrane sodium conductance (Equation 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.

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 the ohmic terms plus the extra sodium contribution:

$$j_r = \left(\sum_{\text{species } \ell} (\psi_{\text{in}} - \psi_{\ell}^{\text{Nernst}}) g_{\ell}^0 \right) + (\psi_{\text{in}} - \psi_{\text{Na}^+}^{\text{Nernst}}) B v^2. \quad (12.8)$$

As in Your Turn 11F (page 171), the first term in Equation 12.8 can be rewritten as $g_{\text{tot}}^0 v$. Letting H denote the constant $\psi_{\text{Na}^+}^{\text{Nernst}} - \psi^0$, we can also rewrite the last term as $(v - H) B v^2$, obtaining

$$j_r = v g_{\text{tot}}^0 + (v - H) B v^2. \quad (12.9)$$

Figure 12.4b shows 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 , and v_2 , where v_1 and v_2 equal $\frac{1}{2}(H \mp \sqrt{H^2 - 4g_{\text{tot}}^0/B})$, respectively. At small depolarization ($v \approx 0$), the sodium permeability 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 other⁶ stable fixed point v_2 . Our program is now to make the appropriate changes to the steps in Section 11.3.3 (page 170).

12.3.3 The nonlinear cable equation

We first substitute Equation 12.9 into the charge balance equation (Equation 11.5, page 165). Some algebra shows that $v_1 v_2 = g_{\text{tot}}^0 / B$, so the equation becomes

$$(\lambda_{\text{cable}})^2 \frac{\partial^2 v}{\partial x^2} - \tau_{\text{cable}} \frac{\partial v}{\partial t} = \frac{v(v - v_1)(v - v_2)}{(v_1 v_2)}. \quad \text{nonlinear cable equation}$$

(12.10)

Unlike the linear cable equation, Equation 12.10 is not equivalent to a diffusion equation.⁷ In general, it's very difficult to solve nonlinear, multivariable 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 ϑ by a function $\tilde{v}(t)$ of *one* variable, via $v(x, t) = \tilde{v}(t - (x/\vartheta))$. Substituting into Equation 12.10 leads to an *ordinary* (one-variable) differential equation:

$$\left(\frac{\lambda_{\text{cable}}}{\vartheta} \right)^2 \frac{d^2 \tilde{v}}{dt^2} - \tau_{\text{cable}} \frac{d\tilde{v}}{dt} = \frac{\tilde{v}(\tilde{v} - v_1)(\tilde{v} - v_2)}{v_1 v_2}. \quad (12.11)$$

We can tidy up the equation by defining the dimensionless quantities $\bar{v} \equiv \tilde{v}/v_2$, $y \equiv -\vartheta t/\lambda_{\text{cable}}$, $s \equiv v_2/v_1$, and $\gamma \equiv \tau_{\text{cable}}\vartheta/\lambda_{\text{cable}}$, finding

$$\frac{d^2 \bar{v}}{dy^2} = -\gamma \frac{d\bar{v}}{dy} + s\bar{v}^3 - (1 + s)\bar{v}^2 + \bar{v}. \quad (12.12)$$

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 γ 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 γ). You'll find in Problem 12.1 that choosing

$$\vartheta = \pm \frac{\lambda_{\text{cable}}}{\tau_{\text{cable}}} \sqrt{\frac{2}{s}} \left(\frac{s}{2} - 1 \right) \quad (12.13)$$

does yield a traveling wave solution (the solid curves in Figure 12.5).

⁶The value v_1 is an unstable fixed point, because small deviations above or below it get driven to larger deviations (Figure 12.4b).

⁷Contrast Section 11.3.3 (page 170).

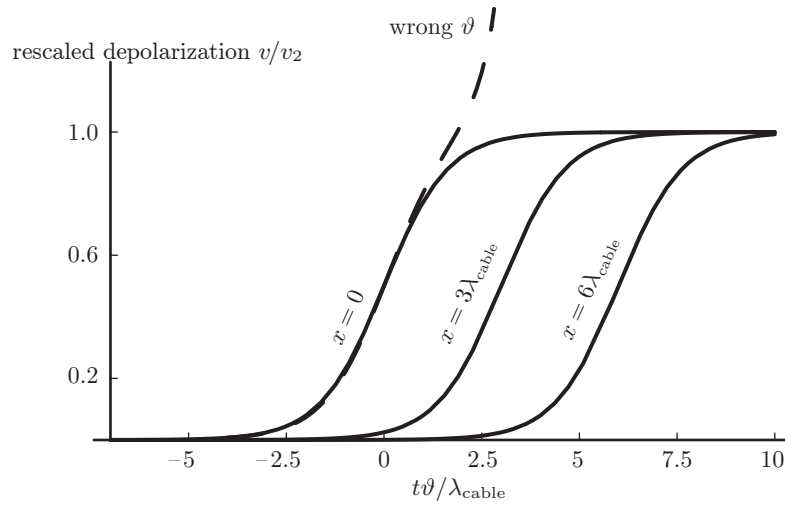


Figure 12.5: [Mathematical functions.] **Traveling wave solution to the nonlinear cable equation** (see Problem 12.1). The membrane potential relative to rest, $v(x, t)$, is shown as a function of time at three different fixed locations (*three solid curves*). Points at larger x see the wave go by at later times, so this wave is traveling in the $+\hat{x}$ 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 ϑ different from that in Equation 12.13; this solution is singular. Time is expressed as multiples of $\lambda_{\text{cable}}/\vartheta$. The depolarization v is expressed as multiples of v_2 .

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 linear one (Equation 11.7, page 165), 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.* (12.14)
- *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 nerve impulses. 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 ϑ is proportional to $\lambda_{\text{cable}}/\tau_{\text{cable}} = \sqrt{\alpha\kappa g_{\text{tot}}^0/(2C^2)}$ times a factor independent of the axon's radius a . 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 $\vartheta \propto \sqrt{a}$. This prediction is roughly borne out in experimental data.⁸ 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 m s^{-1} .

An axon carries electrochemical signals, but does so far more slowly than a coaxial cable.

In the mechanical analogy, the wave speed is proportional to the density of stored energy divided by a friction constant. Both κ and g_{tot} are inverse resistances, so $\sqrt{\kappa g_{\text{tot}}}$ in our expression for ϑ is indeed an “inverse friction”-type constant. In addition, the formula $\mathcal{E}/\Sigma = \frac{1}{2}q^2/(\mathcal{C}\Sigma^2)$ for the stored electrostatic energy in a capacitor shows that it is proportional to $1/\mathcal{C}$. Thus, the prefactor in Equation 12.13 has the overall form expected from the mechanical analogy.

T2 Section 12.3' (page 185) 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, nerve impulses, so critical for all multicellular animals, have turned out to be a physics problem. That is, a handful of classes of actors, obeying rules that can be characterized with simple functions, could be assembled as elements of a mathematical model that made many 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.⁹

T2 Section 12.4' (page 186) mentions more details about realistic axon models.

FURTHER READING

Semipopular:

“Dancing Zombie Squid Explained” www.youtube.com/watch?v=JGPfSSU1ReM

Intermediate:

Neurons: Raman & Ferster, 2021; see also Phillips et al., 2012.

The simplified treatment discussed here, and more realistic models, appear in Baylor, 2020; Nelson, 2020; Bressloff, 2014, §2.2, Keener & Sneyd, 2009, §6.2; Benedek & Villars, 2000. Nelson, 2020, §12.3.1 outlines how Hodgkin and Huxley managed to measure the conductance functions experimentally.

ocw.mit.edu/courses/brain-and-cognitive-sciences/9-40-introduction-to-neural-computation-spring-2018/lecture-videos/4.-hodgkin-huxley-model-part-1/

ocw.mit.edu/courses/brain-and-cognitive-sciences/9-40-introduction-to-neural-computation-spring-2018/lecture-videos/5.-hodgkin-huxley-model-part-2

⁸ **T2** Strictly speaking, our result applies only to “unmyelinated” axons.

⁹ See Media 4 for one example.

Technical:

Artificial axon recapitulating action potentials: Ariyaratne & Zocchi, 2016.

Even plants have action potentials: Hedrich & Neher, 2018.

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 generic, here is a physically inspired argument. Begin with Equation 12.10 (page 181). 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 $v(t, x) = \tilde{v}(t - x/\vartheta)$ where $\tilde{v}(t) \rightarrow 0$ as $t \rightarrow -\infty$ and $\tilde{v}(t) \rightarrow v_2$ as $t \rightarrow +\infty$. The wave velocity ϑ 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:

$$y = -|\vartheta|t/\lambda_{\text{cable}}, \quad \frac{d}{dt} = -\frac{|\vartheta|}{\lambda} \frac{d}{dy}.$$

As a function of y , our desired behavior is that $\tilde{v}(y) \rightarrow 0$ as $t \rightarrow +\infty$ and so on. Now multiply both sides of Equation 12.11 by $d\tilde{v}/dy$ and rearrange to find

$$\frac{d}{dy} \left[\frac{1}{2} \left(\frac{d\tilde{v}}{dy} \right)^2 + U(\tilde{v}) \right] = -\gamma \left(\frac{d\tilde{v}}{dy} \right)^2, \quad (12.15)$$

where

$$U(\tilde{v}) = -\frac{1}{v_1 v_2} \left[\frac{1}{4} \tilde{v}^4 - \frac{1}{3} (v_1 + v_2) \tilde{v}^3 \right] - \frac{1}{2} \tilde{v}^2, \quad \text{and} \quad \gamma = \tau_{\text{cable}} |\vartheta| / \lambda_{\text{cable}}. \quad (12.16)$$

Similar manipulations continue to work for any voltage gating function with the general form in Figure 12.4 (page 180), 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 an appeal to mechanics. Think of a roller-coaster car, rolling with "position" \tilde{v} 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 a world where roller coasters are immersed in a viscous fluid). Our roller coaster starts at "time" $y \rightarrow -\infty$ on top of a hill ($\tilde{v} = v_2$). 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 179), 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 $(B/g_{\text{tot}}^0)(40 \text{ mV})^2 = 52$, and then $v_1 = 0.3 \text{ mV}$ and¹⁰ $v_2 = 100 \text{ mV}$.

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 $\tilde{v} \rightarrow -\infty$ 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 γ , because it contains the unknown propagation speed ϑ . 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 $\tilde{v} = v_1$. If the friction is too small, it will overshoot $\tilde{v} = 0$ 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 ($\tilde{v} = 0$)!

¹⁰The 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 later cut off the rise of potential.

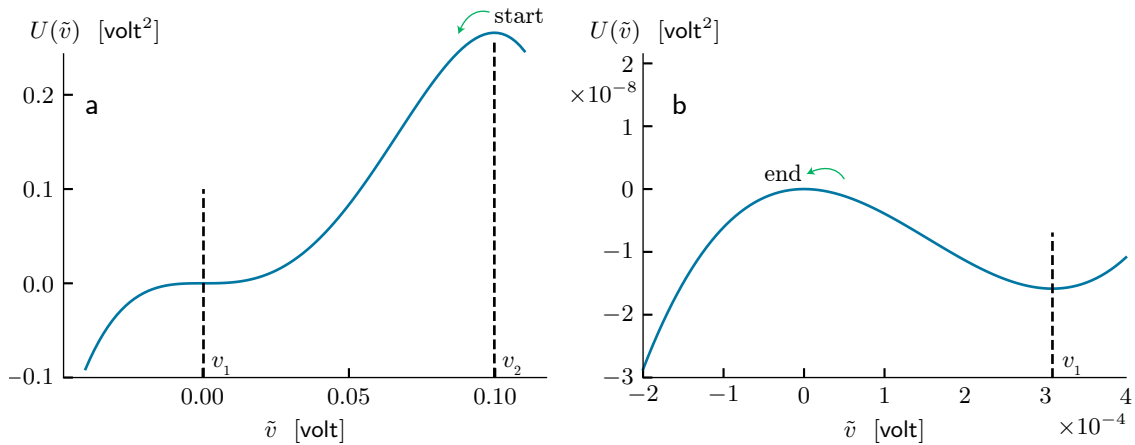


Figure 12.6: [Mathematical functions.] **Effective potential for Equation 12.16.** For illustration, total resting membrane conductance was set to $5\Omega^{-1}\text{m}^{-2}$ and $H = \psi_{\text{Na}^+}^{\text{Nernst}} - \psi^0 = 100\text{mV}$. (a) The desired solution starts at $\tilde{v} = v_2$. (b) Magnified form of the left side of (a). The desired solution coasts to a stop at the left hilltop ($\tilde{v} = 0$).

T₂

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 8.5a (page 120), 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 175).

Most neural action potentials are biphasic (hyperpolarization follows a depolarizing spike).

The word “later” reminds us that channels do not actually respond instantly to the current membrane potential; 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 instantly, 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**.¹¹ Here is will explore another context for them. The cables that send Internet between cities are not wires at

¹¹Some authors reserve this word for the special case of an “exactly integrable” system.

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.

PROBLEMS

12.1 *Analytic solution for simplified action potential*

Show that the function $\bar{v}(y) = (1 + e^{\alpha y})^{-1}$ solves Equation 12.12 (page 181), if we take the parameter γ to be given by $\sqrt{2/s}(\frac{s}{2} - 1)$. Hence derive the speed of the action potential (Equation 12.13, page 181). α is another constant, which you are to find.

12.2 [Not ready yet.]

CHAPTER 13

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.¹ More precisely, this chapter discusses tensors in three-dimensional space, abbreviated **3-tensors**.²

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”), modeled on the tangent to a curve at a point. It can be specified by giving three numbers (its **components**), after first choosing a cartesian coordinate system on space. The three components $\{\vec{r}_i\}$ of a 3-vector \vec{r} carry one coordinate index, hence the name “rank 1.”

Equally, we 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,

¹Tensor calculus was developed gradually, starting with G. Ricci-Curbastro around 1890.

²[\[T2\]](#) Also, Section 7.2.3’ (page 108) constructed the fundamental forms of a two-dimensional surface; they are **2-tensor fields** on a curved space.

where $\{\vec{a}_i\}$ is a set of three constants. Either way, we need three numbers to specify an object in this class.

From now on, 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 two closely related roles in pre-Einstein physics:³

- A tensor may express a linear, vector-valued function of another vector.
- A tensor may express a scalar-valued function of a vector that is quadratic, or a scalar 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 two 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 specified by components (an array of ordinary numbers).

13.3.1 A tensor can represent 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 to ensure that the wheel's center of mass (CM) lies 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 is constantly subjected to sideways forces, which would wear out the bearings and so on 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 generates *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 ω , we define the **angular velocity** $\vec{\omega}$ as the product of ω 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_ℓ momentarily located at positions $\vec{r}_{(\ell)}$ relative to a reference point fixed in the body. Then the resulting **angular momentum** \vec{L} has cartesian components that are linear functions of those of $\vec{\omega}$, and therefore may be written⁴ as $\vec{L}_i = \vec{J}_{ij}\vec{\omega}_j$, or more simply $\vec{L} = \vec{J} \cdot \vec{\omega}$. Here the **moment of inertia tensor** \vec{J} is a set of quantities with two indices, and hence is said to have rank 2.

We can compactly express \vec{J} by the formula

$$\vec{J} = \sum_{\ell} m_{\ell} [\|\vec{r}_{(\ell)}\|^2 \mathbf{1} - \vec{r}_{(\ell)} \otimes \vec{r}_{(\ell)}]. \quad (13.1)$$

³Physicists also began to use tensors in four or more dimensions after H. Minkowski reformulated Einstein's ideas in this language (Chapter 32).

⁴As 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.

The symbol $\overset{\leftrightarrow}{\mathbb{1}}$ represents the **identity tensor**, whose entries (“components”) are the identity matrix. Equivalently, $\overset{\leftrightarrow}{\mathbb{1}}$ 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 \vec{a} and returns a rescaled version of $\vec{r}_{(\ell)}$:

$$(v \otimes v) \cdot \vec{a} = v(v \cdot \vec{a}).$$

The new vector again depends linearly on \vec{a} . Just as we can represent a vector by its cartesian components $\vec{r}_1 = \hat{x} \cdot \vec{r}$, and so on, so also the dyad product has the nine components

$$[\vec{r} \otimes \vec{r}]_{ij} = \hat{e}_{(i)} \cdot ((\vec{r} \otimes \vec{r}) \cdot \hat{e}_{(j)}) = (\vec{r}(\vec{r} \cdot \hat{e}_{(j)}))_i = r_i r_j = \begin{bmatrix} x^2 & xy & xz \\ yx & y^2 & yz \\ zx & zy & z^2 \end{bmatrix}_{ij}, \quad i, j = 1, 2, 3. \quad (13.2)$$

Thus, each of the two terms of Equation 13.1 has a 3×3 array of components, and hence so does their sum $\overset{\leftrightarrow}{\mathbb{J}}$.

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. (13.3)

Yet another view is to regard the components as forming a column vector $[\vec{r}]$; then the usual rules of matrix multiplication give that $[\vec{r}][\vec{r}]^t$ is a 3×3 matrix containing the components of the dyad product.

A tensor whose matrix of components is symmetric, for example $\overset{\leftrightarrow}{\mathbb{J}}$, will itself be called a **symmetric** tensor. If we set the two indices equal (consider only diagonal elements) and sum them, then the result is a single number called the **trace** of the tensor.⁶

Your Turn 13A

- Use first-year physics formulas to find an expression for \vec{L} in terms of $\vec{\omega}$, $\{\vec{r}_{(\ell)}\}$, and $\{m_\ell\}$. Then rearrange as needed to obtain Equation 13.1.
- Although $\overset{\leftrightarrow}{\mathbb{J}}$ is symmetric, show that it need not be traceless (in contrast to the electric quadrupole moment tensor).
- Show that if $\vec{\omega}$ is an eigenvector of $\overset{\leftrightarrow}{\mathbb{J}}$, then the body can spin freely about that axis without wobbling (precessing).

Note that although \vec{L} depends linearly on the components of $\vec{\omega}$, it need not point parallel to $\vec{\omega}$. If not, then in rigid rotation \vec{L} will trace out a cone with $\vec{\omega}$ as its axis. That time dependence implies the unwanted torque mentioned in the automotive example.

⁵Some 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.

⁶We encountered the trace in Sections 3.2 and 7.2.3.

Also note that, although \vec{L} and $\vec{\omega}$ both change sign if we switch to a left-handed coordinate system, nevertheless, the relation between them is unaffected.

Your Turn 13B

- Indeed, Equation 13.1 does not contain any cross products. Where did they go?
- 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 any two vectors is defined similarly: $\vec{a} \otimes \vec{b}$ is the tensor that eats any \vec{v} and returns

$$(\vec{a} \otimes \vec{b}) \cdot \vec{v} = \vec{a}(\vec{b} \cdot \vec{v}).$$

Each of the three components of the right side of this expression is a linear function of the component of \vec{v} . Notice that $\vec{a} \otimes \vec{b}$ is not necessarily the same function as $\vec{b} \otimes \vec{a}$: *The dyad product is not commutative.*

As before, we can construct a 3×3 array of ordinary numbers by letting $\vec{a} \otimes \vec{b}$ eat each of the coordinate axes and expanding the three resulting vectors in components.

Your Turn 13C

- Show that the resulting matrix, which can be written as $[\vec{a} \otimes \vec{b}]_{ij}$, has entries given by the products $a_i b_j$ in row i and column j . Compare this definition to the special case Equation 13.2. This matrix will only be symmetric if \vec{a} is a scalar multiple of \vec{b} (or if either \vec{a} or \vec{b} is zero).
- Show that the components of $\vec{b} \otimes \vec{a}$ form a matrix that is the transpose of $[\vec{a} \otimes \vec{b}]_{ij}$.

Still more generally, 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 expressible as a dyad product.

13.3.3 Tensors arise naturally throughout physics: some examples

- 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 $\vec{f} = \vec{\zeta} \cdot \vec{v} + \dots$, involving a **viscous drag tensor**.⁷ The fact that \vec{f} need not be parallel to \vec{v} is the key to bacterial locomotion (Figure 13.1).

⁷For 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

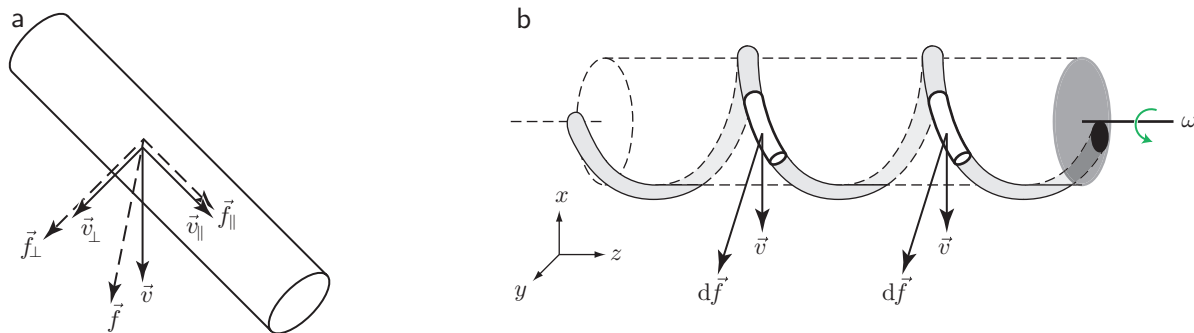


Figure 13.1: [Schematic.] **Principle of flagellar propulsion in bacteria.** (a) A thin rod is dragged through viscous fluid. The force required to get velocity \vec{v} is not parallel to \vec{v} , because the drag coefficient is larger in the perpendicular direction. (b) A thin, rigid, helical rod is cranked about its helix axis at angular speed ω . 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. [From Nelson, 2020.]

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 equilibrium displaced by some \vec{r} , such that $\vec{f} = -\vec{K} \cdot \vec{r}$. Here the **spring constant 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 **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 **induced dipole moment**. In the linear regime,⁸

$$\vec{D}_E = \vec{\alpha} \cdot \vec{E},$$

where $\vec{\alpha} = q^2[\vec{K}]^{-1}$ is called the **polarizability tensor**. Because molecular polarizability gives rise to dielectric susceptibility and thence 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 $\vec{j} = \kappa\vec{E}$, we have $\vec{j} = \vec{\kappa} \cdot \vec{E}$, where $\vec{\kappa}$ is called the **conductivity tensor**. Similarly, any molecule

Molecular polarizability is in general anisotropic.

The conductivity of an ohmic medium is in general anisotropic.

velocity; that term can dominate for large bodies moving rapidly through a low-viscosity medium (for example, for wind resistance on a car).

⁸Even if polarization is nonlinear in applied field, we can use the first-order part of its Taylor expansion to define a tensor.

⁹See Sections 6.5.2–6.5.3. The tensor character of permittivity will enter our discussions of birefringence in Chapter 50.

or ion that moves diffusively has a **mobility tensor**, which need not be a scalar if the medium is anisotropic.

5. The net force $d\vec{f}_{1\rightarrow 2}$ 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,

$$d\vec{f}_{1\rightarrow 2} = \overset{\leftrightarrow}{T} \cdot d\vec{\Sigma}_{1\rightarrow 2}, \quad (13.4)$$

where $\overset{\leftrightarrow}{T}$ is a symmetric rank-2 tensor and $d\vec{\Sigma}_{1\rightarrow 2}$ is directed along the perpendicular from 1 to 2.

6. **[T2]** 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 $\overset{\leftrightarrow}{T}$ 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, *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, an instance of Newton's Third Law. But Equation 13.4 involves a 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 $\overset{\leftrightarrow}{T}$ the **momentum flux 3-tensor**.¹¹

Your Turn 13D

- Still thinking about still air in a room, suppose that $d\vec{\Sigma}_{1\rightarrow 2}$ is oriented along the $+z$ axis. Make a connection between $\overset{\leftrightarrow}{T}_{zz}$ and air pressure.
- 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 $\overset{\leftrightarrow}{T}_{zz}$ is the same as in (a).
- 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 $\overset{\leftrightarrow}{T}_{zz}$.

¹⁰See Problem 14.3.

¹¹Many 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."

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 3.¹²

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 represent a scalar-valued quadratic function of a vector

If \vec{T} is any tensor of rank 2, then $f(\vec{v}) = \vec{v} \cdot \vec{T} \cdot \vec{v}$ 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**.¹³ Its components in any cartesian system are given by the **Kronecker symbol** δ_{ij} .

Here are some more examples of this idea:

Your Turn 13E

- Show that the kinetic energy of a spinning rigid body is $\frac{1}{2}\vec{\omega} \cdot \vec{J} \cdot \vec{\omega}$, where \vec{J} is the moment of inertia tensor introduced earlier.
- Show that the rate at which work is done pulling a rigid object through viscous fluid equals $\vec{v} \cdot \vec{\zeta} \cdot \vec{v}$.
- Show that the potential energy stored by the spring system is $\frac{1}{2}\vec{r} \cdot \vec{K} \cdot \vec{r}$. Similarly to the kinetic energy of a rigid body, we again see that a symmetric 3-tensor of rank 2 can be used to specify a quadratic function of a vector.
- Explain why the dissipated power density in a general ohmic material is $\vec{E} \cdot \vec{\kappa} \cdot \vec{E}$, another quadratic function of a vector. 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. That's why the moment of inertia, quadrupole, and metric tensors all have this property.

The electric quadrupole moment \vec{Q}_E also defines a contribution to the far potential that depends quadratically on \vec{r} (see the third term of the far potential, Equation 3.2, page 37). Also, like the examples above, \vec{Q}_E has a coordinate representation as a 3×3 matrix, which 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, it is **traceless**; that is, $\text{Tr} \vec{Q}_E = 0$.

¹²Other 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.

¹³This same tensor, regarded as a linear function of a vector, was called $\vec{\mathbf{I}}$ in Equation 13.1.

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 ℓ changes from $\vec{r}(\ell)$ to $\vec{r}(\ell) + d\vec{r}(\ell)$, where¹⁴

$$d\vec{r}(\ell) = d\vec{\Omega} \cdot \vec{r}(\ell), \quad \text{with} \quad \vec{\Omega}_{ij} = d\theta \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}_{ij}. \quad (13.5)$$

This linear function of $\vec{r}(\ell)$ 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.

[T2] Section 13.3' (page 199) offers a connection to quantum mechanics.

13.4 RANK THREE

13.4.1 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 \hat{n} that is perpendicular to the plane that they span and is chosen using the right-hand rule. Let Σ be the area of the parallelogram with the two given vectors as edges, and define the **cross product** as $\hat{n}\Sigma$. This new vector is linear in each of the two that we began with; for example:

- If we double either vector, then Σ doubles and \hat{n} is unchanged, so $\hat{n}\Sigma$ doubles.
- If we replace either vector by its negative, then Σ is unchanged but \hat{n} reverses, so $\hat{n}\Sigma$ 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 number. 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.

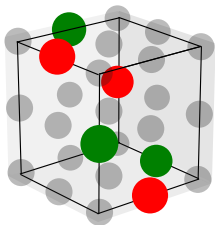
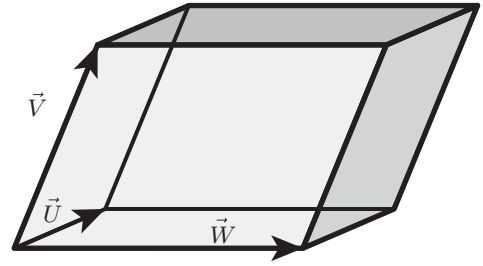


Fig. 0.3 (page 6)

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 $\epsilon(\vec{U}, \vec{V}, \vec{W})$ in Equation 13.6 is positive.



13.4.2 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 $\sigma = -1$):

$$\epsilon(\vec{U}, \vec{V}, \vec{W}) = v\sigma. \quad (13.6)$$

Exchanging any two of the three vectors leaves v unchanged while reversing the sign of σ , so we say that ϵ is *totally antisymmetric*.¹⁵

To see that Equation 13.6 yields a function that is linear in all three of its vector arguments, note:

- If we double the length of any input vector, then v doubles and σ is unchanged, so $v\sigma$ doubles.
- If we replace any input vector by its negative, then v is unchanged but σ is replaced by its negative, so $v\sigma$ 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.

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 v unchanged but changes σ by the sign of the permutation, again like ϵ_{ijk} , so again we find that the Levi-Civita symbols are components of a totally antisymmetric, third-rank, 3-tensor.

When there are more than 2 indices (rank higher than 2), it's too cumbersome to put any glyph above the symbol to indicate tensoriality. Also, in this situation we will never drop the indices, so their presence suffices to announce that ϵ_{ijk} are the components of a 3-tensor of rank three.

¹⁴We encountered this relation earlier in Equation 3.12 (page 44).

¹⁵A totally *symmetric* rank-3 tensor would be *unchanged* under exchange of any of its inputs, just as in rank two.

13.5 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 \mathbf{B} defined in Equation 7.3 (page 100) 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 we also now see that there are **tensor fields**. Later chapters will make extensive use of this concept. In fact, the polarizability 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=f51iqUk0ZTw.

Intermediate:

General: Neuenschwander, 2015; Arfken et al., 2013; Cahill, 2013; Fleisch, 2012; Stone & Goldbart, 2009.

3-tensors as functions of vectors: Thorne & Blandford, 2017, chap. 1.

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.

T₂

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 yet.]

T₂

13.3'a Tensors in quantum mechanics

Quantum mechanics introduces a tensor of rank 2 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, echoing the corresponding statement in the main text about dyad products.

There are also tensor operators in quantum mechanics and tensor representations of internal symmetry groups in high energy physics. All are subject to similar analyses. There are also generalizations to handle intrinsic particle spin, called "spinors" (Section 34.7'b, page 469).

13.3'b Another concept of rank

In this book, the "rank" of a tensor \mathbb{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 \mathbb{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.

PROBLEMS

13.1 *Octahedron I*

A mass distribution consists of six equal point masses m placed at the vertices of an octahedron: $\vec{r}_{(\pm 1)} = (\pm a, 0, 0)$, $\vec{r}_{(\pm 2)} = (0, \pm a, 0)$, $\vec{r}_{(\pm 3)} = (0, 0, \pm a)$. 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.¹⁶

- a. A charge distribution consists of six single point charges e placed at the vertices of an octahedron: $\vec{r}_{(\pm 1)} = (\pm a, 0, 0)$, $\vec{r}_{(\pm 2)} = (0, \pm a, 0)$, $\vec{r}_{(\pm 3)} = (0, 0, \pm a)$. 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: $\vec{r}_{(\pm 1)} = (\pm a, 0, 0)$, $\vec{r}_{(\pm 2)} = (0, \pm a, 0)$. A neutralizing charge $-4e$ is placed at the origin. Find the electric dipole and quadrupole moments.

¹⁶See Section 3.6.4, page 41.

CHAPTER 14

Tensors from Heaven

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 invariant, changes if we consider a different object.
- Two of the examples were different: The 3D metric tensor (Section 13.3.4, page 195) 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 (right-handed) 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 constant tensor \vec{K} as a function that eats a displacement vector and returns a number, the stored potential energy $\frac{1}{2}\Delta\vec{r}\cdot\vec{K}\cdot\Delta\vec{r}$. This function is quadratic in the components of \vec{r} . It can be represented in any coordinate system by a matrix of ordinary numbers. We call those numbers the **components** of \vec{K} in the chosen coordinate system, and denote them by \vec{K}_{ij} . It's important that the nine *numbers* \vec{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

$$\vec{r}'_a = S_{ai}\vec{r}_i. \quad (14.1)$$

Then the same spring potential energy function as before can also be written as $\frac{1}{2} \Delta \vec{r}' \cdot \vec{K}' \cdot \Delta \vec{r}'$, where the new components are determined by

$$\vec{r}_i \vec{K}_{ij} \vec{r}_j = \vec{r}'_a \vec{K}'_{ab} \vec{r}'_b = \vec{r} \cdot (\mathbf{S}^t \vec{K}' \mathbf{S}) \cdot \vec{r}.$$

This must hold for any spring displacement, so $\vec{K} = \mathbf{S}^t \vec{K}' \mathbf{S}$, or

$$\vec{K}'_{ab} = S_{ai} S_{bj} \vec{K}_{ij}. \quad (14.2)$$

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 system¹). What's special about these "cartesian" coordinate systems is that the distance-squared between two points always takes the pythagorean form²

$$\|\Delta \vec{r}\|^2 = \sum_i \Delta \vec{r}_i \Delta \vec{r}_i \quad \text{in cartesian coordinates.} \quad (14.3)$$

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 \mathbf{S} is specifically an **orthogonal matrix**, that is, one for which

$$\mathbf{S} \mathbf{S}^t = \mathbf{S}^t \mathbf{S} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (14.4)$$

Then the new coordinates again have the property that the length-squared of a vector equals $\Delta \vec{r}'_a \Delta \vec{r}'_a$, which has the same form as Equation 14.3. For future use, note that Equation 14.4 implies

$(\det \mathbf{S})^2 = 1, \text{ and hence } \det \mathbf{S} = \pm 1 \text{ for an orthogonal matrix.}$

(14.5)

The case with positive determinant contains all rotations; the other case consists of spatial inversion combined with a rotation.³

¹Later we will upgrade to coordinate systems on space and time (four dimensions). For now, we consider three-dimensional space only.

²If 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. If we use tensors on a non-euclidean space, or with curvilinear coordinates, the distinction becomes essential.

³Some 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 and 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 constant tensor $\vec{\vec{K}}$, as any set of nine numbers that depend on a choice of cartesian coordinates and that transform like the components of $\vec{r} \otimes \vec{r}$. 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 $\vec{\vec{\mathbb{I}}}$ to be that 2-tensor (14.6) whose components in this coordinate system are δ_{ij} .

The corresponding quadratic function defined by $\vec{\vec{\mathbb{I}}}$ 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 $\vec{\vec{\mathbb{I}}}$? To investigate, let's see how the components of your \vec{g} look in your friend's (primed) coordinate system:

$$\vec{\vec{\mathbb{I}}}'_{ab} = S_{ai} S_{bj} \delta_{ij} = S_{aj} S_{bj} = [SS^t]_{ab} = \delta_{ab}, \quad (14.7)$$

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. Of course, in this case that conclusion is a tautology, not a surprise, because we explicitly restricted attention to those "good" coordinate systems for which it is true. 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 ϵ are the same in any right-handed cartesian system

Section 13.4 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 ± 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? Then you must show that if we start in one cartesian system, then transform to any other, that the components are *numerically the same* as before. Then the tensor that they define won't actually depend on my 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, but the details are interesting (in part because we will later use the same approach to generalize to four dimensions).

Suppose that we have a space that is euclidean, and moreover we have agreed

that one of the cartesian coordinate systems will be called “right-handed.”⁴ We now define a 3-tensor by stating its components as in Section 0.2.2 (page 5):

$$\begin{aligned}\varepsilon_{ijk} &= 0 \text{ if any two of the indices are equal;} \\ \varepsilon_{ijk} &= +1 \text{ if } i, j, k \text{ are an even permutation of } 1, 2, 3; \\ \varepsilon_{ijk} &= -1 \text{ if } i, j, k \text{ are an odd permutation of } 1, 2, 3.\end{aligned}\tag{14.8}$$

Next, we must calculate the new components

$$\varepsilon'_{abc} = S_{ai}S_{bj}S_{ck}\varepsilon_{ijk}.\tag{14.9}$$

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, ε_{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 ε'_{abc} are equal, in agreement with 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, ε'_{123} . And of the 27 terms being summed in Equation 14.9, *all but six are zero*:

$$\begin{aligned}\varepsilon'_{123} &= S_{11}S_{22}S_{33} + S_{12}S_{23}S_{31} + S_{13}S_{21}S_{32} - S_{11}S_{23}S_{32} - S_{13}S_{22}S_{31} - S_{12}S_{21}S_{33} \\ &= \det S.\end{aligned}$$

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,

- The determinant must always equal ± 1 ;
- It's $+1$ for the identity operator (rotation by zero degrees); and
- It cannot change discontinuously;⁵ but
- 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, $\varepsilon'_{123} = +1$, completing the proof that all components are the same in any right-handed system.

14.3.2 Components only specify a unique ϵ after a right-hand convention is chosen

Had we used Equation 14.8 in conjunction with a *left*-handed system, then we would have defined a *different* Levi-Civita tensor, equal to minus the one in Equation 14.8.

⁴Mathematicians call this binary choice an “orientation” on 3-space. It has nothing to do with your hands, which side of your body your heart is on, nor the shape of your DNA. *Any* cartesian coordinate system may be singled out and given this status.

⁵After all, the determinant of a matrix is just a polynomial in the entries of that matrix.

You can confirm that by re-expressing it in terms of a right-handed system, because in that calculation, $\det \mathbf{S} = -1$.⁶ 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 $d^2\vec{\Sigma}$) requires that we commit to a convention about which is our “right” hand. For this reason, some books refer to the “Levi-Civita *pseudotensor*.”⁷ We’ll instead take the viewpoint that ε is a perfectly well-defined 3-tensor, once we have made a choice for which coordinate systems we will call right-handed.

14.3.3 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 $\delta_{ij}\delta_{kl}$), 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 \vec{v} lying in the xy plane (Figure 14.1). Let θ be the angle between \vec{u} and \vec{v} . Thus, $\vec{u} = (1, 0, 0)$ and $\vec{v} = (v \cos \theta, v \sin \theta, 0)$. The sum $\vec{u}_i\vec{v}_i = uv \cos \theta$ 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 $\varepsilon_{ijk}\vec{u}_j\vec{v}_k$ using any right-handed coordinate system we like. The three

⁶An orthogonal matrix with determinant -1 corresponds to a rotation combined with a *reflection* through a plane, or through a point, and therefore reverses the handedness of a coordinate system. Any two left-handed coordinate systems can be continuously connected by a family of rotations, and you can easily find examples where the determinant is -1 , so all must have this property.

⁷Similarly, a vector quantity that depends on a choice of handedness is sometimes disparaged by the prefix “pseudo.” For example, Chapter 15 will package the three numbers needed to represent a magnetic field as \vec{B}_i , which is sometimes called a “pseudovector”; also the usual components of angular momentum, angular velocity, and torque are pseudovectors. There are even pseudoscalars, single quantities that change sign upon change of handedness, such as the field that when quantized represents the pion.

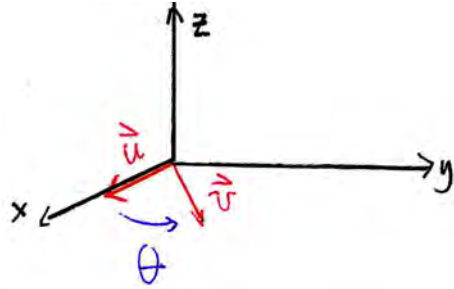


Figure 14.1: In this right-handed coordinate system, the angle from \vec{u} to \vec{v} is positive; the angle from \vec{v} to \vec{u} is negative

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 $\beta\varepsilon_{ijk}\vec{u}_j\vec{u}_k$. 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} parallel to \vec{u} and \vec{v} in the xy plane (Figure 14.1). This time, however, we must be careful to specify that θ is the angle from \vec{u} to \vec{v} , and that θ is taken to be positive if that angle is counterclockwise when viewed along the z axis from positive toward negative values of z (Figure 14.1). Then again $\vec{u} = (u, 0, 0)$ and $\vec{v} = (v \cos \theta, v \sin \theta, 0)$ and

$$(\vec{u} \times \vec{v})_3 = \varepsilon_{31k}u\vec{v}_k = \varepsilon_{312}u(v \sin \theta) = uv \sin \theta, \quad (14.10)$$

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 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 note that

$$\varepsilon_{ijk}\varepsilon_{ijk} = \sum_{\text{permutations}} (\pm 1)^2 = 6. \quad (14.11)$$

Next, try the same expression but don’t set the last indices equal nor sum them: $\varepsilon_{ijk}\varepsilon_{ij\ell}$ is an invariant symmetric tensor of rank 2, so it must be a multiple of $\delta_{k\ell}$. To find the constant of proportionality, set $k = \ell$, sum over i , and compare to Equation 14.11.

This gives

$$\varepsilon_{ijk}\varepsilon_{ijl} = 2\delta_{kl}. \quad (14.12)$$

Finally, try not setting the last *two* indices equal: $\varepsilon_{ijk}\varepsilon_{iml}$ is an invariant tensor of rank 4, and it's antisymmetric upon exchange of jk as well as $m\ell$. But it's symmetric if we swap jk with $m\ell$. Suppose that $j = 1, k = 2$; then only one term of the sum over i is nonzero, namely $i = 3$. This in turn implies that m, ℓ must be either 12 or 21. For all those reasons, we must have

$$\varepsilon_{ijk}\varepsilon_{iml} = M(\delta_{jm}\delta_{k\ell} - \delta_{j\ell}\delta_{km}) \quad \text{for some constant } M. \quad (14.13)$$

To evaluate M , this time set $m = j$ and $\ell = k$, sum both, and again compare to Equation 14.11:

$$6 = \varepsilon_{ijk}\varepsilon_{ijk} = M(\delta_{jj}\delta_{kk} - \delta_{jk}\delta_{kj}) = M(3 \cdot 3 - \delta_{jj}) = 6M.$$

Thus, $M = 1$ in Equation 14.13:

$$\varepsilon_{ijk}\varepsilon_{iml} = M(\delta_{jm}\delta_{k\ell} - \delta_{j\ell}\delta_{km}). \quad (14.14)$$

Your Turn 14A

Try using one of the three preceding identities to get a familiar formula for $\vec{u} \times (\vec{v} \times \vec{w})$.

Your Turn 14B

The arguments given above may seem too informal.

- Rederive Equation 14.12 directly, as follows: In the summation, the index pair ij can only take the six possible values (12), (21), (13), (31), (23), (32). For each of the corresponding terms, list the possible values of k and ℓ to which that term could contribute, and in this way verify the identity.
- Rederive Equation 14.13 directly, as follows: There are three terms in the sum. For each one, enumerate the possible values of the index pairs jk and $m\ell$ to which that term could contribute, and in this way verify the identity.

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 ε to express cross products will become clear. Also, the approach used in this book 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 concept of cross product. The argument is the same as the one in Section 14.3.1.

Finally, understanding ε as a tensor will prove valuable as we seek to reformulate electrodynamics *without* any cross products, thereby making its inversion invariance obvious.

FURTHER READING

Intermediate:

Neuenschwander, 2015; Arfken et al., 2013; Cahill, 2013; Fleisch, 2012; Stone & Goldbart, 2009.

T2 Twisted objects: Burke, 1985.

T₂

14.3' Spatial inversion invariance

1. The appearance of the Levi-Civita tensor in a law of Nature should bother you! Classical mechanics and electrodynamics are supposed to be *invariant* under spatial inversions, so why do we need any right-hand rule (or equivalently any choice of right-handed coordinates) to formulate them? The answer is: We don't. Both classical mechanics and electrodynamics can be expressed completely without ever introducing cross products or pseudoquantities. In fact, doing this for electrodynamics, and hence making its invariance under spatial inversions ("parity invariance") manifest, is one of our goals of this book.⁸

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 neutrino has a preferred helicity. There is a spin operator analogous to ε that appears in the weak interaction, that changes under spatial inversion, and that cannot be removed by redefining things.

2. We don't get a Levi-Civita tensor until we select a "handedness," that is, select one privileged class of cartesian coordinate systems that we call "right-handed." Mathematicians call this binary choice an "orientation," but beware: That term can lead to confusion with the everyday sense of the word, which is 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 handedness convention of space."

3. What does "from Heaven" mean? Our constructions all relied on choosing cartesian coordinates. In fact, with some 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 $\vec{\mathbf{I}}$ 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 plus a distinction between left- and right-handed coordinate systems: The geometric construction of Section 13.4 works on any such space and does not require any coordinate choice.

T₂

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 (they "point").
2. The Levi-Civita tensor, and things constructed with its assistance, are ambiguous (ill-defined) until we choose an orientation (choice of which hand is "right"). 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

⁸We'll begin in Chapter 15. Doing the same thing for rigid-body dynamics is similarly rewarding; see Problem 15.4.

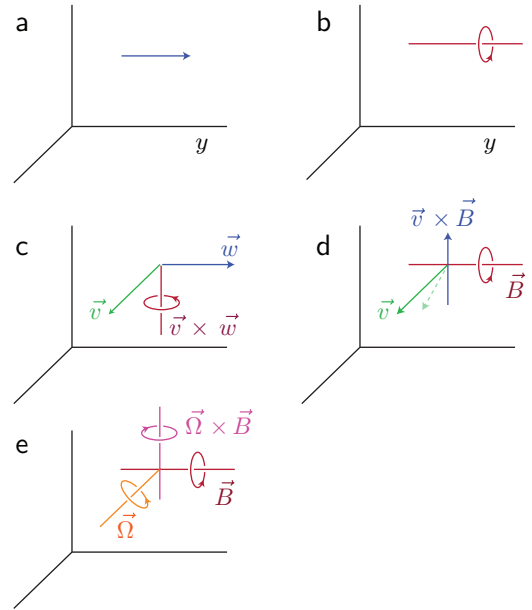


Figure 14.2: [Diagrams.] **Twisted vectors and their operations.** (a) An ordinary vector. The reflection $y \rightarrow -y$ turns it into minus itself. (b) A twisted vector. The reflection $y \rightarrow -y$ leaves it unchanged. (c) Cross product of vector with vector yields a twisted vector (see text). (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).

“twisted” to specify these objects.⁹ Thus, angular momentum and magnetic induction \vec{B} are twisted vectors, whereas velocity and force are ordinary vectors.

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.¹⁰

⁹Others speak of “tensor densities.”

¹⁰This construction also lets us associate an antisymmetric rank-2 tensor $\vec{\omega}$ to any twisted vector

- 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. We regard the magnetic field as an ordinary vector defined with the help of some choice of right-hand convention, and similarly for cross products. Eventually we will eliminate “pseudo” quantities from our formulation of electrodynamics altogether.

\vec{B} and vice versa: The tensor takes any vector \vec{v} and returns the vector $\vec{\omega} \cdot \vec{v} = \frac{1}{2} \vec{v} \times \vec{B}$, which is Equation 15.3 (page 214).

PROBLEMS

14.1 *Dots and crosses*

Prove the identity $(\vec{A} \times \vec{B}) \cdot (\vec{C} \times \vec{D}) = (\vec{A} \cdot \vec{C})(\vec{B} \cdot \vec{D}) - (\vec{A} \cdot \vec{D})(\vec{B} \cdot \vec{C})$.

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 seems reasonable to add, “That average will always be a constant times the identity tensor.” Let’s prove this.

Note first that the rotational average must itself be a rotationally-invariant, symmetric 3-tensor. Call it $\vec{\vec{A}}$; then its matrix of components in some coordinate system must have the property that $\mathbf{S}^t \mathbf{A} \mathbf{S} = \mathbf{A}$ for any rotation matrix \mathbf{S} . In particular, this property holds for any *infinitesimal* rotation. Recall from Equation 13.5 (page 196) that an infinitesimal rotation is given by $\mathbf{S} = \mathbf{1} + \epsilon \mathbf{T} + \mathcal{O}(\epsilon^2)$, where \mathbf{T} is an antisymmetric matrix and $\mathbf{1}$ is the identity matrix.

Work out the consequences of invariance under such transformations (to order ϵ) and prove that \mathbf{A} is a constant times $\mathbf{1}$.

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 $\vec{v}(\vec{r})$ (the “order parameter”). The energy cost to be in one of these states is some analytic, local, rotationally invariant function of \vec{v} and its derivatives, integrated over space. Because it’s analytic, we can expand that function in Taylor series as a polynomial in the components of \vec{v} . Clearly the part of this function with no derivatives must involve only *even powers* of the components \vec{v}_i . This trivial fact has profound consequences for the phase-transition behavior of ferromagnets.

Now suppose someone else tell you that some kind of matter (a “nematic liquid crystal”) has states characterized by a spatially varying, symmetric, traceless rank-2 tensor $\vec{\vec{M}}$. The free energy cost to be in one of these states is some analytic, local, rotationally invariant function of $\vec{\vec{M}}$ and its derivatives, integrated over space. Because it’s analytic, we can expand that function in Taylor series as a polynomial in the components of $\vec{\vec{M}}$. The part of this function with no derivatives must be at least quadratic in the components of $\vec{\vec{M}}$ (why?).

Only a few terms are allowed in the free energy function of a nematic liquid crystal.

Do: Now find all possible contributions to the free energy cost function (if any) that are quadratic or cubic in the components of $\vec{\vec{M}}$ (again, only find terms with no derivatives). Your answer has profound consequences for the phase-transition behavior of nematic liquid crystals.

CHAPTER 15

Magnetostatics

Oersted 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, Oersted was apparently all “thumbs” in the lab, and all his experiments had to be carried out by his students and assistants.

— *R. M. Clegg*

You are quite right to say that it is inconceivable that 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 [Oersted’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. This simplification is justified if charges are motionless, and 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 executing 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 yet.].

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 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 on a test charge (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:¹

$$(\vec{\text{force}}) = 2q\vec{\omega} \cdot \vec{v}, \text{ that is, } f_i = 2q\vec{\omega}_{ij}v_j. \quad (15.1)$$

After someone sets up a current distribution in the lab, we can operationally measure the resulting field $\vec{\omega}$ by throwing a lot of charged test bodies and seeing how they accelerate.

Magnetic force depends linearly on velocity and is directed perpendicular to it.

Moreover, experimentally the new force has another unusual property:²

- The force is always perpendicular to the test charge's velocity.

This observation implies that the current specifically creates an *antisymmetric* rank-two tensor $\vec{\omega}$. To see this, think about two velocities \vec{v} and \vec{u} . Then $\vec{\omega} \cdot (\vec{v} + \vec{u})$ must be perpendicular to $(\vec{v} + \vec{u})$:

$$0 = (\vec{v} + \vec{u}) \cdot \vec{\omega} \cdot (\vec{v} + \vec{u}) = \vec{v} \cdot (\vec{\omega} \cdot \vec{v}) + \vec{u} \cdot (\vec{\omega} \cdot \vec{u}) + \vec{v} \cdot \vec{\omega} \cdot \vec{u} + \vec{u} \cdot \vec{\omega} \cdot \vec{v}.$$

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 $\vec{\omega}$ 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: Define the three quantities

$$\vec{B}_i = \varepsilon_{ijk}\vec{\omega}_{jk}. \quad (15.2)$$

In a sense, we lose nothing by this reformulation, because it is invertible: We can always recover $\vec{\omega}$ from \vec{B} :

Your Turn 15A

Show that

$$\vec{\omega}_{im} = \frac{1}{2}\varepsilon_{kim}\vec{B}_k. \quad (15.3)$$

(Where did the factor of 1/2 come from?)

Introductory texts formulate magnetism in terms of \vec{B} , and so will we at first. But there is a price to pay for this approach:

¹We 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 it makes another 2 elsewhere go away.

²See Media 6.

- Equation 15.2 introduces 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 $\vec{f} = q\vec{v} \times \vec{B}$.
- 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.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.

[T₂] Section 15.2’ (page 226) raises a puzzle about velocity-dependent forces.

15.3 VECTOR POTENTIAL

15.3.1 No scalar potential this time

In electrostatics, the four equations $\vec{\nabla} \cdot \vec{E} = \rho_q/\epsilon_0$ and $\vec{\nabla} \times \vec{E} = \vec{0}$ 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, $\vec{\nabla} \times \vec{E} = \vec{0}$, in terms of ψ , so we could just substitute $\vec{E} = -\vec{\nabla}\psi$ into the Gauss law and *forget* Faraday. Can we duplicate that victory?

At first it looks bad. The magnetic field is not curl-free: Ampère’s law says $\vec{\nabla} \times \vec{B} \neq 0$. 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, $\vec{\nabla} \cdot \vec{B} = 0$, 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

³ **[T₂]** 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 Maxwell’s equations in vacuum are invariant (Section 34.9’, page 470).

($u\vec{r}$). Make sure that you understand how the chain rule implies that

$$\frac{\partial g}{\partial u} = \frac{\partial f}{\partial \vec{r}_i} \Big|_{u\vec{r}} \frac{\partial(u\vec{r}_i)}{\partial u} = \vec{r} \cdot \vec{\nabla} f \Big|_{u\vec{r}} \quad (15.4)$$

$$\frac{\partial g}{\partial \vec{r}_1} = \frac{\partial f}{\partial \vec{r}_i} \Big|_{u\vec{r}} \frac{\partial(u\vec{r}_i)}{\partial \vec{r}_1} = (\vec{\nabla}_i f \Big|_{u\vec{r}})(u\delta_{i1}) = u\vec{\nabla}_1 f \Big|_{u\vec{r}}, \quad (15.5)$$

and similar results for $\partial g/\partial \vec{r}_{2,3}$. Think about how the indices match on each side of these formulas.

15.3.3 Revisit electrostatics

The magnetic Gauss law $\vec{\nabla} \cdot \vec{B} = 0$ looks pretty different from $\vec{\nabla} \times \vec{E} = \vec{0}$, 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 $\vec{\nabla} \times \vec{E} = \vec{0}$, 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 ψ .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 ψ 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.

Show that the curl-free condition is equivalent to

Ex.
$$\vec{\nabla}_i \vec{E}_j - \vec{\nabla}_j \vec{E}_i = 0 \text{ for any } i \text{ and } j \text{ (stationary case)}. \quad (15.6)$$

Solution: One way is to write out explicitly each component of the equation, for example, $(\vec{\nabla} \times \vec{E})_1 = \vec{0}$ and so on. But let's get some practice with Levi-Civita identities: Take the curl and contract it with ε :

$$0 = \varepsilon_{ijk}(\vec{\nabla} \times \vec{E})_k = (\varepsilon_{ijk}\varepsilon_{klm})\vec{\nabla}_\ell \vec{E}_m.$$

Now use the identity Equation 14.13 (page 207) to simplify the factor in parentheses:

$$= (\delta_{i\ell}\delta_{jm} - \delta_{im}\delta_{j\ell})\vec{\nabla}_\ell \vec{E}_m = \vec{\nabla}_i \vec{E}_j - \vec{\nabla}_j \vec{E}_i.$$

As in Chapter 2, we integrate \vec{E} along the chosen path from a reference point (for simplicity the origin) to a desired field point \vec{r} . We can express that path by the formula $u\vec{r}$ where u ranges from 0 to 1. Substituting into Equation 2.2 (page 28) gives

$$\psi(\vec{r}) = - \int_0^1 (\vec{r} du) \cdot \vec{E}(u\vec{r}). \quad (15.7)$$

In this expression, \vec{r} is held constant during the integration over u . Then the negative

⁴Section 2.2.1 (page 27).

gradient is (see Equation 15.5)

$$\begin{aligned} -\frac{\partial\psi}{\partial\vec{r}_i} &= \int_0^1 du \left[\vec{E}_m(u\vec{r}) \frac{\partial\vec{r}_m}{\partial\vec{r}_i} + \vec{r}_m \frac{\partial\vec{E}_m}{\partial\vec{r}_k} \Big|_{u\vec{r}} \frac{\partial(u\vec{r}_k)}{\partial\vec{r}_i} \right] \\ &= \int_0^1 du \left[\vec{E}_i(u\vec{r}) + u\vec{r}_m \frac{\partial\vec{E}_m}{\partial\vec{r}_i} \Big|_{u\vec{r}} \right]. \end{aligned} \quad (15.8)$$

In the last term, we may replace $\partial\vec{E}_m/\partial\vec{r}_i$ by $\partial\vec{E}_i/\partial\vec{r}_m$, thanks to Equation 15.6. We can now use Equation 15.4 and the Fundamental Theorem of Calculus to find

$$-\vec{\nabla}_i\psi|_{\vec{r}} = \int_0^1 du \frac{d}{du} [u\vec{E}_i(u\vec{r})] = u\vec{E}_i(u\vec{r})|_0^1 = \vec{E}_i(\vec{r}).$$

Once again:

- We have established the potential representation for electrostatics.
- It relies on the curl-free condition, even though we did not explicitly use Stokes's theorem.
- There is an ambiguity in ψ : Adding a constant to ψ , 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 $\vec{\nabla} \times \vec{E} \neq 0$.⁶

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 15B

Use Equation 15.2 to show that the magnetic Gauss law is equivalent to

$$\varepsilon_{imk} \vec{\nabla}_k \vec{\omega}_{im} = 0. \quad (15.9)$$

That is, when we take all the first derivatives of $\vec{\omega}_{im}$ and *antisymmetrize*, the result is always zero. This resembles Equation 15.6, albeit with an extra index.

⁵In fact, Chapter 2 found a complete, general solution to electrostatics with a specified charge distribution.

⁶Chapter 18 will pick up this loose thread.

Your Turn 15C

Show that, of the six nonzero terms in Equation 15.9, half are redundant; that is, it may be written as

$$\vec{\nabla}_k \vec{\omega}_{im} + (2 \text{ cyclic permutations}) = 0 \quad (15.10)$$

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, define⁷

$$\vec{A}_i(\vec{r}) = 2 \int_0^1 (u \vec{r}_m \, du) \vec{\omega}_{mi}(u \vec{r}). \quad (15.11)$$

We want the curl of this new vector field, or equivalently

$$\begin{aligned} \vec{\nabla}_k \vec{A}_i - \vec{\nabla}_i \vec{A}_k &= 2 \int_0^1 du u \left[\frac{\partial \vec{r}_m}{\partial \vec{r}_k} \vec{\omega}_{mi}(u \vec{r}) + \vec{r}_m \frac{\partial \vec{\omega}_{mi}}{\partial \vec{r}_n} \Big|_{u \vec{r}} \frac{\partial (u \vec{r}_n)}{\partial \vec{r}_k} \right] - (i \rightleftharpoons k) \\ &= 2 \int_0^1 du u \left[(\vec{\omega}_{ki} - \vec{\omega}_{ik}) + u \vec{r}_m \left(\frac{\partial \vec{\omega}_{mi}}{\partial \vec{r}_k} - \frac{\partial \vec{\omega}_{mk}}{\partial \vec{r}_i} \right) \Big|_{u \vec{r}} \right]. \end{aligned}$$

The first two terms can be written as $2\vec{\omega}_{ki}$. The last two terms can be simplified by using Equation 15.10: They equal $-\vec{\nabla}_m \vec{\omega}_{ik} \Big|_{u \vec{r}}$.

Analogously to Equation 15.8, we therefore get

$$\begin{aligned} &= 2 \int_0^1 du \left[2u \vec{\omega}_{ki}(u \vec{r}) - u^2 \vec{r}_m \frac{\partial \vec{\omega}_{ik}}{\partial \vec{r}_m} \Big|_{u \vec{r}} \right] \\ &= 2 \int_0^1 du \frac{\partial}{\partial u} \left[u^2 \vec{\omega}_{ki}(u \vec{r}) \right] = 2u^2 \vec{\omega}_{ki}(u \vec{r}) \Big|_0^1 = 2\vec{\omega}_{ki}(\vec{r}). \end{aligned} \quad (15.12)$$

Now tidy things up by recalling the formula for curl and Equation 15.2:

$$(\vec{\nabla} \times \vec{A})_m = \varepsilon_{mki} \vec{\nabla}_k \vec{A}_i = \frac{1}{2} \varepsilon_{mki} (\vec{\nabla}_k \vec{A}_i - \vec{\nabla}_i \vec{A}_k) = \varepsilon_{mki} \vec{\omega}_{ki} = \vec{B}_m.$$

Indeed, Equation 15.11 has constructed a vector field \vec{A} whose curl equals \vec{B} . So we'll call \vec{A} the **magnetic vector potential**.

Our payoff for this level of abstraction is that the result we proved works in *any number* of dimensions:

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).

Poincaré lemma

(15.13)

Later, when we need this result in four dimensions, we won't need to prove it again.⁸

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 former and \underline{A} for the latter.

⁸It 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.

15.4 GAUGE INVARIANCE AND COULOMB GAUGE

We have found the general solution to the magnetic Gauss law, so we can just substitute $\vec{\nabla} \times \vec{A}$ 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, $\vec{A} \rightarrow \tilde{\vec{A}} = \vec{A} + \vec{\nabla}\Xi$, 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 **gauge invariance**, though maybe “redundancy” would have been a better term to use. The substitution $\vec{A} \rightarrow \tilde{\vec{A}}$ is called a **gauge transformation** of the vector potential. This is much more freedom than what we had in electrostatics, where adding a *constant* to ψ 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 (**gauge fixing**). For example, we can always insist that \vec{A} obeys

$$\vec{\nabla} \cdot \vec{A} = 0. \quad \text{Coulomb gauge condition} \quad (15.14)$$

To see this, suppose that we represent a \vec{B} field by some vector potential that doesn’t satisfy Coulomb gauge. If we then gauge transform it we get $\vec{\nabla} \cdot \tilde{\vec{A}} = \vec{\nabla} \cdot \vec{A} + \nabla^2 \Xi$. We just need to choose Ξ to be a function of position that solves the Poisson equation⁹ with source given by $\vec{\nabla} \cdot \vec{A}$. After that gauge transformation, $\tilde{\vec{A}}$ is in Coulomb gauge.

15.5 BACK TO PHYSICS

15.5.1 Steady currents

To avoid distraction from electrostatics, let’s temporarily assume that there is no free net charge ($\rho_q(\vec{r}) = 0$); 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 also temporarily restricts to situations with *steady* motion ($\partial\vec{j}/\partial t = 0$). Thus, our system will be invariant under time *translation* (it is **stationary**), though not under time *reversal* (it is not **static**). Somewhat inconsistently, the study of such situations is often called **magnetostatics**.

Stationarity only be 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 situation just described, they simplify

⁹Section 2.4.2 (page 30) found the general solution to the Poisson equation.

to just

$$\begin{aligned}
 \vec{\nabla} \cdot \vec{E} &= \rho_q / \epsilon_0 = 0 && \text{Gauss (no net charge)} \\
 \vec{\nabla} \cdot \vec{B} &= 0 && \text{Gauss} \\
 \vec{\nabla} \times \vec{B} &= \mu_0 \vec{j} && \text{Ampère (stationary case)} \\
 \vec{\nabla} \times \vec{E} &= \vec{0}. && \text{Faraday (stationary case)}
 \end{aligned} \tag{15.15}$$

These equations have decoupled into two that involve \vec{E} only, 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.

15.5.2 Axial symmetry suggests a solution to the Oersted 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 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:

$$\begin{aligned}
 \text{Ampère: } \int d^2\vec{\Sigma} \cdot (\vec{\nabla} \times \vec{B}) &= \mu_0 \int d^2\vec{\Sigma} \cdot \vec{j} = \mu_0 I \\
 \text{Stokes: } = \oint_{\text{rim}} \vec{B}(r_*) \cdot d\vec{r}_* &= \int_0^{2\pi} (w d\varphi) \vec{B} \cdot \hat{\phi} = 2\pi w f(w).
 \end{aligned}$$

We conclude that $f(w) = \mu_0 I / (2\pi w)$ for any w larger than the wire's radius, and hence that

$$\vec{B}(\vec{r}) = \hat{\phi} \frac{\mu_0 I}{2\pi \|\vec{r}\|} \tag{15.16}$$

works—the famous answer.

Your Turn 15D

That answer looks bad at $\|\vec{r}\| = 0$. Is there a problem? [*Hint*: Think back to Section 1.5, page 19.]

Other problems are harder than this one, however. We need a more systematic approach.

15.5.3 The electrostatic Green function 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 15E

To see the power of this observation, first show that Ampère's law may be written as¹⁰

$$\nabla^2 \vec{A} = -\mu_0 \vec{j} \quad \text{in Coulomb gauge} \quad (15.17)$$

That scary vector partial differential equation has magically separated into *three independent copies* of the Poisson equation. And we already know how to solve the Poisson equation, from electrostatics (Equation 2.6, page 30)! For each component of \vec{j} , compute

$$\vec{A}_i(\vec{r}) = \mu_0 \int d^3 r_* \frac{\vec{j}_i(\vec{r}_*)}{4\pi \|\vec{r} - \vec{r}_*\|}. \quad (15.18)$$

So we just *finished* magnetostatics, for situations where we are given a stationary 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!

Your Turn 15F

Work out the divergence of the expression in Equation 15.18. [*Hint:* Use the continuity equation to show that $\vec{\nabla} \cdot \vec{j}$ must always be zero in a steady situation.]

15.5.5 Some of the equations are vacuous, 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 13). And yet, we reformulated electrostatics as one equation in one unknown: the Poisson equation (Equation 2.4, page 28). 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 four unknown potential functions.

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 $0 = 0$ 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

$$\vec{\nabla} \cdot (\vec{\nabla} \times \vec{B}) = \vec{\nabla} \cdot \vec{j} \quad (\text{stationary case}).$$

¹⁰The 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.

The left side is identically zero, regardless of what \vec{B} may be; the right side is also identically 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 Oersted, 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 Oersted 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

$$\vec{j}(\vec{r}) = I\delta^{(2)}(\vec{r}_\perp)\hat{z}. \quad (15.19)$$

Here \vec{r}_\perp denotes the two-component vector $[\begin{smallmatrix} x \\ y \end{smallmatrix}]$. Each delta function contributes a dimension \mathbb{L}^{-1} , so this expression has dimensions appropriate for a charge flux.¹¹ We already found the resulting magnetic field in Section 15.5.1.

Your Turn 15G

- 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 Good Guess, then check and adjust it.]
- Confirm that the vector potential you found really is in Coulomb gauge, as we argued generally must be the case.
- 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 Oersted problem and find the magnetic field created by an *arbitrary* current distribution.

Your Turn 15H

Show that the curl of Equation 15.18 is

$$\vec{B}(\vec{r}) = \frac{\mu_0}{4\pi} \int d^3r_* \vec{j}(\vec{r}_*) \times \frac{\vec{r} - \vec{r}_*}{\|\vec{r} - \vec{r}_*\|^3}. \quad \text{stationary case} \quad (15.20)$$

This is a generalization of the usual Biot–Savart formula to cover an arbitrary current distribution (not necessarily confined to a thin wire).

¹¹See Section 0.3.8 (page 10).

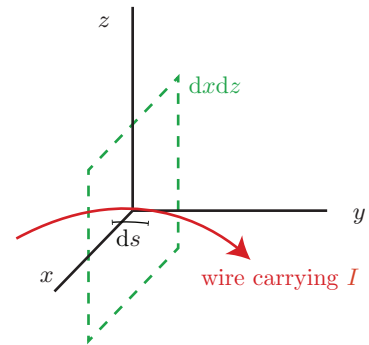


Figure 15.1: Thin-wire idealization.

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 through any cross-section of the wire has everywhere a constant value. 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 \hat{y} . Chapter 8 explained how to find the y -component of the charge flux: Find the net charge crossing the surface element shown in the figure, from smaller to larger y , during time dt . That charge equals $I dt$ if the element $dx dz$ includes the wire (at the origin); otherwise, it’s zero. Idea 8.5 (page 114) defined \vec{j}_2 as a function that, when integrated over $dx dz dt$, yields this charge. Thus, our chunk has

$$\vec{j}_2(t, x, 0, z) = I \delta(x) \delta(z) \hat{y}. \quad (15.21)$$

Think about why this formula has the units appropriate for a charge flux.

We can now 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

$$d\vec{j}(\vec{r}) = I \delta^{(3)}(\vec{r} - \vec{\ell}(s)) \frac{d\vec{\ell}}{ds} ds. \quad \text{short segment of thin wire} \quad (15.22)$$

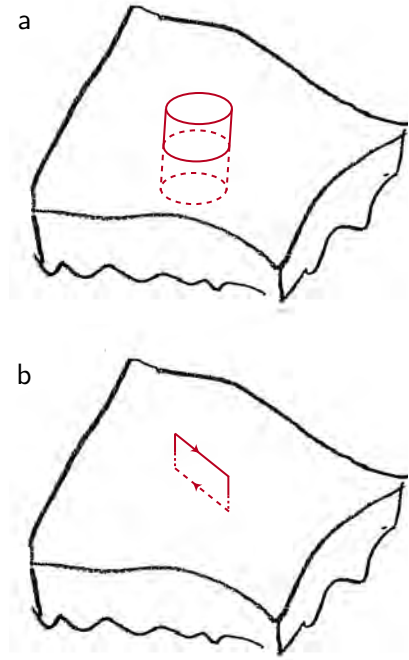
This formula has the same dimensions as Equation 15.21, but it’s no longer restricted to any special coordinate system, nor to one particular point on the wire. To get the charge flux for the entire wire, integrate Equation 15.22 over its entire arc length.

Your Turn 15I

Substitute Equation 15.22 into Equation 15.20 and recover the usual form of the Biot–Savart law.

Figure 15.2: [Sketches.] **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 \vec{B} perpendicular to the surface must be the same just inside and outside the material (Equation 15.23).

(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 component of \vec{B} parallel to the surface may jump if there is a surface current layer (Equation 15.24).



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 electrostatics¹² but without any dependence on the behavior of charges or currents at the surface:

$$\Delta \vec{B}_\perp = 0. \quad (15.23)$$

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 $\vec{B}_\perp = 0$ 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 currents confined to the surfaces of one or both of the media, so let $\vec{j}^{(2D)}$ denote the net 2D charge flux¹³ (with units A/m).

Your Turn 15J

Show that

$$\Delta \vec{B}_\parallel = \mu_0 \vec{j}^{(2D)} \times \hat{n}. \quad (15.24)$$

Here $\Delta \vec{B}_\parallel = (\vec{B}^{[2]} - \vec{B}^{[1]})_\parallel$ and \hat{n} is the unit perpendicular vector pointing from region 1 to region 2.

¹²See Figure 15.2a; compare Section 6.10 (page 86).

¹³Some books call this quantity “surface current density.”

15.8 MAGNETOENCEPHALOGRAPHY

[Not ready yet.]

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 that involves a potential (in this case a vector potential), and it *always works*, even in nonstationary situations (because $\vec{\nabla} \cdot \vec{B} = 0$ always). Chapter 34 will find an even more powerful object that combines the vector potential with electric potential, and that, unlike our previous construction of ψ , remains valid beyond statics.

[T2] Section 15.9'a (page 226) 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.

[T2] Differential forms and Poincaré lemma: Spivak, 1999, vol. 1; Hubbard & Hubbard, 2007; Burke, 1985.

Technical:

Parker bound: Parker, 2007.

T₂

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.

T₂

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 \vec{B} implies that $\vec{\nabla} \cdot \vec{B} \neq 0$ 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

Indirect observation of large-scale galactic magnetic fields sets a stringent bound on magnetic monopole density.

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 such fields are observed (hence not screened) then implies a limit on the abundance of free magnetic charges. For a review of magnetic monopoles and flux limits, see, e.g, J Preskill, *Annu. Rev. Nucl. Part. Sci.* **34**:46(1984).

15.9'b Elimination of pseudovectors

The main text pointed out the conceptual benefits of formulating magnetic effects in terms of the tensor $\vec{\omega}$, not the traditional \vec{B} . For example, to measure $\vec{\omega}$ 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 $\vec{E} + c\vec{B}$, 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 \vec{B} . Third, and most important, everybody does agree

¹⁴See Problem 17.2.

¹⁵There may be collective excitations in condensed matter with this character.

¹⁶Admittedly, some exotic articles do introduce $\vec{E} + ic\vec{B}$. But the inversion invariance of the resulting formulas is hidden.

that \vec{B} has got to be scrapped when we unify 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 secretly invariant. So we should expect that there is a way to make the invariance explicit in each object separately.

15.9'c 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 such a tensor form a similar object of rank $p + 1$, 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 $\omega = dA$ for some $(p - 1)$ -form A . There is an important caveat "locally" means that this result is valid only on a contractible region of space. (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, and different choices of path are not 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 ψ 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 $\omega = dA$. There's an ambiguity: Because $d^2 = 0$, we may add any gradient $d\Xi$ to the vector potential A without altering dA . That's gauge invariance.

Your Turn 15K

Find equally elegant forms of Ampère's law and the Stokes theorem.

Your Turn 15L

Show that changing the base point used in Equation 15.11 results in a gauge-transformed \vec{A} .

PROBLEMS

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 $0.5 \mu\text{V}/\text{m}$, 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 5 shows an electric motor consisting of a button magnet suspended from a 1.5 volt 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*

Figure 15.3a shows a demonstration experiment involving salt, pepper, and an overhead projector. It may seem remarkable that those tiny little ions could pull hard enough on the surrounding water to get it into bulk (macroscopic) motion. 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 $L = 5 \text{ cm}$. The plates have width $w = 5 \text{ cm}$ and are immersed in a solution with depth $h = 1 \text{ cm}$. The water between the plates contains sodium and chloride ions, each at number density (ions per volume) c_{ion} . Each ion carries electric charge $\pm e$, where e is the charge on a proton (the pepper is irrelevant). The solution consists of about one gram of NaCl dissolved in volume Lhw of water.

A total current of $I = 1 \text{ A}$ 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,

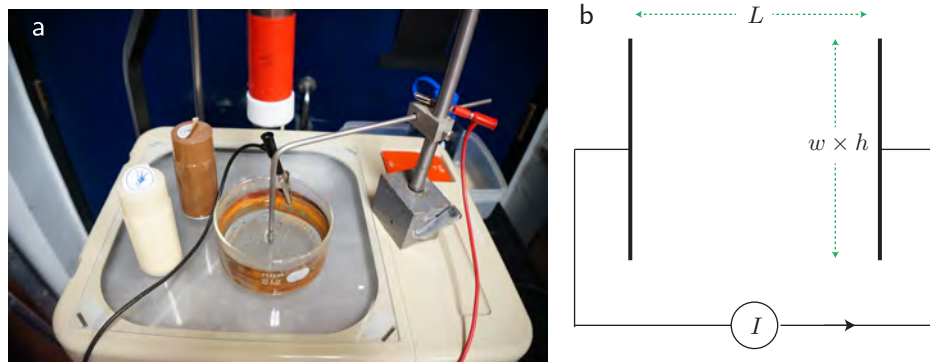


Figure 15.3: An experiment to demonstrate magnetoelectrophoresis. (a) Pepper is sprinkled on the surface of a salt-water solution to visualize bulk flow. A central electrode sends direct current radially outward to a ring-shaped electrode at the rim of the dish. A magnet pole can be brought toward the surface (*red*). See also Media 6. (b) Simplified geometry for Problem 15.3.

all $-$ charges originally in a layer with this 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.

- 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.)
- Now imagine applying a uniform magnetic field \vec{B} perpendicular to the plane of the picture, with strength $B = 0.03 \text{ T}$. 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.)
- 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 6?

15.4 Parity I

- The Maxwell/Lorentz equations, in the traditional form (Equations 0.1–0.4, page 2), are manifestly invariant under spatial rotations, because they involve constructions that are themselves invariant (dot product, cross product, curl, divergence). They 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 “right”; and
 - They involve cross products, which also depend on the same conventional choice.

Find a reformulation of these equations that involves no Levi-Civita tensors, by re-expressing the magnetic field \vec{B} in terms of $\vec{\omega}_{ij}$ (defined operationally by Equation 15.1, page 214) and simplifying with identities such as the ones in Section 14.5.2 (page 206). Hence, render the equations in manifestly parity-invariant form.

- [T2]** Similar criticisms can be raised for rigid body mechanics, which is also parity invariant, yet full of cross products:

$$\begin{aligned} \vec{v}_{(\ell)} &= \vec{\omega} \times \vec{r}_{(\ell)} && \text{velocity from angular velocity and position} \\ \vec{L} &= \vec{J} \cdot \vec{\omega} && \text{angular momentum from angular velocity} \\ \vec{\tau} &= \sum_{\ell} \vec{r}_{(\ell)} \times \vec{F}_{(\ell)} && \text{torque from force and position} \\ d\vec{L}/dt &= \vec{\tau} && \text{Newton's law of motion.} \end{aligned}$$

Construct second-rank tensors $\vec{\vec{\Omega}}$, $\vec{\vec{\Lambda}}$, and $\vec{\vec{T}}$ 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. [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 yet.]

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 $(0, 0, \pm w/2)$.

- Explain why the \vec{B} field 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 $\vec{0}$ between the coils.
- 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.
- 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 yz plane, itself lies in that plane. Hence the streamline that passes through any point in that plane remains confined to it.¹⁷ 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.
- Learn how to use a computer to create 3D streamplots, and show a representative sample throughout space for the same system. Look at various viewing angles till 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.]

¹⁷Section 0.3.1 (page 7).

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

You have surely been told that fundamental equations of physics are valid in any set of units.¹ Indeed, that is the case for classical mechanics. So you have a right to be puzzled when an author writes down Maxwell's 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 quotation marks remind us of this fact.

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. (16.1)

Once you understand that there are three distinct points, conversions become 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 TIME, LENGTH, AND MASS

16.2.1 Base units in mechanics

Just about every useful thing you've ever learned about units in mechanics can be systematized via a simple maxim:

*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*

¹Appendix 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.

representing an unknown quantity, just as we use the letter x for an unknown number.

Again: The units are part of the quantity. We carry these 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 $1 \text{ inch} \approx 2.54 \text{ cm}$. Dividing both sides of this formula by the numeric part 2.54, we find $0.39 \text{ inch} \approx 1 \text{ cm}$, and so on.

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 make a choice of **base unit**. Some options include:

1. Choose arbitrarily, for example, multiples of the king's foot for length. The SI does take this approach, but for time: The second is defined by declaring that an arbitrary (but convenient) physical phenomenon has a frequency with a specific, exact numerical value.³
2. Alternatively, choose a unit in such a way that some fundamental physical constant has a simple numerical part. For example, once we define the second we could agree to measure all lengths in light-seconds, defined as $(1 \text{ s})c$. The speed of light is exactly one light-second per second.
3. Or choose a unit in such a way that some physical constant has an exact numerical part. The SI does take this approach for length and mass: It currently defines the meter in terms of the second by requiring that the speed of light be exactly

$$c = 299\,792\,458 \text{ m/s.} \quad (16.2)$$

It similarly defines the kilogram in terms of the meter and second by requiring that the (unreduced) Planck constant h be exactly $6.626\,070\,15 \cdot 10^{-34} \text{ kg m}^2 \text{ s}^{-1}$.

Other quantities, like force, can then be expressed as products of base units raised to various powers.

16.2.2 Elimination of units is an abbreviation

Again, a physical quantity like force, or a constant of Nature like Newton's gravitational constant, is the product of a pure number times some units. The numerical part changes if we change units. For example, $c \approx 3.0 \cdot 10^5 \text{ km/s} \approx 186\,000 \text{ mile/hour}$. We refer collectively to **m**, **cm**, **inches** . . . as different units for the same "dimension," which we denote generically by \mathbb{L} . Similarly time and mass have generic dimensions called \mathbb{T} and \mathbb{M} respectively. We'll adhere to an approach that we may call:

A. Carry all units: Any valid formula, like $f = ma$, involves an equality between quantities with the same dimensions, *and is valid in any set of units*. Definitions such

³The second is defined as being equal to the time duration of exactly 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the fundamental unperturbed ground-state of the cesium-133 atom. That strange value was chosen to get a human-sized unit that is nearly equal to an older definition of the second.

as $1 \text{ m} = 100 \text{ cm}$ are themselves valid formulas because both sides have the same dimensions.

But some people get tired of writing units all the time. Here are two other options for how to proceed:

B1. Eliminate all units: Alternatively, we could choose once and for all a set of base units, and agree to express everything in terms of them. For example, if we choose SI base units, then in place of acceleration a , we define $\bar{a} = a/(1 \text{ m/s}^2)$, which is the numerical part of a . Similarly $\bar{c} \approx 3 \cdot 10^8$. Now our formulas, expressed in terms of the barred quantities, contain no units at all. But those formulas are only valid if we consistently use the stated system.

Suppose that we were asked to find a force. After we do our calculations, we wind up with a numerical value for \bar{f} . Knowing the meaning of force, we interpret this number as the actual force in newtons, because the combination of SI base units with the dimensions of force (MLT^{-2}) is kg m/s^2 , which is called newton. We get the same final answer as in approach **A**—if we didn't make any errors along the way.

The virtue of approach **B** is that formulas are compact. Moreover, if we follow option **2** above and choose our base unit of length to be $(1 \text{ s})c$, not meters, then we find $\bar{c} = 1$, so *we can drop all the factors of \bar{c} from our formulas*, abbreviating them still further. The disadvantage of approach **B** is that we forfeit the real benefit of dimensional analysis: Dimensional analysis expresses certain homogeneity (rescaling) properties of Nature, which appear as redundancies we can use to spot our errors. Eliminating units removes this helpful mechanism for checking our work.

We can compromise and take this process only partway:

B2. Eliminate some units: We again agree to measure time in some unit xx and length in $(\text{xx})c$, but we don't commit to any particular unit for mass. For each physical quantity X , we define \bar{X} to be X divided by as many powers of c as are needed to eliminate the length dimensions only. Thus, all barred quantities have dimensions that are powers of T and M . We again have the virtue that $\bar{c} = 1$, so we needn't write it. But we have also retained some of our error-checking abilities.

For example, we have force $\bar{f} = f/c$, mass $\bar{m} = m$, acceleration $\bar{a} = a/c$, and energy $\bar{\mathcal{E}} = \mathcal{E}/c^2$. Then some famous formulas become

$$\bar{f} = \bar{m}\bar{a}; \quad \bar{\mathcal{E}} = \bar{m}.$$

*This book uses approach **A** exclusively.* Students often tacitly use **B**, for example, when working exams, and so miss errors they would have caught by carrying the units explicitly. Generally **A** involves more writing, but it's worth it.

Section 16.3.4 below outlines how “gaussian” authors use a variant of approach **B2**.

16.3 UNITS IN ELECTRODYNAMICS

Charge is incommensurable with time, length, and mass, so we have to make an arbitrary choice of base unit, and also assign a new dimension symbol \mathbb{Q} for it. There's no human intuition for charge, so we don't feel constrained to use “human sized” units.

Here are the Maxwell equations as stated in the Prologue:

$$\begin{aligned}\vec{\nabla} \cdot \vec{E} &= \rho_q / \epsilon_0 && \text{Gauss} \\ \vec{\nabla} \cdot \vec{B} &= 0 && \text{Gauss} \\ \vec{\nabla} \times \vec{E} + \partial \vec{B} / \partial t &= \vec{0} && \text{Faraday} \\ \vec{\nabla} \times \vec{B} - \mu_0 \epsilon_0 \partial \vec{E} / \partial t &= \mu_0 \vec{j} && \text{Ampère.}\end{aligned}\quad [0.1-0.4, \text{ page 2}]$$

and the Lorentz force law:

$$d\vec{p}/dt = q(\vec{E} + \vec{v} \times \vec{B}). \quad [0.5, \text{ page 3}]$$

Two constants of Nature, ϵ_0 (the **electric permittivity of vacuum**) and μ_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:

$$\vec{E} \sim \frac{\text{ML}}{\text{T}^2 \text{Q}} \quad \text{and} \quad \vec{B} \sim \vec{E}/c \sim \frac{\text{M}}{\text{TQ}}.$$

Here “ \sim ” means “has the same dimensions as.”⁴

The Gauss and Ampère laws then give the units of ϵ_0 and μ_0 :

$$\epsilon_0 \sim \frac{\text{Q}}{\text{L}^3} \frac{\text{T}^2 \text{Q}}{\text{M}} \sim \frac{\text{Q}^2 \text{T}^2}{\text{L}^3 \text{M}} \quad \mu_0 \sim \frac{\text{M}}{\text{TQL}} \frac{\text{L}^2 \text{T}}{\text{Q}} \sim \frac{\text{ML}}{\text{Q}^2}.$$

Because these physical constants involve charge dimensions, their numerical parts will depend on what we choose as our unit of charge. We can use this freedom to arrange that the numerical part of *either* ϵ_0 or the proton charge e has an exactly specified numerical part (option **3** in Section 16.2.2).⁵ Once we do that, then there’s no more freedom; the *other one* has numerical parts set by Nature, which we can only measure and quote to a certain number of significant figures.

16.3.1 The SI base unit of charge is the coulomb

Following approach **3** above, the SI declares that the proton charge e is

$$e = 1.602\,176\,634 \cdot 10^{-19} \text{ coul.} \quad \text{exact}$$

That strange, but exact, multiple was chosen to make this definition nearly equivalent to an older one.⁶ The strange exponent has the convenient consequence⁷ that typical atomic and molecular-bond energies are around $1\text{eV} = (e)(1\text{J coul}^{-1})$.

That choice exhausts all our freedom to set units, so the numerical values of ϵ_0 and μ_0 need to be measured in the lab; they cannot have declared exact values. Thus,

⁴The symbol “ \approx ” means “is approximately equal to.”

⁵ μ_0 follows whatever status we gave to ϵ_0 , via Equation 16.3 and Equation 16.2.

⁶See Further Reading.

⁷Also, 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 volt mains). Also, the total charge delivered in a lightning strike is of order 1 coul.

Table 16.1: Derived units.

Quantity	Symbol	Unit Name	Abbrev.	Alternative	Alternative
charge	q	coulomb	coul	C	
current	I	ampere	A	coul/s	
magnetic induction	\vec{B}	tesla	T	kg/(coul s)	volt s/m ²
electric field	\vec{E}	—	—	kg m/(coul s ²)	volt/m
electric potential	ψ	volt	volt	V	J/coul
charge density	ρ_q	—	—	coul/m ³	
charge flux	\vec{j}	—	—	A/m ²	coul/m ² s
inductance	L	henry	H	kg m ² /coul ²	J/A ²
capacitance	C	farad	F	s ² coul ² /(kg m ²)	coul ² /J or coul/volt
electric dipole moment	\mathcal{D}_E	debye	debye	D	10 ⁻¹⁸ statcoul cm
				$\approx 10^{-21}$ coul m ² s ⁻¹ /c	$\approx (0.021 \text{ nm})e$
resistance	R	ohm	Ω	volt/A	J s/coul ²
conductance	G	siemens	Ω^{-1}	S	mho = \mathcal{U}

μ_0 and ϵ_0 are on equivalent logical footing. Their values are not independent, however; Chapter 18 will show that they are related by

$$c \equiv (\mu_0 \epsilon_0)^{-1/2} \sim \mathbb{L}\mathbb{T}^{-1}. \quad (16.3)$$

Thus, measuring one gives the other one, because c is exact. Recent values are

$$\mu_0 \approx (4\pi)(1.000\,000\,000\,82 \cdot 10^{-7}) \text{ m kg coul}^{-2} \quad (16.4)$$

and hence, via Equations 16.3 and 16.2,

$$\epsilon_0 \approx 8.854187817 \cdot 10^{-12} \text{ coul}^2 \text{ N}^{-1} \text{ m}^{-2}.$$

T2 Section 16.3.1' (page 240) asks, "Why the proton?"

16.3.2 Derived SI units

Starting from the base units coul, m, kg, and s, various useful combinations have been given names (Table 16.1).⁸⁹

Hence ϵ_0 can also be written as $\approx 8.85 \cdot 10^{-12}$ F/m, and $\mu_0 \approx 4\pi \cdot 10^{-7}$ H/m = $4\pi \cdot 10^{-7}$ N/A².

16.3.3 The gaussian base unit of charge is the statcoulomb

The gaussian system uses base units cm, g and s. There are several cgs-based systems; the most common one is often called "gaussian units."¹⁰ This time, we follow approach

⁸SI style guides say to use V and C; this book instead uses volt and coul rather than risking the confusion of a one-letter abbreviation. On the chalkboard, V could look like a volume; C could look like capacitance, or the speed of light.

⁹Some books also introduce the unit weber, abbreviated Wb, for magnetic induction times area, via $\text{Wb} = \text{kg m}^2/(\text{coul s}) = \text{volt s}$.

¹⁰Maxwell and F. Jenkin had more to do with developing this system than Gauss (Maxwell & Jenkin, 1865).

2 above, and set the base unit of charge (the “statcoulomb”) by requiring that ϵ_0 (not e) have an exact numerical part:

$$\epsilon_0 = \frac{1}{4\pi} \frac{\text{statcoul}^2 \text{s}^2}{\text{g cm}^3} \quad \text{exact.} \quad (16.5)$$

In this system, it’s e that has an approximate, measured value. We determine μ_0 by using Equation 16.3:

$$\mu_0 = \frac{4\pi}{c^2} \frac{\text{g cm}^3}{\text{statcoul}^2 \text{s}^2}.$$

Combining Equations 16.4, 16.3, and 16.5 yields

$$1 \text{ statcoul} \approx (0.1 \text{ m/s})c^{-1} \text{ coul} \approx \frac{1}{3 \cdot 10^9} \text{ coul.} \quad (16.6)$$

We then can express charge density in $\text{statcoul}/\text{cm}^3$ and charge flux in $\text{statcoul}/(\text{cm}^2 \text{ s})$, and so on.

Another useful unit conversion involves electrostatic potential. The SI unit is volt = J/coul. The corresponding gaussian unit is $\text{statvolt} = \text{erg}/\text{statcoul}$.

Your Turn 16A

Find the relation between these units.

16.3.4 The gaussian system involves two additional conventions

We can deal more briskly with points (*ii–iii*) in Idea 16.1.

Modified \check{B} field

“Gaussian” authors also redefine the magnetic induction, introducing a physically different quantity that we will call

$$\check{B} \equiv c\vec{B}.$$

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 \check{B} with no particular identifying phrase.

We can use \check{B} in any system of units, and indeed, we’ll occasionally find it convenient even in SI units, because it has the same units as \vec{E} .¹¹ Similarly, we will sometimes use a modified magnetic dipole moment¹² defined by $\check{D}_M = \vec{D}_M/c$.

Maxwell’s equations can then be written without explicitly mentioning μ_0 :

$$\vec{\nabla} \cdot \vec{E} = \rho_q/\epsilon_0 \quad (16.7)$$

$$\vec{\nabla} \cdot \check{B} = 0 \quad (16.8)$$

$$\vec{\nabla} \times \vec{E} + \frac{1}{c} \frac{\partial \check{B}}{\partial t} = 0 \quad (16.9)$$

$$\vec{\nabla} \times \check{B} - \frac{1}{c} \frac{\partial \vec{E}}{\partial t} = \frac{1}{\epsilon_0} \vec{j}, \quad (16.10)$$

¹¹To 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 $\check{B}_{(B)}$, and no electric field, if $\|\vec{E}_{(A)}\| = \|\check{B}_{(B)}\|$.

¹²Chapter 17 will introduce magnetic dipole moment. Section 49.6.2 will define a similarly modified *density* of moment.

and the Lorentz force law says

$$\frac{\partial \vec{p}}{\partial t} = q \left(\vec{E} + \frac{\vec{v}}{c} \times \vec{B} \right). \quad (16.11)$$

These equations are still valid in any system of units; in gaussian base units, we have the numerical values $\epsilon_0 = \frac{1}{4\pi} \frac{\text{statcoul}^2 \cdot \text{s}^2}{\text{g} \cdot \text{cm}^3}$ and $c \approx 3 \cdot 10^{10} \text{ cm s}^{-1}$.

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 16.6 to show that the SI equivalents of these units are

$$1 \text{ gauss} \approx c \cdot 10^{-4} \text{ T}$$

$$1 \text{ statvolt/cm} \approx 3 \cdot 10^4 \text{ volt/m.}$$

More precisely, a field $\vec{B} = 1 \text{ T}$ corresponds to $\vec{B} = 10^4 \text{ gauss}$.

Elimination of charge units

We could stop there. But “gaussian” authors take one more step. So far we have stayed with what Section 16.2.2 called approach **A**, but now we switch to:

B2: Eliminate charge units: For each physical quantity X , “gaussian” authors define \bar{X} to be X divided by as many powers of $(\text{statcoul})(\text{s})(\text{g cm}^3)^{-1/2}$ as are needed to eliminate the \mathbb{Q} dimensions.¹³ Why this crazy choice? With this choice, $\bar{\epsilon}_0$ becomes *dimensionless*, with exact numerical value equal to $1/(4\pi)$; it has also been purged of all units.

Thus, all barred quantities have dimensions that are powers of L, M, and T only: We have eliminated charge units. The vacuum Maxwell equations now take the elegant form

$$\begin{aligned} \vec{\nabla} \cdot \vec{\bar{E}} &= 4\pi \rho_q \\ \vec{\nabla} \cdot \vec{\bar{B}} &= 0 \\ \vec{\nabla} \times \vec{\bar{E}} + \frac{1}{c} \frac{\partial \vec{\bar{B}}}{\partial t} &= 0 && \text{ (“gaussian” units)} \\ \vec{\nabla} \times \vec{\bar{B}} - \frac{1}{c} \frac{\partial \vec{\bar{E}}}{\partial t} &= \frac{4\pi}{c} \vec{j} \\ \frac{d\vec{p}}{dt} &= \bar{q} \left(\vec{\bar{E}} + \frac{\vec{v}}{c} \times \vec{\bar{B}} \right). \end{aligned}$$

Then we get Coulomb’s Law in the ultra-simple form $\bar{\psi}(\vec{r}) = \bar{q}/\|\vec{r}\|$, 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).

¹³This step does not change the numerical part of X if we’ve expressed it in the base units **cm**, **g**, **s**, and **statcoul**. That’s because this factor’s numerical part equals one in that case.

“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 $C = 1 \text{ F}$ corresponds to the reduced quantity $\bar{C} = 9 \cdot 10^{11} \text{ cm}$.”

16.3.5 What is an “esu”?

We have seen that “gaussian units” eliminate the dimension \mathbb{Q} , though they still retain the familiar \mathbb{L} , \mathbb{T} , and \mathbb{M} .

Incredibly, however, it is commonplace for authors not to state any specific units. Instead they often just write something like “esu” for everything, which roughly means “whichever of those units is appropriate for this quantity in the system I’m using.” You’re supposed to 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.

16.4 REMARKS

It is humbling to note that electrodynamics was only a small part of Maxwell’s short professional life (think kinetic theory of gases; math theory of color vision; math theory 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 they have the same physical content in any language.

The “gaussian” unit system eliminates one of the two independent constants of Nature in Maxwell’s equations: Instead of ϵ_0 and μ_0 , all we now have is c . Some people find this beautiful. If you instead think that making fewer errors in your work is beautiful, then don’t eliminate units.

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 Maxwell’s equations in terms of \check{B} instead of \vec{B} (Equations 16.7–16.10), regardless of whether we measure \check{B} in **gauss** or in $\text{T} \cdot c$. Ultimately we’ll construct a single, unified “Faraday tensor” out of the components of \vec{E} and \check{B} .

Finally, don’t try looking on Amazon for a “statvoltmeter” or an “statammeter.”¹⁴ 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.

¹⁴Especially don’t ask for an “abammeter.”

Intermediate:

Historical: Maxwell & Jenkin, 1865.

Technical:

Current definitions of SI units: www.bipm.org/en/publications/si-brochure. (Prior to 2019, the SI defined the coulomb by giving μ_0 , not e , an exact conventional value: The constant in Equation 16.4 was exactly 1).

T_2

16.3.1' 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.

 T_2

16.3.3'a Planck units

The SI sets one base unit arbitrarily (the second), then fixes the others by requiring that fundamental constants have exact values. In principle, one need not stop there, because there is one more fundamental constant: Newton's gravitational constant. Requiring that G_N have an exact value would finish constraining all base units. (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 , \hbar , 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 \hbar was understood).

16.3.3'b Elimination of more units

Regardless of the SI's decision, in gravitational physics, many authors take elimination of units one step further, agreeing to measure time in units of $(1 \text{ m})/c$ (not s) and mass in $(1 \text{ m})c^2/G_N$ (not kg). Barred quantities are obtained by dividing physical quantities by enough powers of c and G_N to eliminate *both* \mathbb{M} and \mathbb{T} , leaving only \mathbb{L} . In this scheme, $\bar{G}_N = 1$ and $\bar{c} = 1$, so both can be dropped from formulas.

In high-energy physics, many authors choose instead to eliminate both \mathbb{L} and \mathbb{T} , leaving only \mathbb{M} , which they typically measure in GeV/c^2 . They set up barred quantities by demanding that $\bar{c} = 1$ and $\bar{\hbar} = 1$, leading to confusion when they talk to gravitational physicists.

PROBLEMS

16.1 Dimensional shortcut

Background: A magnet is dropped through a nonmagnetic, but conducting, tube. Friction with the tube's walls is negligible. But instead of increasing without bound, the magnet's velocity saturates¹⁵ at a surprisingly small value v_* . At this terminal velocity, the release of gravitational potential energy does not go into increasing the magnet's kinetic energy; instead, it all goes into ohmic heating of the tube, via induced “eddy” currents.

Eddy currents slow the fall of a magnet through a conducting tube.

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 $\vec{\mathcal{D}}_M$. Actually the dipole moment always enters into formulas multiplied by μ_0 , so let $X = \mu_0 \|\vec{\mathcal{D}}_M\|$. In the limit $X \rightarrow 0$, there's no effect and (in vacuum) the falling magnet's velocity increases without limit, that is, $v_* \rightarrow \infty$.
- (ii) The effect depends on the electrical conductivity κ of the material constituting the tube. In the limit $\kappa \rightarrow 0$, eddy currents are suppressed, so again $v_* \rightarrow \infty$.
- (iii) The terminal velocity depends 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_* 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 tube.¹⁶
- (v) Here are some typical values: A stack of button magnets like the one used in Media 7 has magnetic moment $\approx 0.3 \text{ A m}^2$ and mass $\approx 7 \text{ g}$. The conductivity of aluminum is $\kappa \approx 5 \cdot 10^7 \text{ } \Omega^{-1} \text{ m}^{-1}$. A typical demo apparatus has diameter $L \approx 1 \text{ cm}$.

Now take these steps:

- a. Find a combination of the relevant constants X , κ , F , and L that has the dimensions of a velocity.
- b. Confirm that the formula you found in (a) has the expected behaviors listed in (ii–iv) above.
- c. Evaluate the formula for v_* numerically with the values given in (v) above.

16.2 Units: conductivity

- a. Infer the units of conductivity κ from the formula $\vec{j} = \kappa \vec{E}$. Infer the units of resistance from the formula $\Delta\psi = IR$.
- b. Use dimensional analysis to guess the relation between κ and R for a long wire of length L and cross section A .

¹⁵See Media 7.

¹⁶You can neglect possible dependences on other dimensions, for example, on the thickness of the tube's wall. (That is, pretend it's infinitely thick, a long straight hole bored into a big solid block of metal.)

- 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 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:

- “ $1 \Omega = ? \text{ s/cm}$.”
- “ $1 \text{ H} = ? \text{ s}^2/\text{cm}$.”
- “ $1 \text{ farad} = ? \text{ cm}$.”

Also fill in the missing numbers (that is, derive them).

CHAPTER 17

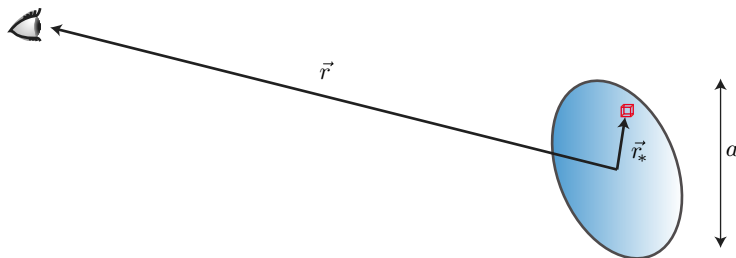
Magnetostatic Multipole Expansion

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 far away, as an expansion in powers of a/r . Again place the origin of coordinates at some fixed point inside the source. We'll again exploit Taylor's theorem for $\|\vec{r} - \vec{r}_*\|^{-1}$, 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: Another multipole expansion organizes the far fields by their falloff with distance.



17.2 TENSOR PRELIMINARIES

First, recall¹ that a stationary source must have $\vec{\nabla} \cdot \vec{j} = 0$. So

$$0 = \int d^3r (\vec{r}_i) (\vec{\nabla} \cdot \vec{j}) = - \int d^3r \vec{j}_k (\vec{\nabla}_k \vec{r}_i) = - \int d^3r \vec{j}_i. \quad (17.1)$$

(The boundary term is zero because we assumed a localized source.) We conclude that each cartesian component of \vec{j} , when integrated over the source, yields zero.

Similarly,

$$\begin{aligned} 0 &= \int d^3r (\vec{r}_k \vec{r}_i) \vec{\nabla} \cdot \vec{j} = - \int d^3r \vec{j}_m \vec{\nabla}_m (\vec{r}_k \vec{r}_i) \\ &= \int d^3r (\delta_{mk} \vec{r}_i \vec{j}_m + \delta_{mi} \vec{r}_k \vec{j}_m) = \int d^3r (\vec{r}_i \vec{j}_k + \vec{r}_k \vec{j}_i). \end{aligned} \quad (17.2)$$

¹Section 8.4 (page 114).

Define the **magnetic dipole moment tensor** as the first moment of \vec{j} :

$$\vec{\Gamma} = \int d^3r \vec{r} \otimes \vec{j}. \quad (17.3)$$

The charge flux is a vector field, but after the integral, $\vec{\Gamma}$ 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}_* 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 221), much as we did in electrostatics, thus gives

$$\vec{A}(\vec{r}) = \frac{\mu_0}{4\pi r} \int d^3r_* \vec{j}(\vec{r}_*) \left(1 + \frac{\vec{r} \cdot \vec{r}_*}{r^2} + \dots \right). \quad (17.4)$$

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 stationary current distribution never creates a “magnetic monopole” field.³

The definition Equation 17.3 lets us rephrase the second term:

$$\vec{A}(\vec{r}) = \frac{\mu_0}{4\pi r^3} \vec{r} \cdot \vec{\Gamma} + \dots \quad (17.5)$$

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 a potential into a product of universal, standard functions of \vec{r} (here the three functions $\mu_0 \vec{r} / (4\pi r^3)$) multiplied by some constants characterizing the source (here the components of $\vec{\Gamma}$).

Just three constants suffice to characterize the far fields of even a complicated stationary current distribution.

Although $\vec{\Gamma}$ 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.

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

²Section 8.7 (page 117) introduced this problem.

³A magnetic monopole field is, however mathematically imaginable; see Problem 17.2.

analogous to Equations 15.3 and 15.2 (page 214):

$$\vec{\Gamma}_{in} = \varepsilon_{ink} \vec{\mathcal{D}}_{M,k} \quad \text{where} \quad \vec{\mathcal{D}}_M = \frac{1}{2} \int d^3 r_* (\vec{r}_* \times \vec{j}(\vec{r}_*)). \quad (17.6)$$

The three numbers $\vec{\mathcal{D}}_M$ 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

$$\vec{A}^{\text{MD}}(\vec{r}) = \frac{\mu_0}{4\pi} \frac{\vec{\mathcal{D}}_M \times \hat{r}}{r^2}. \quad (17.7)$$

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

- 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 $\vec{\mathcal{D}}_M$ for this current distribution. [*Hint*: Use Equation 15.22 to find the charge flux and substitute into Equation 17.6.]
- Also, compute the curl of Equation 17.7 to find the \vec{B} field far away from a current source, to leading nontrivial order in a/r . Comment on the parallel between your answer and Your Turn 3B (page 40).

17.4 HIGHER MOMENTS

17.4.1 The magnetic quadrupole potential falls faster than the dipole

Naturally, 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 $\int d^3 r \vec{r}_i \vec{r}_k \vec{j}_m$. This tensor is clearly symmetric on its first two indices, so we might imagine that it would have $\frac{3(3+1)}{2} \times 3 = 18$ independent entries. But once again, current conservation imposes a condition that reduces this number. Moreover, not every possible combination of second moments actually contributes to the far field. Here are the details.

Begin by extending the argument in Section 17.2:

$$\begin{aligned}
 0 &= \int d^3r (\vec{r}_k \vec{r}_i \vec{r}_n) \vec{\nabla} \cdot \vec{j} = - \int d^3r \vec{j}_m \vec{\nabla}_m (\vec{r}_k \vec{r}_i \vec{r}_n) \\
 &= \int d^3r (\delta_{mk} \vec{r}_i \vec{r}_n \vec{j}_m + \delta_{mi} \vec{r}_k \vec{r}_n \vec{j}_m + \delta_{mn} \vec{r}_k \vec{r}_i \vec{j}_m) = \int d^3r (\vec{r}_i \vec{r}_n \vec{j}_k + \vec{r}_k \vec{r}_n \vec{j}_i + \vec{r}_i \vec{r}_k \vec{j}_n).
 \end{aligned} \tag{17.8}$$

That is, the totally symmetrized part of the second moment of the charge flux is zero.

Next, analogously to the dipole moment, define the **magnetic quadrupole moment tensor** tensor via

$$\vec{\vec{Q}}_M = \frac{2}{3} \int d^3r_* (\vec{r}_* \times \vec{j}(\vec{r}_*)) \vec{r}_*. \tag{17.9}$$

Your Turn 17C

Show that $\vec{\vec{Q}}_M$ is a traceless tensor.

Your Turn 17D

Extending the analogy to magnetic dipoles, use Equation 17.8 to show that

$$\int d^3r_* \vec{j}_i \vec{r}_{*,\ell} \vec{r}_{*,k} = \frac{1}{2} (\varepsilon_{in\ell} \vec{\vec{Q}}_{M,nk} + \varepsilon_{ink} \vec{\vec{Q}}_{M,n\ell}). \tag{17.10}$$

Now substitute into your result from Your Turn 17B to find

$$\vec{A}_i^{\text{MQ}}(\vec{r}) = \frac{\mu_0}{8\pi r^5} (3\vec{r}_k \vec{r}_\ell - r^2 \delta_{k\ell}) \varepsilon_{in\ell} \vec{\vec{Q}}_{M,nk}. \tag{17.11}$$

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; because $\vec{\vec{Q}}_M$ is also traceless, we see that only *five* independent numbers determine the magnetic quadrupole fields in magnetostatics, analogously to the five independent entries of the electric quadrupole moment tensor.⁴

Your Turn 17E

A circular loop of wire carries current I and sits in the xy plane centered on the origin. Cook up 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 41).]

⁴Changing the antisymmetric part of $\vec{\vec{Q}}_M$ may however produce a gauge transformation on the vector potential.

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 some 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.⁵

17.5 MAGNETIC DIPOLE IN AN EXTERNAL FIELD

We now find magnetic analogs of some results in Section 3.7 (page 43).

17.5.1 Force and torque on a dipole of fixed strength

Force

Consider a current distribution that can translate or rotate in space but is otherwise rigid: All current elements are steady in time and maintain fixed spatial relations with each other. A macroscopic example could be to imagine a stiff loop of wire with a constant-current source. Also, some individual molecules can create a permanent magnetic moment because of persistent currents in their electron state.

This current distribution under study is immersed in an external static magnetic field \vec{B}^{ext} , which varies with a characteristic length scale much bigger than the size of the distribution itself. We have $\vec{\nabla} \times \vec{B}^{\text{ext}} = 0$ inside the current distribution, because whatever the source of the external field, it doesn't overlap that distribution.

Choose an origin of coordinates somewhere inside the current distribution. Any internal forces must add up to zero. The Lorentz force law applied to each current element gives⁶

$$\vec{f} = \int d^3r_* \vec{j}(\vec{r}_*) \times \vec{B}^{\text{ext}}(\vec{r}_*).$$

Similarly to Section 3.7, we now make a Taylor expansion of the external field near the reference point: $\vec{B}^{\text{ext}}(\vec{r}_*) = \vec{B}^{\text{ext}}(\vec{0}) + \dots$. Then

$$\vec{f}_i = \varepsilon_{ink} \left[\vec{B}_k^{\text{ext}}(\vec{0}) \int d^3r_* \vec{j}_n(\vec{r}_*) + \frac{\partial \vec{B}_k^{\text{ext}}}{\partial \vec{r}_m} \Big|_{\vec{0}} \int d^3r_* r_{*m} \vec{j}_n(\vec{r}_*) \right] + \dots$$

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:

$$\vec{f}_i = \varepsilon_{ink} \frac{\partial \vec{B}_k^{\text{ext}}}{\partial \vec{r}_m} \Big|_{\vec{0}} \varepsilon_{mnj} \vec{D}_{M,j}$$

⁵See Problem 17.7.

⁶We'll revisit this argument in more detail later (Equation 35.5, page 482).

$$\begin{aligned}
 &= \frac{\partial \vec{B}_k^{\text{ext}}}{\partial \vec{r}_m} \Big|_0 (\delta_{im} \delta_{kj} - \delta_{ij} \delta_{mk}) \vec{\mathcal{D}}_{M,j} \\
 &= \vec{\mathcal{D}}_{M,j} \vec{\nabla}_i \vec{B}_j^{\text{ext}} - \vec{\mathcal{D}}_{M,i} \vec{\nabla} \cdot \vec{B}.
 \end{aligned}
 \tag{17.12}$$

A fixed magnetic dipole in a nonconstant magnetic field experiences an orientation-dependent force.

If the dipole moment is fixed (independent of the dipole’s position), then we can rewrite the last expression as

$$\vec{f} = \vec{\nabla} (\vec{B}^{\text{ext}} \cdot \vec{\mathcal{D}}_M).
 \tag{17.13}$$

Similarly to the electric dipole case, so too a rigid magnetic dipole feels no net force in a uniform magnetic field.

Torque

We can also work out the torque on this rigid current distribution:

$$\vec{\tau} = \int d^3 r_* \vec{r}_* \times (\vec{j}(\vec{r}_*) \times \vec{B}^{\text{ext}}).$$

For example,

$$\vec{\tau}_3 = \int d^3 r_* [\vec{j}_3(\vec{r}_*) (\vec{r}_* \cdot \vec{B}^{\text{ext}}) - \vec{B}_3^{\text{ext}} (\vec{r}_* \cdot \vec{j}(\vec{r}_*))].$$

First consider the terms without derivatives of the external field:

$$\vec{B}_i^{(0)} \int d^3 r_* \vec{j}_3(\vec{r}_*) \vec{r}_{*i} - \vec{B}_3^{(0)} \int d^3 r_* \vec{j}_i(\vec{r}_*) \vec{r}_{*i}.$$

The second term is zero because the magnetic dipole moment tensor is antisymmetric. The first term can be written in terms of the moment using Equation 17.6 as

$$= \vec{B}_i^{(0)} \frac{1}{2} \epsilon_{3ik} \int d^3 r_* (\vec{j} \times \vec{r}_*)_k$$

A fixed dipole in a nonconstant external field experiences an aligning torque.

$$\vec{\tau} = -(\vec{B}^{(0)} \times \vec{\mathcal{D}}_M).$$

We conclude that a free magnetic dipole of fixed strength in an external field experiences a torque tending to align its moment with the \vec{B} field. Once it is aligned, Equation 17.13 shows that it also feels a force driving it to a region of higher $\|\vec{B}\|$. You’ll explore a practical application of these results to manipulation of micrometer objects in Problem 17.3.

A beam of atoms with net magnetic moment will split as it travels through a region of nonuniform magnetic field.

Note that a quantum-mechanical spin cannot freely “reorient,” due to spatial quantization. Thus, a single neutron, which has a permanent magnetic dipole moment, (or a neutral atom such as silver) will migrate along *or against* the gradient of $\|\vec{B}\|$ depending on its spin state: The Stern–Gerlach effect (1922).⁷ Even particles currently thought to be fundamental, like the electron and muon, have permanent intrinsic magnetic dipole moments.

17.5.2 Diamagnetism, paramagnetism, ferromagnetism

Just as some 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, giving rise to a magnetic dipole moment. Bulk materials containing such molecules can develop a density of magnetic dipole moment throughout their volume.⁸ 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 there is a nonzero net magnetic dipole moment density even at zero applied field, we call the material **ferromagnetic**.

17.5.3 Purification of oxygen via diamagnetic forces

[Not ready yet.]

17.5.4 Magnetic levitation of macroscopic objects at room temperature

[Not ready yet.] See Media 8 = <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.

Raab & de Lange, 2005. Levitation of single cells: Durmus et al., 2015. Other continuous monitoring applications: Mirica et al., 2010.

Diamagnetic enrichment of oxygen from air: <https://patents.google.com/patent/US7771509B1/en>; Rybak et al., 2011; Cai et al., 2007; Madaeni et al., 2011; Nakano & Shiraishi, 2004; Hajduk et al., 2013.

Magnetic tweezers: Lionnet et al., 2012b; Lionnet et al., 2012a.

PROBLEMS

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

⁷Spin physics was born when Stern and Gerlach were astonished to find an even number of discrete spin states, not the odd number predicted from the theory of orbital angular momentum. [T2] Section 34.11' (page 471) will discuss how they fit into the tensor analysis methods of this book.

⁸See Problems 17.1 and 17.3. Chapter 49 will also develop this idea.

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 assume that this means 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.⁹

The cells are then placed in a magnetic field gradient, and the resulting force is used to manipulate the cell. What is the force if 100 of these particles are attached to a cell that is in a magnetic field of 1 T, with gradient 10 T/m?

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 , θ , φ be spherical polar coordinates.

- 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

$$\vec{A} = g\hat{\varphi} \frac{\cos\theta}{r \sin\theta}. \quad (17.14)$$

Here g is an overall constant and $\hat{\varphi}$ is the unit vector in the azimuthal direction. Compute the magnetic field corresponding to this vector potential as follows. First re-express \vec{A} as a scalar function times $\vec{\nabla}\varphi$ using your result in (a).

- Prove the identity $\vec{\nabla} \times (f \cdot \vec{V}) = (\vec{\nabla}f) \times \vec{V} + f \cdot \vec{\nabla} \times \vec{V}$ for any scalar function f and vector field \vec{V} .
- Use (a–c) to compute the curl of \vec{A} and interpret the result.
- Not surprisingly, the expression Equation 17.14 is singular at $r = 0$. But it’s *also* bad all along the polar axis! Show that the two modified expressions

$$\vec{A}(\pm) = g\hat{\varphi} \frac{\pm 1 + \cos\theta}{r \sin\theta} \quad (17.15)$$

differ from Equation 17.14 only by gauge transformations, and hence describe the same magnetic field.

- Show that one of the vector potentials Equation 17.15 is nonsingular all along the half-line $\theta = 0$, whereas the other one is nonsingular all along $\theta = \pi$. Thus, Equation 17.15 offers us vector potentials whose nonsingular domains jointly cover all of space except the origin (where there is a real singularity).

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.

⁹ [T2] 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 $\vec{D}_M = v\vec{B}/\mu_0$.

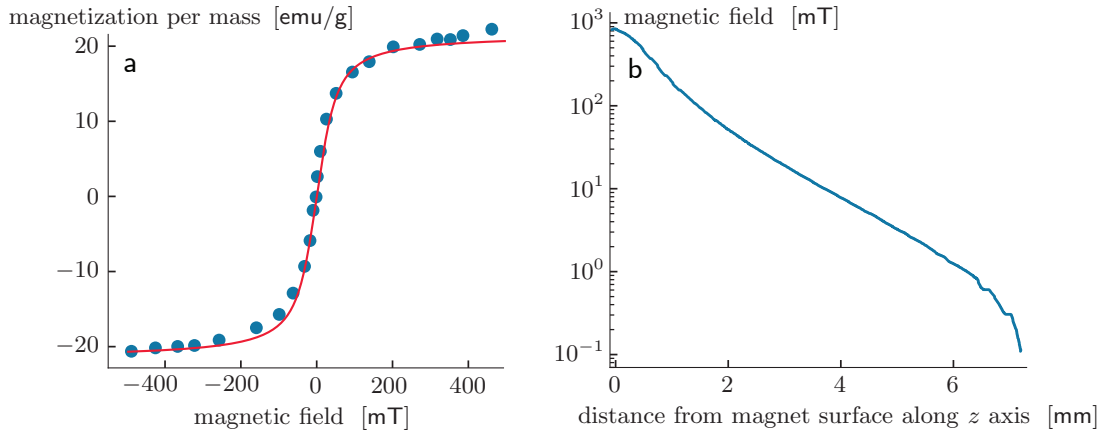


Figure 17.1: [Experimental data.] **See Problem 17.3.** (a) *Dots:* Mass density of induced magnetic dipole moment for beads of diameter $1\ \mu\text{m}$ and mass about $10^{-12}\ \text{g}$. The authors stated that “ $1\ \text{emu} = 10^{-3}\ \text{SI}$.” *Curve:* Fit to the function $M = a(\coth(B/B_0) - B_0/B)$ 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.]

- Apparently emu is some “gaussian” unit for magnetic dipole moment. Figure out the appropriate unit and explain the cryptic notation “ $1\ \text{emu} = 10^{-3}\ \text{SI}$.”
- 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.
- 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_{max} and z_0 .
- 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).¹⁰ Sketch the expected force-versus- z curve for $0 < z < 6\ \text{mm}$.
- For comparison, estimate the *weight* of this bead in air. (It will effectively be a bit reduced in water due to buoyancy.)

17.4 Levitation of single cells

[Not ready yet.]

17.5 Ambidextrous 2

Rederive the results of Section 17.5.1 without making use of the Levi-Civita tensor; that is, formulate them in terms of the magnetic field tensor $\vec{\omega}$ (Equation 15.3) and the magnetic dipole moment tensor $\vec{\Gamma}$ (Equation 17.2).

17.6 Magnetic quadrupole

Derive Equation 17.11. Then work out the curl to find the corresponding \vec{B} field to quadrupole order ($\mathcal{O}(R^{-4})$), and confirm the claim in the chapter that only the symmetric part of $\vec{\mathcal{Q}}_M$ enters your expression.

17.7 Basepoint dependence

¹⁰The dipole moment is *not* constant, so don’t use Equation 17.13.

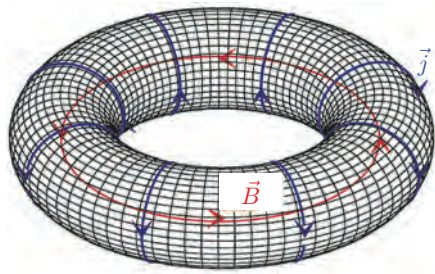


Figure 17.2: [From https://commons.wikimedia.org/wiki/File:Solenoid_currents_inducing_a_toroidal_magnetic_moment.tif]

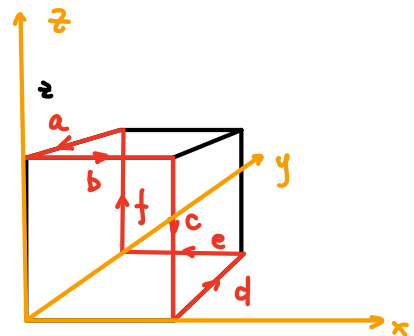


Figure 17.3: Edges a – f of this cube form a closed current loop.

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 \vec{a} . 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

- 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, $\vec{\ell}(u) = \vec{w} + u\hat{v}$, show that we just need to integrate $I\vec{w} \times \hat{v}du$.
- 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

- Suppose that one segment of a wire loop is straight, so that $\vec{\ell}(s) = \vec{r}_0 + s\hat{n}$ for some constant vector \vec{r}_0 and constant unit vector \hat{n} . Derive an expression for the

contribution to magnetic moment (Equation 17.6, page 245) 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.]

CHAPTER 18

Beyond Statics

“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

18.1 FRAMING: SELF-CONSISTENCY

The equations of static electricity and magnetism have a lot of practical implications. We’ve seen how to understand nerve impulses, photocopiers, lightning rods, molecular recognition, and much more with these equations. But charges and currents are not always static, nor even stationary. Finding the right equations required both experiments and the polestar of mathematical *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:

$$\vec{\nabla} \cdot \vec{E} = \rho_q / \epsilon_0 \quad \text{Gauss} \quad (18.1)$$

$$\vec{\nabla} \cdot \vec{B} = 0 \quad \text{Gauss} \quad (18.2)$$

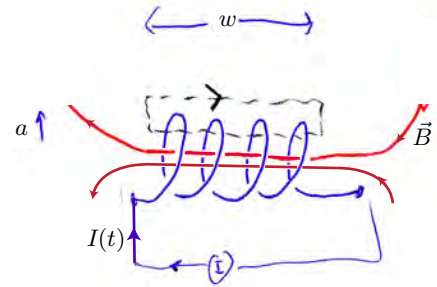
$$\vec{\nabla} \times \vec{E} = \vec{0} \quad (\text{stationary case}) \quad (18.3)$$

$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{j}. \quad \text{Ampère (stationary case)} \quad (18.4)$$

18.2.2 A coil carrying constant current

To anchor all the abstractions that are to come, here is an old result from magneto-statics 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 **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.

Figure 18.1: Solenoid. The dashed rectangle, with boundary orientation shown, can be decomposed into elements $d^2\vec{\Sigma}$ 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.



Each loop makes a magnetic dipole 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;¹ 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

$$\int d^2\vec{\Sigma} \cdot (\vec{\nabla} \times \vec{B}) = \int d^2\vec{\Sigma} \cdot (\mu_0 \vec{j}). \quad (18.5)$$

Stokes's theorem gives the left side as $\oint d\vec{\ell} \cdot \vec{B}$ 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 coil² 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 $d\vec{\ell} \cdot \vec{B} = 0$. And $\vec{B} \approx 0$ outside the cylinder, because the field lines fan out once they exit the ends.

On the right side of Equation 18.5, each time the wire pierces the surface Σ there is a contribution I to the integral. (That's because $\vec{\Sigma}$ and \vec{j} both point into the page at each such point.) Thus, Equation 18.5 becomes

$$Bw = \mu_0 NI, \quad \text{or} \quad B = \mu_0 NI/w, \quad (18.6)$$

a familiar result. 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.

Your Turn 18A

Repeat the argument, but traverse the dashed path in the opposite direction; make sure the physical result doesn't change.

¹Your Turn 0B (page 9).

²We are neglecting end effects.

As an aside, Equation 18.6 is sometimes expressed in terms of the quantity³ $\Phi_B = N\pi a^2 B$ as

$$\Phi_B = LI, \quad (18.7)$$

where we packaged all the constants into a single quantity to describe the coil geometry: the **self-inductance** L . In the situation we are considering, $L = \pi a^2 \mu_0 (N/w)^2 (w)$. 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 TIME-DEPENDENT CURRENTS

18.3.1 Faraday observed an \vec{E} 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, everyone assumed that it would continue to hold 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 $\vec{\nabla} \times \vec{E} = \vec{0}$. 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:

$$\vec{\nabla}_i \vec{E}_j - \vec{\nabla}_j \vec{E}_i = -2 \frac{\partial}{\partial t} \vec{\omega}_{ij}. \quad \text{Faraday} \quad (18.8)$$

The left side is an antisymmetric tensor field, which matches the object $\vec{\omega}$ that naturally describes magnetism.⁶ Equation 18.8 is clearly rotationally invariant (it equates tensors of the same type) and also invariant under spatial inversion (it contains no Levi-Civita tensor nor any “axial vector” quantities). Also the units match on each side.⁷

³Many 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 Φ_B any particular name.

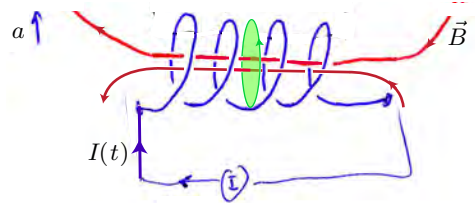
⁴Today we know that the newtonian equations also require modification; in Einstein’s theory those moving planets actually do generate tiny “gravitomagnetic” effects.

⁵See Chapter 36. In the USA, Joseph Henry independently discovered Faraday’s law, but did not publish it promptly.

⁶Equation 15.3 (page 214).

⁷Why is the factor of -2 needed? Chapters 32–34 will argue that the form of this equation, including this factor, is ultimately dictated by Lorentz invariance.

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 $d^2\vec{\Sigma}$ all pointing to the left, integrated over in Equation 18.10.



It's more conventional, however, to contract both sides of the preceding formula with a Levi-Civita tensor, which yields⁸

$$\vec{\nabla} \times \vec{E} = -\frac{\partial}{\partial t} \vec{B}. \quad \text{Faraday} \quad (18.9)$$

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

$$\int d^2\vec{\Sigma} \cdot (\vec{\nabla} \times \vec{E}) = -\frac{d}{dt} \int d^2\vec{\Sigma} \cdot \vec{B}. \quad (18.10)$$

This time, Stokes's theorem gives the left side as $\oint d\vec{\ell} \cdot \vec{E}$. By axial symmetry, the integrand is constant, so we get $2\pi a \vec{E}_\varphi$, where \vec{E}_φ is the component in the direction of current flow.

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

$$-\pi a^2 \frac{dB}{dt} = -\frac{\pi a^2}{w} \mu_0 N \frac{dI}{dt}.$$

⁸Recall Equation 15.2.

Setting this expression equal to the left side of Equation 18.10 gives

$$\vec{E}_\varphi = -\frac{\mu_0 N a}{2w} \frac{dI}{dt}. \quad (18.11)$$

The minus sign says that the induced electric field *opposes changes* in current.⁹ 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 $\rho_q^{(1D)}$. Imagine that charge as subdivided into packets Δq . Thus, there are $2\pi a N \rho_q^{(1D)} / \Delta q$ such packets in the wire. Each feels the same electric field, for a total force of

$$\vec{f}_\varphi = \frac{2\pi a N \rho_q^{(1D)}}{\Delta q} (\vec{E}_\varphi \Delta q).$$

To understand this formula, think of a pipe full of water, acted on by a body force like gravity. Pushing a volume δV 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* element of charge 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 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 $\Delta x = q / \rho_q^{(1D)}$. We can write the work *per unit charge* in terms of the self-inductance (Equation 18.7) as¹⁰

$$- \vec{f}_\varphi \Delta x = \frac{\mu_0 N^2 a^2 \pi}{w} \frac{dI}{dt} = L \frac{dI}{dt}. \quad (18.12)$$

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.

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

⁹Lenz's law again.

¹⁰Some books use the abbreviation "back-EMF" to describe this quantity, but we won't use that term. The F in that abbreviation stands for "force," but force is a vector with units \mathbf{N} ; in contrast, the quantity under discussion is a *scalar* with units volt.

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-induction* in Thomson's model. Incorporating self-induction creates the possibility of true traveling wave solutions, but those solutions again suffer from dispersion (Problems 18.4 and 18.5).

At high frequencies, self-inductance in a cable becomes important, but proper tuning of parameters can cancel out dispersion.

Astonishingly, Heaviside discovered that dispersion could be eliminated by *intentionally introducing leakage* between the conductors of a cable. Problems 18.5–18.6 have the details.¹¹

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:

$$\mathcal{E} = \mu_0 \frac{\pi a^2 N^2}{w} \frac{I^2}{2} = \frac{1}{2} LI^2. \quad (18.13)$$

It's important not to confuse the work \mathcal{E} with “frictional” loss (Joule heating due to ohmic resistance). Resistive losses occur even when current is held steady, and only in resistive 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 . 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 very suggestive result:

$$\mathcal{E} = \frac{\pi a^2 w}{2\mu_0} \|\vec{B}\|^2. \quad (18.14)$$

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 a constant times the field strength squared.

Equation 18.14 is reminiscent of a result we got long ago for capacitors:¹² We found that the energy needed to charge up a capacitor is $\frac{1}{2}\epsilon_0 \|\vec{E}\|^2$ times the *volume* of

¹¹Not so astonishingly, given his personality, Heaviside neglected to patent his very practical discovery, so others made a fortune from it.

¹²See Equation 6.2 (page 75).

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:¹³

The equations of electrodynamics appear to be compatible with energy conservation if we attribute energy density to empty space. (18.15)

In some special cases, we've found energy density $\frac{1}{2}\epsilon_0\vec{E}^2$ (capacitor, no \vec{B}) or \vec{B}^2/μ_0 (solenoid, no \vec{E}).

How are these energies stored? Apparently not in any medium—our coil and capacitor each have nothing inside! 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 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 ‘æther.’ 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.¹⁴ 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 $\vec{\nabla} \times \vec{E} = \vec{0}$ in order to accommodate experimental reality (Michael 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 relation for charge

Hanging Question #D (page 13) raised the issue that we must solve eight Maxwell equations with just six fields \vec{E} and \vec{B} . In statics, Section 15.5.5 (page 221) 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 Faraday's modified equation, Equation 18.9, and see that it's still vacuous (always satisfied, doesn't constrain the fields). But taking the divergence of Equation 18.4 and using the continuity equation now gives

$$0 = \mu_0 \vec{\nabla} \cdot \vec{j} = -\mu_0 \frac{\partial}{\partial t} \rho_q. \quad (18.16)$$

That's just *false* in nonstatic situations, so we have a problem. However, notice that:

¹³This statement parallels Idea 6.3 (page 75).

¹⁴Like many overturned ideas, this one had a long half-life. Lenard, Lorentz, and Michelson reportedly never gave up on it.

- The Gauss law says that the bad right-hand side equals $-\mu_0\epsilon_0\vec{\nabla}\cdot\frac{\partial\vec{E}}{\partial t}$, so
- Modifying Ampère's law could cure this inconsistency, replacing Equation 18.16 by an identity that's always true and rescuing our $8 \rightarrow 6$ argument. The required modification is just¹⁵

$$\vec{\nabla}\times\vec{B}=\mu_0\epsilon_0\frac{\partial\vec{E}}{\partial t}+\mu_0\vec{j}. \quad \text{Ampère} \quad (18.17)$$

We have arrived at Maxwell's famous modification of Ampère's law, 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.

[T2] Section 18.4.1' (page 274) 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:

$$\Delta\vec{B}_\perp=0. \quad \text{always} \quad [15.23, \text{page } 224]$$

$$\hat{n}\cdot(\vec{E}_{\text{vac}}-\vec{E}_{(1)})=-\hat{n}\cdot\vec{P}_{(1)}/\epsilon_0, \quad \text{dielectric/vacuum} \quad [6.19, \text{page } 87]$$

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,

$$\Delta\vec{E}_\parallel=\vec{0} \quad \text{and} \quad [6.21, \text{page } 87]$$

$$\Delta\vec{B}_\parallel=\mu_0\vec{j}^{(2D)}\times\hat{n}, \quad [15.24, \text{page } 224]$$

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 $\vec{E}_\parallel=\Delta\vec{E}_\parallel=\vec{0}$. Moreover, Faraday's law then says that $\partial\vec{B}/\partial t=\vec{0}$ inside. Supposing that the interior magnetic field is zero at some initial time then gives that it is always zero, so $\vec{B}_\perp=\Delta\vec{B}_\perp=0$.

¹⁵Note 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.

¹⁶See Section 18.4.1'b (page 274).

¹⁷And as they appear in the Prologue.

18.5 WAVE SOLUTIONS

18.5.1 About traveling plane waves

In one spatial dimension, we call a function of the form

$$\phi(t, r) = f(r - vx)$$

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 $x - vx = (x + v\Delta t) - v(t + \Delta t)$ for any x and t . Hence, the second snapshot is the *same* function of x as the first, just shifted in space by $\Delta x = v\Delta t$. In the figure, imagine slicing the surface along two lines of constant t : The result in each case is a bump function, just shifted.

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 $f(\hat{k} \cdot \vec{r} - vt)$ has the properties discussed above, where \hat{k} is any unit vector. Such a function is called a **plane wave**, because there is a stack of planes (each perpendicular to \hat{k}), on each of which it is constant. Suppose that we take snapshots at t and $t + \Delta t$. The second will differ by a shift of $\Delta \vec{r} = c(\Delta t)\hat{k}$.

We will often specialize to periodic functions, for example, taking $f(u) = \cos(2\pi\omega u/v)$. Then our function becomes

$$\phi(t, \vec{r}) = \cos(2\pi(\vec{k} \cdot \vec{r} - \omega t)). \quad (18.18)$$

Here ω is any constant and we defined $\vec{k} = \hat{k}\omega/v$. 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 $2\pi v/\omega$.

We call ω the **angular frequency** (dimensions \mathbb{T}^{-1} , SI unit rad/s) and k the **wavenumber** (dimensions \mathbb{L}^{-1} , 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 ω and k . But that risks confusion with the related quantities:

- **circular frequency** $\nu = \omega/(2\pi)$ (dimensions \mathbb{T}^{-1} , SI unit s^{-1} , also called Hz)
- **spectroscopic wavenumber** $k/(2\pi)$ (dimensions \mathbb{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 \mathbb{T} , SI unit s)
- **wavelength** $\lambda = 2\pi/k$ (dimensions \mathbb{L} , SI unit m).

The period is how long you have to wait at fixed position for the wavefront to repeat. The wavelength is how far you have to travel at a fixed instant of time for the wavefront to repeat.

If ϕ is a vector field, we can simply replace f by *three* functions; for a plane wave, each of those functions still depends on just one variable.

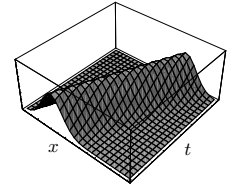


Fig. 11.2b (page 166)

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, let's simplify by looking at empty space, a region with no charges nor currents. Certainly 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

$$-\nabla^2 \vec{E} = -\mu_0 \epsilon_0 \frac{\partial^2}{\partial t^2} \vec{E}. \quad (18.19)$$

Maxwell noticed that this is an example of a *wave equation*. Consider the trial solution

$$\vec{E}(t, \vec{r}) = \vec{E} \cos(kz - \omega t), \quad (18.20)$$

where \vec{E} is any real, constant vector, k is a real constant, and ω is a real positive constant. The wave moves at speed ω/k .

Substituting the trial solution into Equations 18.19 and 18.1 (page 255) gives the conditions for the trial solution to work:

$$k^2 = \mu_0 \epsilon_0 \omega^2 \quad \text{and} \quad \hat{z} \cdot \vec{E} = 0. \quad (18.21)$$

Your Turn 18C

- Confirm that Equations 18.20–18.21 really do yield a solution to all of Maxwell's equations, not just the one combination Equation 18.19. You'll need to find the appropriate $\vec{B}(t, \vec{r})$ first.
- Try generalizing Equation 18.20 to arbitrary waveforms, that is, trial solution

$$\vec{E}(t, \vec{r}) = \vec{E} f(kz - \omega t), \quad \vec{B}(t, \vec{r}) = \vec{B} g(kz - \omega t).$$

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 every frequency travels at the same speed $(\epsilon_0 \mu_0)^{-1/2}$, independent of the amplitude $\|\vec{E}\|$ or frequency. So if we decompose any waveform into Fourier components, after time t they will assemble into the same waveform just shifted in space.

Free-space waves in vacuum follow a dispersion relation with constant velocity (speed does not depend on frequency or polarization).

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 μ_0 and ϵ_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.

¹⁸The derivation of this formula depends on our default choice of cartesian coordinates. The left-hand side would look more complicated in curvilinear coordinates.

18.6 POINTS REMAINING

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 ϵ_0 and μ_0 does lead to the observed value¹⁹ for c . But there are many other 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 Maxwell's 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.
- 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 transverse to the direction of propagation (Equation 18.21). Indeed, we noticed transverse polarization effects in the demo that generated microwaves via electric currents.²⁰ And visible light was already well known in Maxwell's time to display two independent polarizations, a detailed agreement with the electromagnetic theory of light. 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*.
- Notice from Your Turn 18C that \vec{E} and \vec{B} are perpendicular to each other, and each is perpendicular to the direction of motion \hat{z} . Also notice that each varies sinusoidally with time and space, and 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 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 points with zero field.

¹⁹Maxwell was not the first to observe this: See Section 18.6' (page 275).

²⁰See Media 1.

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

$$\Phi_{\vec{k},\omega}(t, \vec{r}) = e^{i(\vec{k}\cdot\vec{r}-\omega t)}. \quad (18.22)$$

Of course, \vec{E} and \vec{B} must still be real-valued vector fields, so in any formula involving $\Phi_{\vec{k},\omega}$ we will eventually need to take the real part to get the physical fields. But in intermediate steps, the complex notation is often nicer. That's because sine and cosine exchange roles under differentiation, whereas the derivative of exponential is always still exponential:

$$\frac{\partial \Phi_{\vec{k},\omega}}{\partial t} = -i\omega \Phi_{\vec{k},\omega}, \quad \vec{\nabla}_j \Phi_{\vec{k},\omega} = i\vec{k}_j \Phi_{\vec{k},\omega}.$$

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

$$\vec{E}(t, \vec{r}) = \frac{1}{2} \vec{E} \Phi_{\vec{k},\omega}(t, \vec{r}) + \text{c.c.} \quad (18.23)$$

The notation "c.c." denotes the complex conjugate of whatever precedes it, and guarantees that the overall expression is real.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 $\vec{E}(t, \vec{r})$.

The two extensions we are considering are that:

- The wavevector \vec{k} need not point along \hat{z} .
- The polarization vector \vec{E} need not be real. Write it as $\vec{E}^{(R)} + i\vec{E}^{(I)}$.

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 $\vec{\nabla} \Phi_{\vec{k},\omega} \rightarrow i\vec{k} \Phi_{\vec{k},\omega}$, so Equations 18.22 and 18.23 give

$$0 = \frac{1}{2} i\vec{k} \cdot \vec{E} \Phi_{\vec{k},\omega} + \text{c.c.} = \frac{1}{2} i\vec{k} \cdot [\vec{E}^{(R)} i \sin(\dots) + i\vec{E}^{(I)} \cos(\dots)] + (\text{two more terms}) + \text{c.c.} \quad (18.24)$$

The ellipses denote $\vec{k} \cdot \vec{r} - \omega t$. The third and fourth terms on the right 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 $\sin(\dots)$ and $\cos(\dots)$ *separately vanish*. So each of $\vec{k} \cdot \vec{E}^{(R)} = 0$ and $\vec{k} \cdot \vec{E}^{(I)} = 0$ must hold, or

$$\vec{k} \cdot \vec{E} = 0. \quad (18.25)$$

²¹Beware 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. We will always write such expressions in full.

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 $\frac{1}{2}\bar{b}\Phi_{\vec{k},\omega} + \text{c.c.} = 0$, where \bar{b} is some complex constant, to the conclusion²² that $\bar{b} = 0$.

For the special case where $\vec{E} = \hat{z}$, Equation 18.25 is the same transversality condition that we found earlier (Equation 18.21).

18.7.2 Faraday law

If \vec{E} is a plane wave, it seems a reasonable guess that \vec{B} will be too, so extend the trial solution:

$$\vec{B}(t, \vec{r}) = \frac{1}{2}\vec{B}\Phi_{\vec{k},\omega}(t, \vec{r}) + \text{c.c.},$$

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 the coefficients \vec{B} .

Again every $\vec{\nabla}$ becomes a factor of $\pm i\vec{k}$, and also $\partial/\partial t$ becomes $\mp i\omega$. Thus, Faraday becomes

$$\frac{1}{2}i\vec{k} \times \vec{E}\Phi_{\vec{k},\omega} + \text{c.c.} = -(-i\omega)\vec{B}\Phi_{\vec{k},\omega} + \text{c.c.}$$

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 times \vec{E} . These results generalize what you found in Your Turn 18C.

The electric and magnetic fields of a plane wave are perpendicular to the direction of travel and to each other.

For linearly polarized light, both fields vanish on a common stack of planes.

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

$$i\vec{k} \times \vec{B}\Phi_{\vec{k},\omega} + \text{c.c.} = c^{-2}(-i\omega)\vec{E}\Phi_{\vec{k},\omega} + \text{c.c.}$$

$$\vec{k} \times \left(\frac{\vec{k}}{\omega} \times \vec{E} \right) = -c^{-2}\omega\vec{E}.$$

Your Turn 18D

Simplify the triple cross product to show that $ck = \omega$ as before (Equation 18.21).

²²Nonlinear expressions will require more care; see Section 18.10.

18.7.5 Traveling wave with attenuation

In the preceding sections \vec{k} and ω 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 a wave could describe a signal traveling through a cable with a current leak that gradually saps its strength.

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** $ck = \omega$.

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 ψ 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} .²³ However, we no longer have $\vec{\nabla} \times \vec{E} = \vec{0}$, so electrons feel a nonconservative force,²⁴ unlike in statics. That is, there is no function whose gradient is minus the electric field. Nevertheless, Faraday's law²⁵ says that there *is* a vector quantity whose curl equals zero, namely $\vec{E} + \frac{\partial \vec{A}}{\partial t}$, so we *can* construct a function whose negative gradient equals that quantity. We'll continue to call it the "scalar potential" ψ , but keep in mind that ψ *can no longer be interpreted as potential energy* of a test body per unit charge. In short, we can always find potential functions such that²⁶

$$\vec{E} = -\vec{\nabla}\psi - \frac{\partial \vec{A}}{\partial t} \quad \text{and} \quad \vec{B} = \vec{\nabla} \times \vec{A}. \quad (18.26)$$

Equations 18.26 let us express six unknown fields (\vec{E} and \vec{B}) in terms of just *four* unknown potentials (\vec{A} and ψ), a significant simplification. We will soon see that further simplifications arise when we substitute this representation into Maxwell's 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.²⁷ Does it still hold good?

²³Section 15.3.5 (page 218).

²⁴Section 18.3.2 (page 258).

²⁵Equation 18.9 (page 258).

²⁶This result addresses Hanging Question #F (page 21).

²⁷Section 15.4 (page 219).

Your Turn 18E

Show that the substitutions

$$\vec{A} \rightarrow \vec{A} + \vec{\nabla}\Xi, \quad \psi \rightarrow \psi - \frac{\partial\Xi}{\partial t} \quad \text{gauge transformation} \quad (18.27)$$

doesn't change the electric or magnetic fields in Equation 18.26. Here Ξ is any scalar function of space and time.

Thus again, the potentials are not uniquely specified 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:

$$\vec{\nabla} \cdot \vec{A} = 0. \quad [15.14, \text{page 219}]$$

The proof that this is always possible locally is the same as it was in statics (Section 15.3.5), because we have not modified the gauge transformation formula for \vec{A} (the first of Equations 18.27 is the same as the formula in Section 15.4, page 219).

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 now an identity, so we can forget the magnetic Gauss law. Also, Faraday's law becomes an identity, so forget it too. We are left with *four equations in the four unknowns* \vec{A} and ψ :²⁸

$$\begin{aligned} \nabla^2\psi &= -\rho_q/\epsilon_0 && \text{(electric Gauss, Coulomb gauge), and} \\ \nabla^2\vec{A} &= -\mu_0\vec{j} + \mu_0\epsilon_0\left(\vec{\nabla}\frac{\partial\psi}{\partial t} + \frac{\partial^2\vec{A}}{\partial t^2}\right). && \text{(Ampère, Coulomb gauge)} \end{aligned}$$

It's tempting to say that we have just found another resolution of Hanging Question #D (page 13) (“eight equations in six unknowns”), 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 the second equation, and substitute the first, we find that this combination is vacuously satisfied; it does not constrain the potentials. So effectively, we do have four independent equations in four unknowns.

18.8.3 Coulomb gauge can be augmented if charge density is zero

We can simplify still more if we're studying a region with zero net charge density.²⁹ (There can still be *currents*, however, as in a neutral wire.)

Even if we restrict to Coulomb gauge, we *still* have some further freedom to apply certain gauge transformations, because transforming with any function Ξ that obeys $\nabla^2\Xi = 0$ will not spoil the Coulomb gauge condition. Let's try

$$\Xi(t, \vec{r}) = \int_{t_0}^t dt_* \psi(t_*, \vec{r}). \quad (18.28)$$

²⁸The derivation of these formulas depends on our default choice of cartesian coordinates. The left-hand sides would look more complicated in curvilinear coordinates.

²⁹Actually, Chapter 38 will achieve a similar simplification even with charges present, but we don't need that much power yet.

Your Turn 18F

Show that, if $\rho_q = 0$ everywhere, then:

- A gauge transformation by the function in Equation 18.28 preserves Coulomb gauge, and
- This gauge 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

- Show that the electric Gauss law is now automatically satisfied (an identity).
- Show that what remains is actually three *decoupled* equations in three unknowns:

$$\nabla^2 \vec{A} = -\mu_0 \vec{j} + \mu_0 \epsilon_0 \frac{\partial^2 \vec{A}}{\partial t^2} \quad \text{in Coulomb gauge extended by } \psi = 0. \quad (18.29)$$

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 18.29 is vacuously satisfied. Moreover, we get the simplification that in this gauge Equations 18.29 decouple, much as they did in magnetostatics.³⁰

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. For example, the plane wave solutions of Equation 18.29 moving along \hat{z} take the form

$$\vec{A}(t, \vec{r}) = \frac{1}{2} \vec{\zeta} \Phi_{\vec{k}, \omega}(t, \vec{r}) + c.c., \quad (18.30)$$

where $\vec{k} = k\hat{z}$, the **polarization vector** $\vec{\zeta}$ is a constant vector in the xy plane, and $\Phi_{\vec{k}, \omega}$ is one of the family of complex traveling waves in Equation 18.22.

Your Turn 18H

Show that more generally, Equation 18.30 gives plane wave solutions moving in *any* direction, as long as \vec{k} and ω obey the **dispersion relation**

$$\|\vec{k}\| = \omega/c \quad \text{where} \quad c = 1/\sqrt{\mu_0 \epsilon_0} \quad (18.31)$$

and $\vec{\zeta}$ is any vector perpendicular to \vec{k} .

³⁰See Your Turn 15E (page 221).

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 \vec{k} , Equation 18.30 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 \vec{k})! The resolution to this puzzle is that Equation 18.29 is only equivalent to the Maxwell equations in Coulomb gauge, and hence our trial solution only works if $\vec{\zeta} \perp \vec{k}$. Thus, *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 18.30, and hence the relation between the polarization vector $\vec{\zeta}$ and the vector \vec{E} appearing in Equation 18.20. Show that as before, \vec{E} , \vec{B} , and \vec{k} are mutually perpendicular.

18.10 COMPLEX POLARIZATIONS

18.10.1 Linear, circular, elliptical

If $\vec{\zeta}$ is a vector with *real* components, then \vec{E} oscillates about the $\pm\vec{\zeta}$ direction; we say the plane wave is **linearly polarized**, because the tip of its \vec{E} vector oscillates back and forth on a line in the plane perpendicular to \vec{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 wave.

Your Turn 18J

If you assumed that $\vec{\zeta}$ was real when you worked Your Turn 18I, work through it again without this assumption. Specifically, work out $\vec{E} \cdot \vec{k}$, $\vec{B} \cdot \vec{k}$, and $\vec{E} \cdot \vec{B}$.

Light can be created in helicity states.

Your Turn 18K

- Consider the wave with $\vec{k} = k\hat{z}$ and $\vec{\zeta} = \hat{x} + i\hat{y}$ (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**.
- Repeat with $\vec{\zeta} \propto \hat{x} + 2i\hat{y}$ and interpret such **elliptically polarized** solutions.

18.10.2 Helicity basis for circular polarization

Starting from a particular \vec{k} , choose a pair of real unit vectors $\hat{\zeta}_{(1)}, \hat{\zeta}_{(2)}$ perpendicular to \vec{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 \vec{k} can be written as a linear combination of these two basis vectors.

Alternatively, we can define complex basis vectors:

$$\hat{\zeta}_{(\pm)} = (\hat{\zeta}_{(1)} \pm i\hat{\zeta}_{(2)})/\sqrt{2}. \quad \text{helicity basis} \quad (18.32)$$

Any polarization vector $\vec{\zeta}$ can be written as a (possibly complex) linear combination of $\hat{\zeta}_{(1,2)}$, or of $\hat{\zeta}_{(\pm)}$. 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**).³¹

18.10.3 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 39).

18.11 PLUS ULTRA

Let’s pause to underscore the character of Maxwell’s advance: Ampère’s law for magnetostatics 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 31) already asked: *Where is the energy at 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

³¹Beware that different authors disagree about the convention for which is positive and which negative.

“luminiferous 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 permissible statement.

T2 Section 18.11’ (page 276) discusses some 20th century developments with a similar flavor.

FURTHER READING

Semipopular:

Zeeman effect: www.youtube.com/watch?v=0zkcB11kgGU

www.youtube.com/watch?v=JV4Fk3VNZqs .

Intermediate:

Maxwell’s thought process: Bork, 2005; Harman, 1998; Siegel, 1991; Hunt, 1991; Buchwald, 1985; Chalmers, 1975; Shapiro, 1973.

T₂**18.4.1'a Connection to ohmic materials**

Maxwell's 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 any case, they do not 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,³² 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 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 energy emerges 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.³³ 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 medium³⁴ might be accompanied by a similar current from the æther that

³²Although effectively reduced by buoyancy.

³³Heaviside uncovered this symmetry, but only much later.

³⁴See Section 49.2.1.

Maxwell and others believed filled empty space.³⁵ In fact, in his first publication introducing a displacement current³⁶ Maxwell does attribute it to distorted ‘æther cells.’

Nor was Maxwell’s original form for the displacement quite correct. He quietly changed it to the present form in a later work,³⁷ 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.³⁸

Nor was Maxwell explaining some existing, definitive experimental result: The constant of proportionality $\mu_0\epsilon_0 \approx 1.1 \cdot 10^{-17} \text{ m}^{-2} \text{ s}^2$ on the new term is extremely small, so no experiment envisioned in his day could directly confirm or refute it.³⁹ 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.⁴⁰

So how on Earth 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:⁴¹ *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. In the 1860s, the infinitely desirable (and widely shared) goal was to unify electromagnetism with optics. 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.⁴² But each of us can develop a smaller version of that sense by studying the thinking of others.

Naturally, lesser minds got hung up on the 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.

T₂

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 high, 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 apparatus wasn’t able to discriminate between infinite speed and the actual value. However, a few years later Ole Römer succeeded with a

³⁵How then could Einstein retain this term *even after denying the reality of the æther*? In the intervening decades, other scientists had developed the more abstract framework used today. Ironically, however, the name “displacement current” inspired by the analogy has stuck.

³⁶Part 3 of “On physical lines,” 1862.

³⁷“Dynamical theory,” 1865.

³⁸The “Treatise,” 1873.

³⁹Actually, 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.

⁴⁰Maxwell 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.”

⁴¹Chalmers, 1975.

⁴²In Maxwell’s case, that “somebody” was himself, but later. Many others contributed later still, notably H. Lorentz.

clever astronomical measurement. Terrestrial measurement had to await another clever idea from Fizeau (Chapter 29).

“Wilhelm Eduard Weber and Rudolf Kohlrausch demonstrated in 1856 that the ratio of electrostatic to electromagnetic units [today $(\epsilon_0\mu_0)^{-1/2}$] produced a number that matched the value of the then known speed of light” [en.wikipedia.org/wiki/Wilhelm_Eduard_Weber]. But “Weber and Kohlrausch measured a quantity larger by a factor of $\sqrt{2}$ than the ratio of the corresponding electrostatic and electromagnetic units. Therefore they obtained the value of the constant $\sqrt{2}c$, the relationship of which to the velocity of light was not immediately apparent. . . . [Instead Kirchhoff made the connection.] However, no special significance was ascribed to this. Weber, in particular, considered that due to the obvious difference in the nature of the phenomena of electrodynamics and of optics the equality of the two constants mentioned above is simply an accidental coincidence” [Shapiro, 1973]. Such a skeptical attitude to a numerical coincidence was permissible—maybe even required—in the absence of any real, 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, Weber and Kohlrausch’s measurements implied $c = 3.10740 \cdot 10^8$ m/s, fairly close to Fizeau’s measured 3.14850. “Maxwell was impressed, as Kirchhoff had been before him, by the close agreement between the electric ratio and the velocity of light, and he did not hesitate [in 1862] to assert the identity of the two phenomena [Whittaker, 1951, p254].” (“He had worked out the formulae in the country, before seeing Weber’s result” [Campbell & Garnett, 1882, p244].) Later (1868), Maxwell and C. Hockin made an improved measurement of $(\epsilon_0\mu_0)^{-1/2} \approx 2.88 \cdot 10^8$ m/s, and compared it to Foucault’s improved measurement of light speed 2.9836.⁴³

T₂

18.11’ On the guidance of mathematical consistency

Maxwell’s first great article⁴⁴ omitted the vacuum displacement charge flux term but explicitly pointed out that the resulting equations imply $\vec{\nabla} \cdot \vec{j} = 0$; he added “we know little of the magnetic effects of any currents which [have $\vec{\nabla} \cdot \vec{j} \neq 0$].” 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, Illiopoulos, Meyer).
- When superstring theorists tell us there must be six 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 intangible echo of great theories you have met, the click of links falling into place automatically.

⁴³Later still, (1892) Abraham obtained a still more precise measurement of $\epsilon_0\mu_0$.

⁴⁴“On Faraday’s lines of force,” 1855–6.

PROBLEMS

18.1 *Parity II*

Work Problem 15.4a (page 229) again, this time for the full Maxwell equations as developed in this chapter.

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 ω_0 , a certain number of radians per second.

At time zero, we turn on a magnetic field \vec{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 is⁴⁵ $1.7 \cdot 10^{-8} \Omega \text{ m}$, and its mass density is $9.0 \cdot 10^3 \text{ kg/m}^3$. 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 $\vec{B}(t) = \hat{z}(B_{(0)} + \delta\vec{B} \cos \omega t)$ is applied. Apply Faraday's law to a circular path in a plane perpendicular to \vec{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 κ . Use the ohmic relation (Equation 8.9, page 116) to find the average power dissipated in the conductor per volume. Actually, the RF signal is not continuous; it consists of pulses of duration Δt 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 ρ_m and κ , R , ω , $\delta\vec{B}$, Δt , and 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 $\delta\vec{B} = 2\pi/(2\gamma \Delta t)$ and $\omega = \gamma B_{(0)}$, where the “gyromagnetic ratio” γ of a proton is some constant and $B_{(0)}$ is the background magnetic field, a given number. Use this information to eliminate $\delta\vec{B}$ and ω from your formula for SAR.

⁴⁵www.matweb.com. The conductivity is the reciprocal of resistivity.

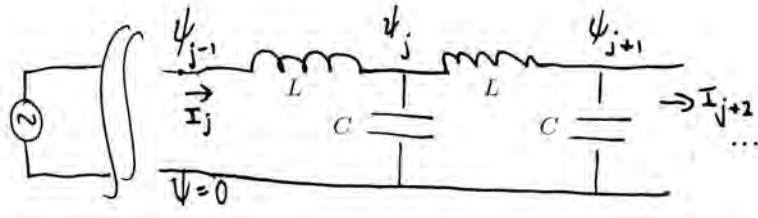


Figure 18.3: See Problem 18.4. Two modules in a long chain.

- e. Now substitute typical human values: $R \leq 0.17 \text{ m}$, $\kappa \approx 0.3 \Omega^{-1} \text{ m}^{-1}$. And use typical instrument values $B_{(0)} \approx 0.5 \text{ T}$ and $T_R \approx 1 \text{ s}$. Also, $\gamma \approx 2.7 \cdot 10^8 \text{ Hz/T}$.
- f. Safety requires that we not heat the patient too much! So demand that $\text{SAR} < 0.4 \text{ W/kg}$. Find the corresponding requirements on Δt and also on $\delta \vec{B}$.

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. Recently, 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 259) for some constant L is called an **inductor**.⁴⁶ 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. 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 \mathbb{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:

$$\psi_j(t) = \frac{1}{2} \bar{\psi}_j e^{-i\omega t} + \text{c.c.} \quad (18.33)$$

- c. Substitute that trial solution to get an infinite set of coupled *algebraic* equations.
- d. It still looks hard, but the equations are invariant under shifting everything one step in space. So our experience with related systems suggests that we make the trial solution

$$\bar{\psi}_j = \psi_0 e^{ijk}, \quad (18.34)$$

where k is some constant. Substitute this into your algebraic equations for a given angular frequency ω and see what k must be in order to get a solution.

- e. If ω lies in a certain range, there will be a real solution k to your condition. Then Equations 18.33–18.34 describe a wave traveling along the chain to infinity. Outside

⁴⁶In particular, an idealized inductor has negligible electrical resistance and capacitance.

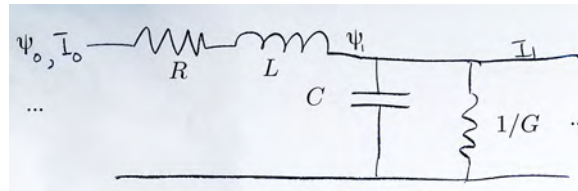


Figure 18.4: See Problem 18.5.

that frequency range, however, there will be no real solution; the transmission line has a **cutoff**. Find the allowed range of frequencies.

18.5 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 circuits, but actually we suppose that all four material properties R , L , C , and G are continuously distributed with densities r , ℓ , c , and g respectively.⁴⁷ Thus, you should initially consider a segment of length Δx , with $R = r\Delta x$ and so on; at an appropriate moment, take the limit $\Delta x \rightarrow 0$.

The line is infinitely long. We suppose that at some point an external agency imposes a harmonic potential $\psi(0; t) = \frac{1}{2}\bar{\psi}e^{-i\omega t} + \text{c.c.}$ We would like to find the solution everywhere. The problem is time-translation invariant, so again a reasonable trial solution is harmonic: $\psi(x; t) = \frac{1}{2}\bar{\psi}(x)e^{-i\omega t} + \text{c.c.}$

- Follow the strategy in Chapter 11 to write a second-order differential equation for $\bar{\psi}(x)$.
- The problem is also spatially translation invariant apart from the imposed boundary condition, so seek a solution of the form $\bar{\psi}(x) = e^{ikx}$ where k is a function of ω that you are to find.
- The dependence of the wavenumber k on the angular frequency is called the cable's **dispersion relation**. Why would it be desirable for k to take the general form $k = \pm(\omega/v_{\text{cable}} + i\lambda)$, where v_{cable} and λ are independent of ω ?
- The desirable condition does not generally hold, but Heaviside found that it does hold if the material parameters r , ℓ , c , and g obey a certain relation. Find that condition.
- Some resistance R 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?

18.6 Realistic transmission line II

This is a continuation of Problem 18.5. There you studied a class of problems described by four parameters r , ℓ , 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,

⁴⁷Note 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 115), and so the leak *conductance* G is proportional to Δx .

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 \mathbb{L} and then let \bar{x} be position divided by that scale. Then we let \bar{k} be k multiplied by that same scale, so that $\bar{k}\bar{x} = kx$. We also find another combination of the parameters with dimensions \mathbb{T} and then let \bar{t} be time divided by that scale. Then we let $\bar{\omega}$ be angular frequency multiplied by that scale, so that $\bar{\omega}\bar{t} = \omega t$.

- a. Give expressions for length and time scales with the property that the dispersion relation becomes

$$\bar{k} = \pm \sqrt{(\bar{\omega} + i)(\bar{\omega} + ig\ell/(rc))}. \quad (18.35)$$

We are interested in a problem where a signal generator fixes a definite potential $\psi_0(\bar{t})$ at the point $\bar{x} = 0$ in a semiinfinite wire. So we make the sign choice above that gives signals that decay as $\bar{x} \rightarrow \infty$ and use the input signal as a boundary condition at $\bar{x} = 0$.

In Problem 18.5 you found some solutions of the form $\frac{1}{2}e^{-i\bar{\omega}\bar{t} + i\bar{k}(\bar{\omega})\bar{x}} + \text{c.c.}$ 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 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

$$\psi_0(\bar{t}) = 2a + \sum_{m=1}^{n_{\max}} \frac{T}{\pi m} \sin(2\pi am/T) (e^{2\pi im\bar{t}/T} + e^{-2\pi im\bar{t}/T})$$

over the range $-0.1 < \bar{t} < 3.1$. Use illustrative parameter values $a = 0.05$, $T = 3$, $n_{\max} = 1000$. 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 18.35 as

$$\bar{k} = \pm(\bar{\omega} + i)\sqrt{1 - ib/(\bar{\omega} + i)}.$$

Here b is a dimensionless combination of r , ℓ , 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?

18.7 Helicity basis

The helicity basis is defined in Equation 18.32, starting from a choice of two vectors $\hat{\zeta}_{(1)}$, $\hat{\zeta}_{(2)}$ perpendicular to each other and to \bar{k} and forming a right-handed triad with it. This construction may sound too arbitrary.

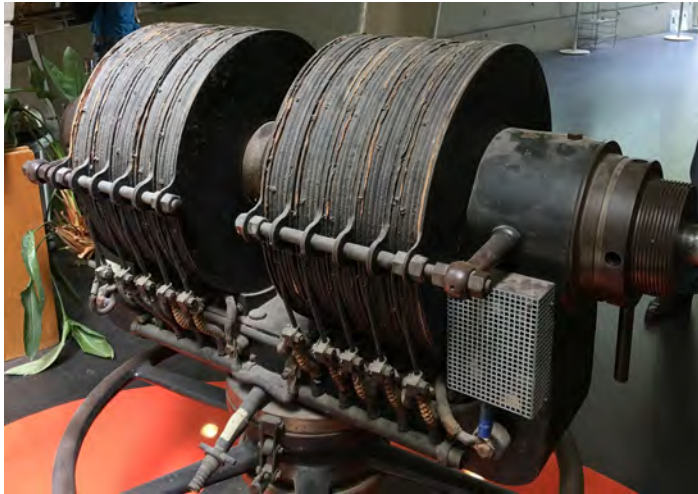


Figure 18.5: Magnet used in Zeeman's experiments.

- a. Show that if we choose a different pair of unit vectors $\hat{\zeta}'_{(1)}, \hat{\zeta}'_{(2)}$, which also make a right-handed, orthonormal triad with \vec{k} , then we get essentially the same helicity basis. That is, $\hat{\zeta}'_{(+)}$ is a scalar constant times $\hat{\zeta}_{(+)}$ and similarly for $\hat{\zeta}'_{(-)}$.

Now establish two properties that will be useful later:

- b. Also show that $\hat{\zeta}_{(\pm)} \cdot \hat{\zeta}_{(\pm)}^* = 1$ and $\hat{\zeta}_{(\pm)} \cdot \hat{\zeta}_{(\mp)}^* = 0$.
- c. Compute $\hat{k} \times \hat{\zeta}_{(\pm)}$ and express it in the helicity basis; show that the helicity basis vectors are eigenvectors of the operator " $\hat{k} \times$." (This operator generates infinitesimal rotation about \hat{k} .)

18.8 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,⁴⁸ 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 $+\hat{z}$ -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

An applied magnetic field splits a single, degenerate atomic spectral line into multiple, nearly degenerate lines.

⁴⁸You 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.

- kinds of polarizations will be emitted. So find the frequencies and corresponding polarizations of radiation seen by an observer located far away on the \hat{z} -axis.
- 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.
 - Evaluate your answer for the frequency shift numerically, assuming $\|\vec{B}\| = 2\text{ T}$. Compare 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.⁴⁹

Some highly magnetized stars (**magnetars**) have much bigger B than what is attainable in the lab, so this effect gives a useful way to establish the value of B on a distant object.

⁴⁹ **[T2]** 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.

CHAPTER 20

First Look at Energy and Momentum Transport by Waves

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

$$\Phi_{\vec{k},\omega}(t, \vec{r}) = e^{i(\vec{k}\cdot\vec{r}-\omega t)}, \quad [18.22, \text{page } 266]$$

and consider a solution to Maxwell's equations that propagates along the $+\hat{z}$ direction and is linearly polarized along \hat{x} :¹

$$\vec{E}_x = \frac{1}{2}i\omega\zeta\Phi_{k\hat{z},\omega} + \text{c.c.}, \quad \vec{B}_y = \frac{1}{2}ik\zeta\Phi_{k\hat{z},\omega} + \text{c.c.} \quad (20.1)$$

Here ζ 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 η . That is, it feels a friction force $-\eta(d\vec{r}_\perp/dt)$.

¹See Section 18.7 (page 266).

In the limit of strong friction, we may neglect inertia in Newton's law of motion and the value of m is irrelevant:

$$0 = -\eta \frac{d\vec{r}_\perp}{dt} + q \left(\vec{E} + \frac{d\vec{r}_\perp}{dt} \times \vec{B} \right)_\perp.$$

The last term on the right equals zero, because $d\vec{r}_\perp/dt$ and \vec{B} both lie in the xy plane, so their cross product has no component in that plane. Thus,

$$\frac{d\vec{r}_\perp}{dt} = \frac{q\vec{E}}{\eta}.$$

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 ζ is real,

$$\mathcal{P} = \vec{f}_\perp \cdot \frac{d\vec{r}_\perp}{dt} = q^2 \|\vec{E}\|^2 / \eta \quad (20.2)$$

$$= \frac{q^2 \omega^2 \zeta}{4\eta} \left\| i\hat{x}e^{-i\omega t} - i\hat{x}e^{+i\omega t} \right\|^2 = \frac{q^2 \omega^2}{\eta} \zeta^2 (\text{Im } e^{-i\omega t})^2. \quad (20.3)$$

This quantity is always greater than or equal to zero. Its time average is

$$\langle \mathcal{P} \rangle = \frac{q^2 \omega^2}{2\eta} \zeta^2. \quad (20.4)$$

Your Turn 20A

- Check that the units in this formula (and every formula) make sense.
- Also, redo this derivation for the more general case in which η is not so huge, so that we must also account for the inertial term $m(d^2\vec{r}_\perp/dt^2)$ in Newton's law. Check that the limits $m \rightarrow 0$ and $\eta \rightarrow \infty$ 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 \vec{k} :

$$\vec{f}_\parallel = q \left(\frac{d\vec{r}_\perp}{dt} \times \vec{B} \right)_\parallel = q \left(\frac{q\vec{E}}{\eta} \times \vec{B} \right)_\parallel. \quad (20.5)$$

Substitute Equation 20.1:

$$\begin{aligned}\vec{f}_{\parallel} &= \frac{q^2\omega k}{\eta} \frac{1}{4} ((\hat{x}i\zeta e^{-i\omega t} + \text{c.c.}) \times (\hat{y}i\zeta e^{-i\omega t} + \text{c.c.}))_{\parallel} \\ &= -\frac{q^2\omega k\zeta^2}{4\eta} (ie^{-i\omega t} + \text{c.c.})^2 = \frac{q^2\omega k\zeta^2}{\eta} (\text{Im } e^{-i\omega t})^2.\end{aligned}\quad (20.6)$$

The time average is then

$$\langle \vec{f}_{\parallel} \rangle = \frac{q^2\omega k\zeta^2}{2\eta}.\quad (20.7)$$

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 $m \rightarrow 0$ and $\eta \rightarrow \infty$ 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, we just found 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
- It also supplements ordinary gas pressure in stars, opposing gravitational collapse (until the nuclear fuel is exhausted).
- It detonates thermonuclear bombs.

Comet tails move to be always directed away from the Sun.

Early Universe expansion was faster than predicted from gas pressure alone.

Fluid equilibrium in stars maintains lower density than predicted from gas pressure alone.

Small objects can be trapped and manipulated with light.

²Chapter 35 will do a more systematic job, at the expense of more abstraction.

³Section 37.4 will give a quantitative formula.

- It allows exquisitely fine manipulation of micrometer-size objects via optical tweezers.⁴
- One day it may even provide a tiny but inexhaustible source of impulse for “solar sail” spacecraft.⁵

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:

$$\frac{\text{rate of energy extraction}}{\text{rate of momentum extraction}} = c. \quad (20.8)$$

Everything specific to our silly little imagined system (amplitude ζ , 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 $\mathcal{E} = \hbar\omega$. 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 $p = \hbar\omega/c$, or

$$\mathcal{E} = pc. \quad (20.9)$$

That result sounds paradoxical: Newtonian mechanics instead says that $\mathcal{E} = p^2/(2m) = pv/2$! Chapter 31 will give Einstein’s resolution to this apparent paradox.

20.4 CIRCULAR AND ELLIPTICAL POLARIZATIONS

Section 18.10.1 (page 271) 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 $\vec{\zeta}$ be any complex vector satisfying $\vec{\zeta} \cdot \vec{k} = 0$.

⁴Section 21.3.2 will describe this technology.

⁵See Problem 20.1.

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 work in the overdamped limit 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.

PROBLEMS

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*, April 2016.

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.”

- 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.
- 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 ω travels along the $+\hat{z}$ direction.

- Write the electric and magnetic fields analogous to Equations 20.1, again parameterized by a single real constant ζ 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

$$m(d^2\vec{r}_\perp/dt^2) = -\eta(d\vec{r}_\perp/dt) + (\text{Lorentz force}).$$

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).

- 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 ζ , η , q , and possibly other constants.
- 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} .
- The wave also pushes the charge in the xy plane, exerting a *torque* τ_z . Find the average of this torque over a cycle.
- The ratio $\langle\tau_z\rangle/\mathcal{P}$ has a remarkably simple form: Find it in terms of the parameters in the problem.
- 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

Circularly polarized light can exert a mechanical torque on objects in its path.

lump of energy \mathcal{E}_* and a lump of angular momentum L_* . 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}_* and L_* using your result in (e). Draw a conclusion about the intrinsic angular momentum carried by one packet.

CHAPTER 21

Ray Optics and the Eikonal

21.1 FRAMING: ALMOST-PLANE WAVES

Real-world problems are mathematically harder than the idealized problems we encounter in our first textbooks. Often, we need some sort of unfair advantage before we can make a dent in a real-world 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. In this chapter, we'll study the propagation of light in media that, while not uniform, at least vary in limiting ways:

- We'll study piecewise uniform media that meet at a sharp, planar boundary (or a nonplanar boundary that is nearly flat on the length scale of wavelength). In this situation, we can join together plane waves in the two media to get an overall solution.
- We'll also study the opposite limit of media whose properties vary slowly (again on length scales much bigger than the wavelength of the light under consideration). Here a class of approximate solutions, which we'll call *almost-plane* waves, is the right construction.

Such situations arise in many practical problems, and let us approach otherwise forbiddingly complex situations.

In everyday life, 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

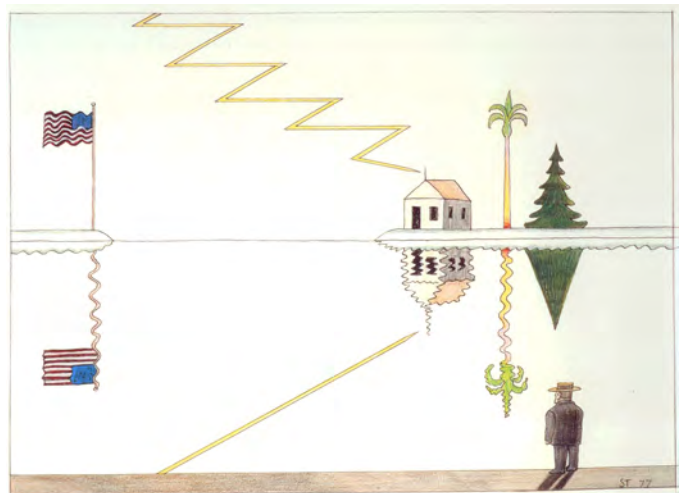
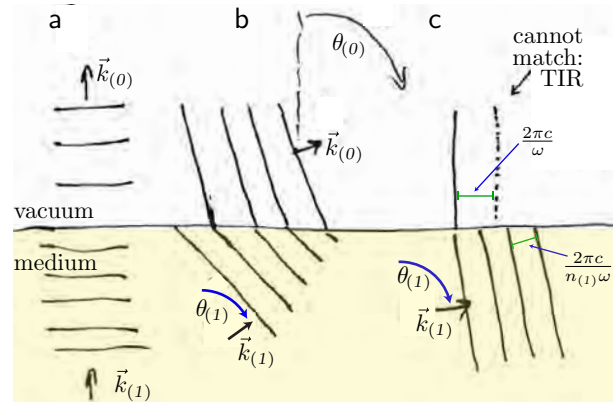


Figure 21.1: [Saul Steinberg.]

Figure 21.2: Refraction versus total internal reflection. 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.) *Left*, perpendicular incidence ($\theta_{(1)} = \theta_{(0)} = 0$). *Center*, the angle $\theta_{(1)}$ is nonzero but less than the critical value. *Right*, no solution is possible when the angle $\theta_{(1)}$ is too large. If the wave is coming from within the medium (traveling upward) as shown, then in this case it cannot escape and must be totally internally reflected.



explore that question, which turns out to make sense in the same limiting situations just described.

Electromagnetic phenomenon: A lens with appropriately graded index can minimize spherical aberration.

Physical idea: Light rays curve gradually in a gradient-index material, just as they bend sharply at a sharp interface.

21.2 LIGHT STILL 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 permittivity¹ $\epsilon_{(1)}$. The assumption of isotropy means that $\epsilon_{(1)}$ is a scalar. The analysis in Chapter 18 showed that there will be transverse wave solutions with dispersion relation $\omega = (c/n_{(1)})\|\vec{k}\|$, where the **refractive index** $n_{(1)} = \sqrt{\epsilon_{(1)}/\epsilon_0}$. For dielectric materials, it is larger than 1.

21.3 PIECEWISE-UNIFORM MEDIUM

21.3.1 The refraction law arises from matching fields across a planar boundary

Consider a sharp, planar junction between an otherwise uniform dielectric medium I and vacuum. (Junctions between two media can be handled similarly.) We assume that the medium and its boundary are not changing in time (think about a chunk of glass). Then Maxwell's 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 $\propto e^{-i\omega t}$.

¹Chapter 49 will justify this prescription in greater detail. For simplicity, we will also assume that $\mu = \mu_0$, but similar formulas ensue if that's not the case.

The coefficients are not constant in *space*, however, due to the boundary, so we *can't* expect solutions with a single overall $e^{i\vec{k}\cdot\vec{r}}$. Separately on each side, however, there are solutions of this form. So consider a trial solution with transverse plane waves on either side of the boundary, with locally constant wavevectors $\vec{k}_{(1)}$ and $\vec{k}_{(0)}$.

Figure 21.2a illustrates the situation when $\vec{k}_{(0)}$ is perpendicular to the interface. The horizontal lines represent planes of constant phase, for example, loci where a linearly-polarized wave has $\vec{E} = \vec{B} = 0$. These “wavefronts” are more widely spaced on the vacuum side because the two regions have the same frequency but different wave speed.

Figure 21.2b shows a more general situation. The component of electric field perpendicular to the interface may change discontinuously as we cross it, due to the possibility of bound charges there, but Faraday’s law shows that the parallel components must be continuous.² Then in particular, the loci of zero parallel \vec{E} field must match up across the boundary. The only way for that condition to be consistent with different wavefront spacing is for the wavevector to *change direction*, as shown.

A plane wave changes direction as it crosses a planar interface between two dielectric media.

Your Turn 21A

Because the frequency is the same on each side, we know from the dispersion relations that $\|\vec{k}_{(1)}\| = n\|\vec{k}_{(0)}\|$.

a. Convince yourself geometrically that the direction of the wavevector must change at a vacuum–medium interface according to

$$\sin \theta_{(0)} = n \sin \theta_{(1)}. \quad (21.1)$$

b. How does this formula change for an interface between two dielectric media?

Figure 21.3a shows one familiar consequence of refraction. It also illustrates a less cluttered visual representation of refraction: Instead of drawing all the wavefronts, the dashed lines in the figure show “rays,” which we provisionally define as piecewise-straight lines that are everywhere parallel to \vec{k} .

21.3.2 Optical tweezers exploit the momentum transfer implied by refraction

Figure 21.4 shows how a spherical object with differing refractive index from its surroundings will feel a net sideways force when it encounters a beam of light. 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 mentioned in Section 20.2.3.

21.3.3 Spherical aberration limits the practical focusing power of glass lenses

The law of refraction is also the basis for the focusing of light by a lens. Figure 21.5a shows an incoming plane wave, represented by a bundle of parallel rays, that impinge on a spherical dielectric object. If the object’s diameter is much bigger than the

²Section 18.4.2 (page 262).

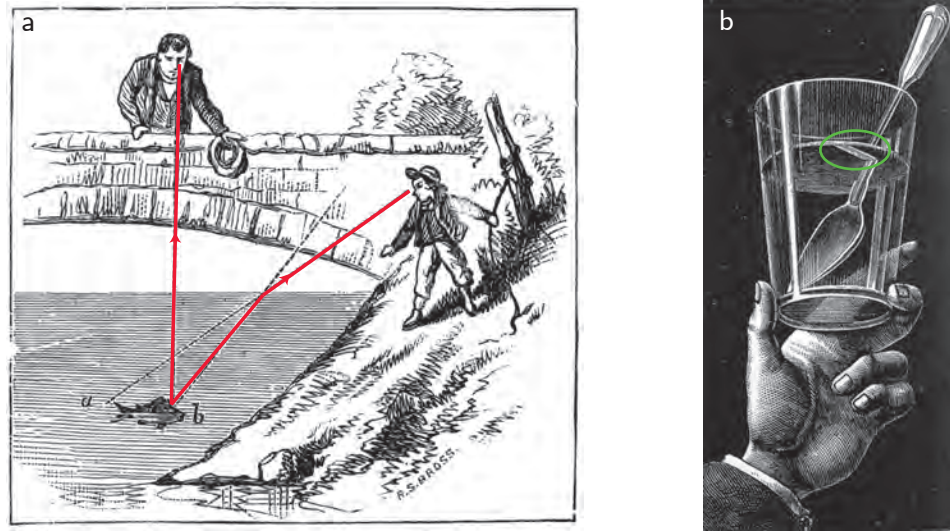
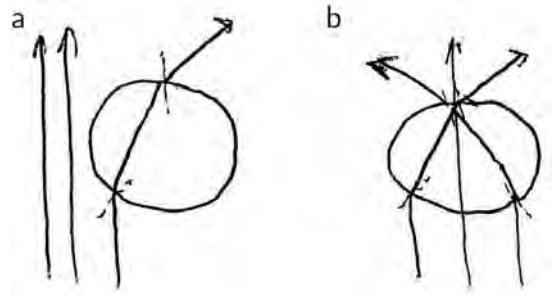


Figure 21.3: Refraction versus total internal reflection. (a) Light reflected from the fish's head travels in every direction along rays (*red*). The observer on the bridge gets an accurate impression of the location of the fish (head is at *b*). The observer on the bank, unconsciously assuming that light travels on straight lines, gets an inaccurate impression (the head seems to be at *a*). (b) Viewed from below, part of the air-water interface appears to be a mirror—an instance of total internal reflection.

Figure 21.4: Basis of optical tweezers. Generation of transverse force on a dielectric sphere by a beam of light, in the ray-optics regime. (a) The sphere is not centered in the beam. Two rays in the beam miss the sphere altogether. One ray is bent, undergoing a change in its momentum (a vector quantity). Newton's third law then requires that the bead receives a continuous momentum transfer (force): It recoils toward the left. (b) If the sphere is centered in the beam, then the central ray is undeflected, and the ones flanking it make cancelling contributions to the net transverse impulse.



wavelength of the light, then we may apply the law of refraction separately to each of the lines shown. We start with a constant field of \vec{k} vectors on the left (a plane wave), convert each to a new direction upon entering the medium via the law of refraction, extend the resulting rays till 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**. However, the focusing is not perfect. The figure shows piecewise-straight lines that bend according to Equation 21.1, with index values appropriate for glass and water. The lines close to

Even if we neglect diffraction, a perfectly spherical lens will not focus light perfectly.

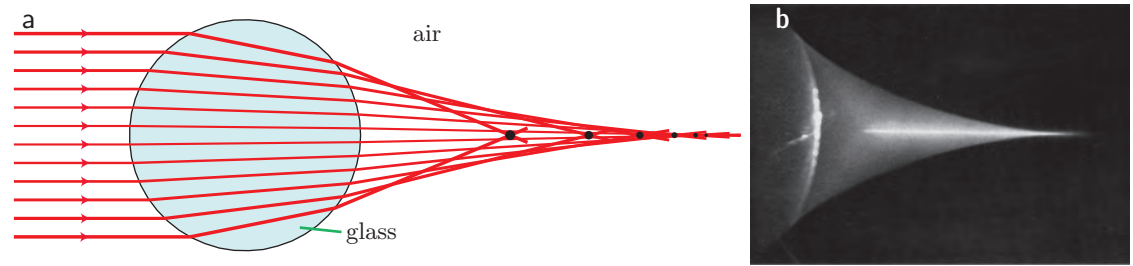


Figure 21.5: Spherical aberration. (a) [Ray diagram.] Parallel rays arriving at a spherical lens, and passing close to its center (*thinner lines*), nearly coincide at a common focus (*smallest dot at far right*). However, rays initially farther from the axis (*heavier lines*) cross it in a spread-out array (*larger dots*). The rays shown were computed by using the law of refraction (Equation 21.1), for a glass sphere immersed in water. (b) [Photograph.] The spread-out focus is visible as the bright line in this photo. [(b) From Cagnet et al., 1962.]

the center do arrive at a common point (far right in the figure), but the ones farther from center do not, a phenomenon called **spherical aberration** that limits the useful light-collecting region of conventional microscope lenses. Section 21.5.2 will outline what can be done about this problem.

21.3.4 Total internal reflection arises when there is no solution to the refraction equation

An interface between two transparent media will act like a mirror at high incidence angle.

Figure 21.2c shows geometrically that there may be no solution of the type described above, if the angle of incidence exceeds a critical value. In terms of your result from Your Turn 21A, $\sin \theta_{(t)}$ must be smaller than $1/n$ because $\sin \theta_{(i)}$ cannot exceed 1. If a plane wave originates in the medium (directed toward the vacuum side) and this condition is violated, then there can be no transmitted plane wave. All incoming energy instead gets reflected back into the medium, a phenomenon called **total internal reflection** (TIR, Figure 21.3b).

Your Turn 21B

- What if a plane wave originates on the vacuum side (\vec{k} directed toward the medium)?
- Imagine yourself submerged in a swimming pool. Looking straight upward, you see the sky. But beyond a certain angle, the surface above you looks like a mirror (try it!). Why?

Light traveling down a thin fiber remains trapped in it.

TIR is the basis for guiding light through glass fibers. As long as the fiber does not bend too sharply, an initially axially propagating wave will remain trapped inside it.³ Such a fiber can carry vastly more data than a coaxial cable because the frequency

³ [T2] 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, Section 12.4'c, page 186).

of visible light is so much higher than the radio frequencies that the coax can carry. Also, a 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 noninvasive medical diagnosis.

When we go beyond ray optics, we'll find that TIR is not quite total after all (Section 49.4, page 610): Small disturbances penetrate less than about a wavelength into the second medium even at high angle of incidence. This “evanescent wave” phenomenon is the basis for an important microscopy technique.

21.4 GRADIENT-INDEX MEDIUM

21.4.1 Rays of light can be regarded as streamlines of energy flux

We can now return to the framing questions (Section 21.1): What is a ray? Extending our provisional definition in Section 21.3.1,

Rays of light are the streamlines of the energy flux for a solution that is locally approximately a plane wave.⁴

Just as a given chamber can have water flowing in various ways, so too a given optical system can have various locally-plane wave solutions, and hence various different families of rays passing through it. The definition makes sense in other ways as well: For example, when we focus light these streamlines converge, leading to high energy density (enough to ignite paper under a magnifying glass, for example).

Consider the streamlines of the energy flux \vec{j}_E . In a uniform, isotropic medium, we found plane wave solutions, for which \vec{k} is a constant. Because the energy flux is always parallel to \vec{k} ,⁵ the streamlines of a plane wave are a family of straight, parallel lines. But this idea has wider usefulness than that one example.

Indeed, in a piecewise-uniform medium the law of refraction can be interpreted 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). We will continue to use this viewpoint when we have a continuously varying refractive index.

21.4.2 Almost-plane waves are a useful idealization when there is a separation of length scales

In a medium whose refractive index changes, but only slowly compared to the wavelength of light, it seems reasonable to look for solutions to Maxwell’s equations that *locally* resemble plane waves, but for which \vec{k}_{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 the ionosphere. Section 21.4.6 will discuss the resulting refraction phenomenon.

⁴See Section 0.3.1 (page 7).

⁵For an anisotropic medium like calcite, we must reconsider this statement.

- The air close to a hot road surface has nonuniform temperature, and hence also density and hence also refractive index, leading to mirage phenomena (see Section 21.5.1).
- Our own eye lenses have this property: Although they are transparent, the index varies continuously from a maximum at the center to a minimum at the surface (see Section 21.5.2).
- Perhaps most exotic, Einstein’s gravity theory predicts that even empty space will behave like an inhomogeneous medium for light, if strong gravitational fields are present (Section 21.5.3).

We might expect some continuous version of the law of refraction to hold in situations like these. Let’s find it.

21.4.3 The eikonal equation controls propagation of an almost-plane wave

Solving vector PDEs without a lot of symmetry is in general difficult. But at least the situations just mentioned are all *stationary*, that is, invariant under time translation, so we can again assume harmonic time dependence for our solutions. Moreover, all of the situations in the preceding list share a convenient aspect: 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, $c/(L_0\omega) \ll 1$. In this regime, it’s reasonable to look for approximate solutions to Maxwell’s equations of **eikonal** form

$$\vec{A} = \frac{1}{2}e^{-i\omega t}\vec{\zeta}(\vec{r})e^{i\omega\beta(\vec{r})/c} + c.c. \quad (21.2)$$

In this expression, $\beta(\vec{r})$ is called the **eikonal function**, or simply “the eikonal.” For a plane wave it would be a linear function, $\hat{k} \cdot \vec{r}$. The other unknown function, $\vec{\zeta}(\vec{r})$, allows for the possibility that the polarization is not constant throughout space, unlike a plane wave. We assume, however, that both β and $\vec{\zeta}$ vary slowly in space, with a characteristic length scale similar to L_0 .

We should ask whether the eikonal trial solution works, to leading order in the small parameter⁶ $c/(L_0\omega)$. We will now develop a framework called **ray optics** that is useful for handling such situations.

Close to any point \vec{r}_* , our trial solution Equation 21.2 thus resembles a plane wave with local wavevector $\vec{k}_{\text{local}} = \frac{\omega}{c}\vec{\nabla}\beta|_{\vec{r}_*}$. In particular, the energy flux everywhere points along $\vec{\nabla}\beta$. So once we establish which particular eikonal functions $\beta(\vec{r})$ give solutions to Maxwell’s equations, we will find examples and compute their gradients. The streamlines of the resulting vector fields will be the rays that we seek.

First, impose Coulomb gauge:⁷

$$0 = \frac{1}{2}e^{-i\omega t}(\vec{\nabla} \cdot \vec{\zeta} + \vec{\zeta} \cdot \frac{i\omega}{c}\vec{\nabla}\beta)e^{i\omega\beta/c} + c.c.$$

We may drop the first term, because the second dominates in the short-wavelength limit. Thus, not surprisingly, $0 = \vec{\zeta} \cdot \vec{k}_{\text{local}}$, just as we found for plane waves.

⁶ **[T2]** One way to describe this short-wavelength limit is to say that we are neglecting diffraction effects; this is the regime where we may hope that a “ray” concept will be useful.

⁷Section 18.8.2 (page 268).

The Maxwell equations then take the form in Equation 18.29 (page 270):

$$\frac{1}{2}e^{-i\omega t} \left(\vec{\nabla}_j \left((\vec{\nabla}_j \vec{\zeta}_i + \frac{i\omega}{c} (\vec{\nabla}_j \beta) \vec{\zeta}_i) e^{i\omega\beta/c} \right) \right) + \text{c.c.} = -\left(\frac{\omega}{c}\right)^2 \frac{1}{2} \vec{\zeta}_i e^{-i\omega t + i\omega\beta/c} + \text{c.c.}$$

Again drop the first term in parentheses on the left, because the other term dominates it for large ω .

$$\frac{i\omega}{c} (\vec{\nabla}_j \beta) \cdot (\vec{\nabla}_j \vec{\zeta}_i) + \frac{i\omega}{c} \vec{\zeta}_i \nabla^2 \beta + \left(\frac{i\omega}{c}\right)^2 \vec{\zeta}_i (\vec{\nabla} \beta)^2 = -\left(\frac{\omega}{c}\right)^2 \vec{\zeta}_i.$$

The last term on the left dominates the others, so we find that our trial solution works if

$$\|\vec{\nabla} \beta\|^2 = 1. \quad \text{eikonal equation in vacuum} \quad (21.3)$$

Some simple solutions to the eikonal equation include $\beta(\vec{r}) = \hat{k} \cdot \vec{r}$ (plane wave) or $\|\vec{r}\|$ (spherical wave). In the former case, the rays 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 Maxwell's equations, which are linear, with the new PDE Equation 21.3, which is *nonlinear*. But we do not always need all the information in the phase function β . Let's now convert our equation into a direct characterization of the rays (streamlines of $\vec{\nabla} \beta$) themselves.⁸

21.4.4 Rays in vacuum

The rays are a family of curves, each of which is everywhere tangent to \vec{k}_{local} . We can write any curve in parametric form as $\vec{\ell}(s)$, where s is arc length. That is, $d\vec{\ell}/ds$ is the field of unit tangent vectors all along the curve (recall Equation 21.3). The tangent must be parallel to $\vec{\nabla} \beta$, which itself is everywhere a unit vector, so

$$\frac{d\vec{\ell}}{ds} = \vec{\nabla} \beta|_{\vec{\ell}(s)} \quad \text{for all } s. \quad (21.4)$$

One way to characterize a curve is to state its **curvature**, that is, how its tangent vector deviates from being a constant. More precisely, we define the curvature vector as the derivative of the unit tangent to the curve with respect to arc length, finding:

$$\text{curvature} = \frac{d^2 \vec{\ell}}{ds^2} = \frac{d}{ds} \left(\vec{\nabla} \beta|_{\vec{\ell}(s)} \right).$$

The right side of this formula is the derivative of a function as we walk along the curve. To evaluate it, we can find the dot product of the gradient (that is, all partial derivatives) with the unit tangent and apply to $\vec{\nabla} \beta$:

$$\frac{d^2 \vec{\ell}_i}{ds^2} = \left(\frac{d\vec{\ell}}{ds} \cdot \vec{\nabla} \right) \vec{\nabla}_i \beta \Big|_{\vec{\ell}} = (\vec{\nabla}_j \beta) (\vec{\nabla}_j (\vec{\nabla}_i \beta)) = (\vec{\nabla}_j \beta) (\vec{\nabla}_i (\vec{\nabla}_j \beta))$$

⁸Note that the polarization vector drops out of Equation 21.3, so we learn nothing about $\vec{\zeta}$ from this approach other than that it must everywhere be perpendicular to $\vec{\nabla} \beta$. 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 behavior.

$$= \frac{1}{2} \vec{\nabla}_i \|\vec{\nabla} \beta\|^2 = 0 \quad \text{by Equation 21.3.}$$

Note that the phase function β has disappeared from this expression; we don't need to solve the eikonal equation after all in order to find the rays. Instead, we conclude that *the curvature is zero*:

$$\text{Light rays in vacuum are straight lines.} \quad (21.5)$$

That makes sense: Ray optics neglects diffraction, and when that approximation holds indeed objects cast sharp shadows. The two illustrative families of solutions found earlier (straight parallel rays and straight radial rays) both obey this rule.

21.4.5 Rays bend continuously in a gradient-index medium

We now consider 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.)

Your Turn 21C

- a. Show that generalizing our previous derivation (Equation 21.3) gives

$$\|\vec{\nabla} \beta\|^2 = n^2. \quad \text{eikonal equation in medium} \quad (21.6)$$

- b. Show that therefore the analog to Equation 21.4 gives the tangent to a ray as

$$\frac{d\vec{\ell}}{ds} = \frac{\vec{\nabla} \beta}{n} \Big|_{\vec{\ell}(s)}. \quad (21.7)$$

Your Turn 21D

- a. Next show

$$\frac{d}{ds} \left(n(\vec{\ell}) \frac{d\vec{\ell}}{ds} \right) = \vec{\nabla} n \Big|_{\vec{\ell}(s)}. \quad \text{ray equation} \quad (21.8)$$

at every position s along a ray.

- b. Check that your result from (a) is compatible with arc length parameterization. That is, show that $\|d\vec{\ell}/ds\|$ remains equal to one if it starts that way.

As in the vacuum case, the ray equation makes no explicit mention of the eikonal function β . It tells us how light rays bend as they pass through a medium—a generalization of the law of refraction. When the ray-optics approximation is justified, this equation reduces Maxwell's partial differential equations to the *ordinary* vector differential equation (21.8), a net simplification.

A bit like Newton's $\vec{f} = d^2\vec{r}/dt^2$, 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.⁹ Solving systems of ODEs numerically is a routine task.

21.4.6 Shortwave radio skip (skywave transmission)

After G. Marconi and others 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 move 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 a variation at high altitude could bounce (“skip”) radio signals at high enough angle of incidence, similarly to the trapping of light in a curved optical fiber (Section 21.3.4). This hypothesis also explained why the effect was more pronounced at night¹⁰ and at short wavelength (see Chapter 54). Together with other improvements 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 them. Let's investigate.

Suppose that $n(x)$ 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 θ_0 with respect to the vertical. Farther along on the ray, $\theta = \cos^{-1}(\hat{z} \cdot d\vec{\ell}/ds)$ may change. If at any point this angle increases to $\pi/2$, then the ray can bounce (or “skip”) back downward.

Taking the dot product of Equation 21.8 with \hat{z} gives

$$\frac{d}{ds}(n \cos \theta) = \frac{dn}{dz} \Big|_{\vec{\ell}(s)}.$$

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

$$n \cos \theta \frac{d}{dz}(n \cos \theta) = \frac{1}{2} \frac{d}{dz}(n^2 \cos^2 \theta) = \frac{1}{2} \frac{d}{dz}(n^2).$$

Thus, $n^2 \cos^2 \theta - n^2$ is a constant along the ray, a generalized law of refraction:

$$\boxed{n \sin \theta = \text{const} \quad \text{if } n \text{ depends only on } z.} \quad (21.9)$$

⁹Unlike newtonian mechanics, the initial *speed* of light is not arbitrary; in our framework, s is always arc length.

¹⁰Too much atmospheric ionization from solar irradiation leads to absorption during the day.



Figure 21.6: Mirages. (a) Frequently the air near a pavement is warmer than that above, leading to a mirage, where light from the sky appears to be coming upward from the ground. Depending on atmospheric index profile, objects can appear upright, inverted, and/or stretched. (b) Less frequently, a temperature inversion can make an object on the surface appear raised (“superior mirage”). [(b) David Morris/Apex.]

In the special case where n changes suddenly at a planar boundary, this result reduces to the usual law of refraction.

More generally, the index of refraction for a plasma can be smaller than one.¹¹ Hence, as a ray ascends to the ionosphere, n decreases from ≈ 1 at the surface. Equation 21.9 then implies that θ will increase; if θ ever increases to $\pi/2$, then the ray can skip back down to Earth.

Radio waves propagating nearly horizontally can reflect from the ionosphere.

21.5 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.

21.5.1 Mirages rely on our brains’ assumptions about light propagation

A sharp thermal gradient can give rise to mirages.

On a long, flat stretch of highway, solar heating creates a layer of air near $z = 0$ 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 the curved path in Figure 21.6. 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 illusion only appears in the distance, not up close. You’ll work out this and other details in Problem 21.3.

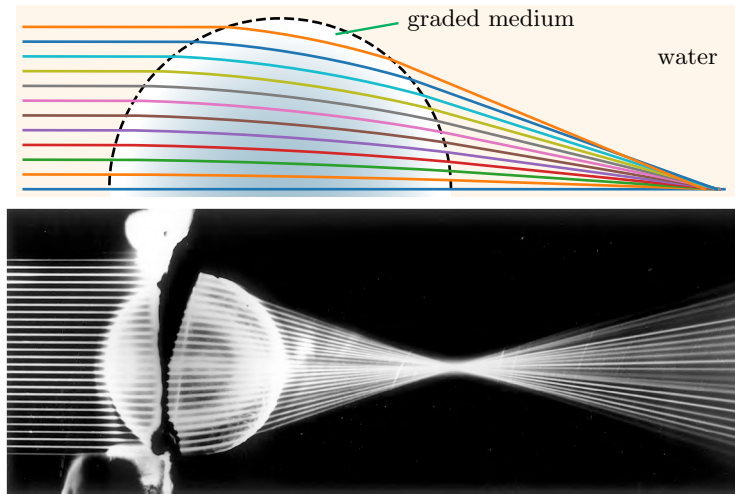


Figure 21.7: [Ray diagram.] **Correction for spherical aberration by a continuously graded refractive index.** (a) A set of parallel incoming rays is shown, with their computed trajectories upon entering the medium. In this case, the rays curve inside the lens, because its refractive index is greater 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 21.4 describes the index function that was used to make this diagram.) (b) Actual light rays traversing the eyelens of an octopus. [From Jagger & Sands, 1999.]

21.5.2 A spherical gradient-index lens can minimize spherical aberration

See Figure 21.7 and Problem 21.4.

21.5.3 Gravitational fields can bend light rays even in vacuum

Einstein's theory of gravitation proposes that space and time can deviate from the cartesian (flat) geometry assumed throughout this book, and that this deviation is responsible for the familiar effects of gravitation. Moreover, because light (and everything else) inhabits spacetime, it, too will be affected by gravitational fields. Of special interest is the fate of a ray that travels through empty space far from any mass, then passes close to a massive object, and finally emerges back into empty space. This ray will be a straight line before and after the flyby, but those two lines may not be parallel, because of the transit through a non-cartesian region during the encounter.

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

$$n_{\text{eff}} \approx 1 - 2\phi_N/c^2 + \dots \quad (21.10)$$

Here ϕ_N is the newtonian gravitational potential far from the mass, and the ellipsis represents terms of higher order in ϕ_N/c^2 .

A gravitational field generates an effective index of refraction.

¹¹Again see Chapter 54.

A spherical lens with appropriately graded index can minimize aberration.

Your Turn 21E

- Equation 21.10 may be unfamiliar to you, so check that the units make sense.
- In the neighborhood of a point mass M , the formula becomes $n_{\text{eff}} \approx 1 + r_*/r$. Look up the mass of our Sun and find a formula for r_* in terms of M/M_{sun} .
- The 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 hence find the maximum value of the effective refractive index along the ray.

21.6 PLUS ULTRA

Erwin Schrödinger was well trained in optics and acoustics. He reasoned that:

- Einstein and de Broglie say that particles correspond to waves.
- Bohr says that in the atomic world, where the length scale is comparable to the de Broglie wavelength, the wave idea explains the observed quantization of energy, analogously to the quantization of acoustic harmonics in an organ pipe.
- It is true that newtonian mechanics seems to rule the macro world.
- But this sounds familiar: Maybe we need to seek a wave equation (not the usual one, but *some* equation with wavy solutions) whose geometric-optics limit gives trajectories that solve Newton's laws (not the law of refraction).

It was already known that, remarkably, newtonian mechanics could be formulated in the way just proposed; that is, its trajectories are the rays solving an eikonal equation. "All" Schrödinger had to do was to find the underlying wave equation and *take it seriously*, even outside of the domain where eikonal approximation holds. This crazy idea needed some interpretation, to be sure. But it worked out well. In fact, it was another of the biggest successful lateral-thinking jumps in scientific history.¹²

FURTHER READING

Semipopular:

Marconi: Larson, 2006.

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, work honored in 2018: www.nobelprize.org/uploads/2018/10/advanced-physicsprize2018.pdf.

Intermediate:

Total internal reflection: Nelson, 2017.

Eikonal approximation: Thorne & Blandford, 2017, chap. 7; Landau & Lifshitz, 1979, chap. 7; Elmore & Heald, 1969, §9.2.

¹²Recall Section 0.4.1 (page 11).

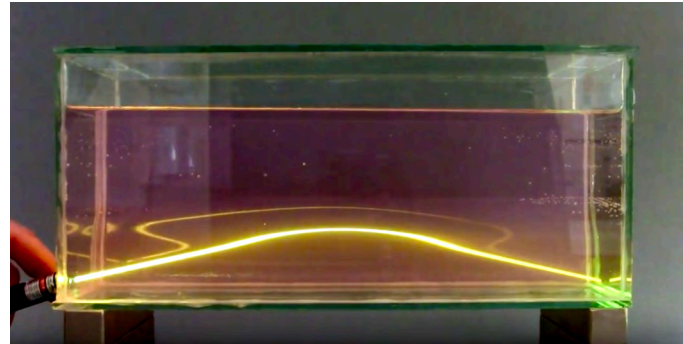


Figure 21.8: See Problem 21.1 and Media 10.

Ray equation: Elmore & Heald, 1969, §9.2.

Mirage: Richey et al., 2006.

Gravitational lensing: Basic: Schutz, 2022, chap. 11. Advanced: Straumann, 2013, chap. 5, Nye, 1999.

Optical tweezers: Jones et al., 2015; Perkins, 2014; van Mameren et al., 2011; Bechhoefer & Wilson, 2002.

Technical:

Optical tweezers: www.cell.com/biophysj/collections/optical-tweezers.

PROBLEMS

21.1 *Poor wandering one*

Figure 21.8 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. What do you think might cause the light to take this bizarre, wandering path?

21.2 *Waves in conductive medium*

An electromagnetic plane wave propagates through vacuum, then enters a medium. The medium is not polarizable ($\epsilon = \epsilon_0$, $\mu = \mu_0$). However, it is electrically conductive, obeying an ohmic relation with conductivity κ :

$$\vec{j} = \kappa \vec{E}.$$

Assume the medium is everywhere electrically neutral.

- a. Find the dispersion relation for plane waves of angular frequency ω traveling through such a medium, and interpret it physically.
- b. 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 Maxwell's 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.

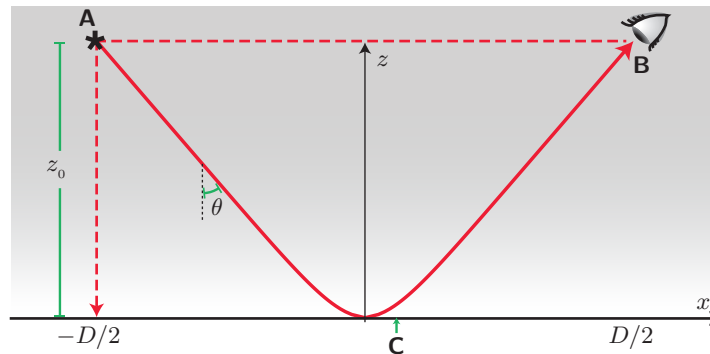


Figure 21.9: [Mathematical function.] **Curved light rays in an inhomogeneous medium.** The x and z axes are not drawn to the same scale. *Dashed lines* show the simple solutions mentioned in Problem 21.3. An observer who assumes straight-line propagation will interpret light from **A** as having come from **C**. The figure is analogous to what we imagined for short-wave radio skip in Section 21.4.6 (page 303), but upside down.

21.3 Mirage

Section 21.5.1 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 index of refraction. This is the special case of a “gradient-index” material whose index depends only on height z .

Section 21.4.6 worked out a general formula for the angle θ that a ray’s trajectory makes with the z -axis (Equation 21.9). This condition has two unsurprising solutions: 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 solutions that are *curved*.

Suppose that the density profile $n(z)$ is strictly increasing as z increases, and that θ starts out tilted downward ($0 < \theta < \pi/2$). Then θ can increase as z decreases, potentially even leveling off ($\theta \rightarrow \pi/2$), as shown in Figure 21.9.

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, $z = h(x)$, where $h(\pm D/2) = z_0$. We wish to find functions $h(x)$ that give solutions to Equation 21.9 subject to these boundary conditions.

- To be specific, suppose that $n(z) = n_\infty(1 - \alpha e^{-z/L})$, where n_∞ is the index of air at 30°C, $n_\infty(1 - \alpha)$ is the index of air at 50°C, $L = 20$ cm, and your eyes are $z_0 = 2$ m off the ground. One can look up these values for the two indices of refraction for visible light:¹³

$$30^\circ\text{C} : n = 1.000262; \quad 50^\circ\text{C} : n = 1.000244.$$

Use Equation 21.9 to see how close θ_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).

- Reformulate Equation 21.9 as a differential equation determining the entire curve; that is, an equation involving dh/dx . Solve it analytically or numerically for the

¹³At 633 nm, 101.3 kPa pressure, 50% relative humidity.

situation discussed above. If any simplifying approximations are valid, go ahead and use them. Use the smallest value of θ_0 for which you found in (a) that a mirage would be possible, 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.)

21.4 Gradient-index lens

Use the ray-optics approximation for this problem, and neglect the possibility of reflection at interfaces. Section 21.4.6 considered light ray trajectories in a nonuniform medium whose refractive index depends on only one cartesian coordinate, the height.¹⁴ 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 center of a spherical lens. Section 21.5.2 mentioned that this situation holds for the eye lenses of fish, and claimed that such nonuniformity can eliminate much of the aberration created by a uniform spherical lens (compare Figure 21.5a to Figure 21.7).

In this problem, you can scale all lengths by the radius a of the sphere, that is, work in terms of $\bar{r} = r/a$ 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 eyes typically have $n_c \approx 1.52$ and $n_p \approx 1.38$, and are immersed in watery media ($n_w \approx 1.33$) on both sides. W. Jagger investigated a nonuniform but spherically symmetric index of refraction profile:

$$n(\bar{r}) \approx n_c(1 + K(0.82\bar{r}^2 + 0.30\bar{r}^6 - 0.12\bar{r}^8)),$$

It will be convenient to define $g(\bar{r}) = n^{-1}(dn/d\bar{r})$.

- Choose cartesian coordinates with the lens center at the origin and a plane passing through that origin, say the xy plane. Write out both components of the ray equation (Equation 21.8, page 302), which determines the streamlines $\vec{\ell}(s)$. It's a pair of coupled, second-order ordinary differential equations in the two coordinates of a curve lying in the chosen plane, $\vec{\ell}_x(s)$ and $\vec{\ell}_y(s)$. Parameterize the curve by arc length s , so that $\|\mathbf{d}\vec{\ell}/ds\| = 1$.
- Now generate a picture similar to Figure 21.7, by constructing a series of solutions to the ray equation. 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").
- Use the law of refraction to find the tangent vector to the ray just *after* it enters the lens.
- 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.
- Follow your solution to find the value \bar{s}_{exit} at which \bar{r} once again reaches the value 1.
- The tangent vector $\mathbf{d}\vec{\ell}/d\bar{s}|_{\bar{s}_{\text{exit}}}$ 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.

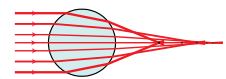


Fig. 21.5a (page 298)

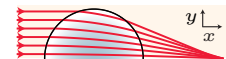


Fig. 21.7 (page 305)

¹⁴See also Problem 21.3.



Figure 21.10: Einstein rings. The arcs at the center of this image from the Hubble Space Telescope are actually the distorted light of distant galaxies, stretched into an Einstein ring 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].

- 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). Repeat for each ray that you wish to trace.

21.5 Gravitational lens

Use the ray-optics approximation for this problem. Section 21.4.6 considered light ray trajectories in a nonuniform medium whose refractive index depends on only one cartesian coordinate, the height.¹⁵ 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 21.10).

Choose cartesian coordinates with the lens center at the origin, and a plane passing through that origin, say the xy plane.

- a. Write out both components of the ray equation (Equation 21.8, page 302), which determines paths $\vec{\ell}(s)$. It's a pair of coupled, second-order ordinary differential equations in the two cartesian coordinates $\vec{\ell}_x(s)$ and $\vec{\ell}_y(s)$ 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, that is, to work in terms of $\bar{s} = s/r_*$ and so on. It will also be convenient to define $g(\bar{r}) = n^{-1}(dn/d\bar{r})$.

- b. Show that $g(\bar{r}) = -1/(\bar{r}^2(1 + 1/\bar{r}))$.
- c. Consider a series of rays that each start at $\bar{x}_0 = -10$, 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 \bar{y}_0 . Because Equation 21.10 is only valid

¹⁵See also Problem 21.3.

- for weak gravitational fields, only examine values of \bar{y}_0 that are greater than (say) 5.
- d. Now generate a picture analogous to Figure 21.7, by having your computer draw your solutions.
 - e. Your trajectories are distinguished by their y_0 values. For each, find the value x_* at which the trajectory hits the symmetry axis $y = 0$ and graph x_* as a function of y_0 .
 - f. Your trajectories become straight lines far from the point mass, and in particular when they hit the symmetry axis. So you can find the angle of approach θ_* at that intersection from your numerical result in (c). This gives an apparent angular location in the sky. By the problem's axial symmetry, the background star appears as a ring with this angular radius: the **Einstein ring** (Figure 21.10). Make a graph of θ_* as a function of y_0 .
 - g. Finally, combine your two previous results to graph θ_* as a function of x_* , that is, apparent angular width 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 24

Partial Polarization

24.1 FRAMING: STOKES PARAMETERS

We found plane-wave solutions to Maxwell's equations. Each such solution had a single, definite wavevector \vec{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, a plane wave 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 if we pass it through a monochromatic filter, any real 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 looks 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 (\vec{E} may be complex), 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.¹ They just measure averages over a

¹Radiotelescopes can in principle measure this, and so pick up more detailed information about the waves they detect than instruments like bolometers or cameras.

time that's 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 286) 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 multiplies $\vec{E}(t)$ by a scalar fraction that depends on ω . A polarizer multiplies it by a matrix 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:²

$$\langle \vec{E}_i e^{-i\omega t} \vec{E}_j e^{-i\omega t} \rangle, \text{ their conjugates, and } \langle \vec{E}_i e^{-i\omega t} \vec{E}_j^* e^{+i\omega t} \rangle \text{ where } i, j = 1, 2.$$

Of these, the first eight average to zero because of their fast time variation.

The remaining four quantities constitute a 2×2 hermitian matrix:

$$\vec{J}_{ij} = \langle \vec{E}_i \vec{E}_j^* \rangle \quad \text{or} \quad \vec{J} = \langle \vec{E} \otimes \vec{E}^* \rangle. \tag{24.2}$$

Although this matrix does not contain enough information to determine \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 2×2 hermitian matrix can be written in terms of four real quantities. A traditional choice is to introduce the four **Stokes parameters**:

$$\vec{J} = \frac{1}{2} \begin{bmatrix} s_0 + s_1 & s_2 - is_3 \\ s_2 + is_3 & s_0 - s_1 \end{bmatrix}. \tag{24.3}$$

Again: The Stokes parameters describe light for the purposes of detectors of the sort used in most optics experiments.³ Note that

$$\det \vec{J} = (s_0^2 - s_1^2 - s_2^2 - s_3^2)/4. \tag{24.4}$$

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.

For optical purposes, a plane wave of light may be described by a point inside an abstract sphere.

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

²The magnetic field of a plane wave just tracks the electric field, so we would learn nothing new by considering terms with \vec{B} .

³Some books factor out the overall normalization and define Stokes parameters as $\xi_1 = s_2/s_0$, $\xi_2 = s_3/s_0$, $\xi_3 = s_1/s_0$.

distribution of electric field vectors. In that language, we propose a classical model of unpolarized light in which the two complex coefficients \vec{E}_1 and \vec{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:

$$\wp(\vec{E}_1, \vec{E}_1^*, \vec{E}_2, \vec{E}_2^*) = f(\|\vec{E}\|^2). \quad \text{unpolarized light} \quad (24.5)$$

Here the length-squared, $\|\vec{E}\|^2$, of a complex vector is understood to mean $\sum_i \vec{E}_i \vec{E}_i^*$. The real function f may be chosen such that $\langle \|\vec{E}\|^2 \rangle$ 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

$$\vec{J} = \frac{\mathbb{1}}{2} \mathcal{V} \int_0^\infty dx f(x^2) x^2. \quad \text{unpolarized light} \quad (24.6)$$

In this expression, $\mathbb{1}$ is the identity tensor in the 2D space of transverse directions and \mathcal{V} is the volume of the unit sphere in four dimensions. For our purposes, the main point is that \vec{J} is proportional to the unit matrix. For example, \vec{J}_{12} is zero by the invariance of Equation 24.5 under reflections in y . Also, symmetry under exchange of x and y gives $\vec{J}_{11} = \vec{J}_{22}$. 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 \vec{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 \vec{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 \vec{J} 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 \vec{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 $\vec{E} = A\hat{x} + Be^{i\delta}\hat{y}$ into the definition of \vec{J} and find the Stokes parameters⁴ in terms of A , B , and δ . 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 a sphere of radius s_0 (see Equation 24.4).

Your Turn 24B

Comment on what parts of the Poincaré sphere 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{J} , 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

$$\vec{J}_{11} = \langle A \cos \theta A \cos \theta \rangle = \frac{1}{2}A^2, \quad (24.7)$$

$$\vec{J}_{12} = \langle A \cos \theta A \sin \theta \rangle = 0, \quad (24.8)$$

and so on. Thus,

$$\vec{J} = \frac{A^2}{2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad (24.9)$$

so this ensemble indeed serves as one realization of unpolarized light. Although not as complete as Equation 24.5 (we omitted circular and elliptically polarized states), this realization 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{J} discards any overall phase, so we don't need to give A and B separate phases.

⁵Nor 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 a polarizing filter performs a *linear projection* on the electric field, that is, the linear operation $\vec{E} \rightarrow \hat{\zeta}(\hat{\zeta}^* \cdot \vec{E})$. Then the corresponding transformation on the polarization tensor \vec{J} is

$$\vec{J} \rightarrow \langle \hat{\zeta}(\hat{\zeta}^* \cdot \vec{E})(\vec{E}^* \cdot \hat{\zeta}) \rangle = (\hat{\zeta} \otimes \hat{\zeta}^*) \cdot \vec{J} \cdot (\hat{\zeta} \otimes \hat{\zeta}^*).$$

Your Turn 24C

- Consider the case of a linear polarizer, that is, $\hat{\zeta} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, acting on unpolarized light. Interpret the new polarization tensor.
- Repeat for a circular polarizer.

Think about how applying various filters to an *arbitrary* \vec{J} , then finding the intensity of the filtered light, lets us deduce the various matrix elements of \vec{J} , and hence the Stokes parameters.

FURTHER READING

Intermediate:

Zangwill, 2013, §16.4; Landau & Lifshitz, 1979, §50; Born & Wolf, 1999, chap. 10; Wolf, 2007, chap. 8.

Technical:

Thompson et al., 2017.

PROBLEMS

24.1 *Stokes*

For a given direction of wave propagation, Equation 24.2 defines a complex 2-tensor $\vec{J}_{ij} = \langle \vec{E}_i \vec{E}_j^* \rangle$ to describe the polarization state of a superposition of plane waves. A special case is a pure (fully-polarized) plane wave, $\vec{E} = \frac{1}{2} \vec{E} e^{-i\omega(t-z/c)} + \text{c.c.}$ Equation 24.3 then repackaged the information in \vec{J} as four real quantities s_α .

- Suppose that we have fully polarized light traveling along the z axis, with $s_0, \dots, s_3 = 3, -1, 2, -2$ (times an overall constant). Find a formula for $\vec{E}(t)$ at the origin of coordinates $\vec{r} = \vec{0}$. Confirm that the tip of the electric field vector sweeps out an ellipse in 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 268) 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):

$$\nabla^2 \psi = -\rho_q / \epsilon_0 \quad (25.1)$$

$$\nabla^2 \vec{A} - c^{-2} \left(\frac{\partial^2}{\partial t^2} \vec{A} + \vec{\nabla} \frac{\partial}{\partial t} \psi \right) = -\mu_0 \vec{j}. \quad (25.2)$$

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**:¹

$$\nabla^2 \vec{A} - c^{-2} \frac{\partial^2}{\partial t^2} \vec{A} = -\mu_0 \vec{j}. \quad \text{Coulomb gauge, no net charge} \quad (25.3)$$

In empty space, we found some simple solutions to this equation: the plane waves. But certainly empty space may instead contain no radiation (fields everywhere zero). We'll now see how, in the presence of accelerating charges, 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.

¹Sometimes 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:²

$$\vec{A}(\vec{r}) = \frac{\mu_0}{4\pi} \int d^3r_* \frac{\vec{j}(\vec{r}_*)}{\|\vec{r} - \vec{r}_*\|}. \quad \text{static case} \quad [15.18, \text{page 221}]$$

Chapter 15 called this expression the Green function solution to the Poisson equation. As usual, call \vec{r} the “field point” and \vec{r}_* the “source point.” Also define $\vec{R} = \vec{r} - \vec{r}_*$, and denote its length by R (no arrow). Then the function $G(\vec{r}, \vec{r}_*) = (4\pi R)^{-1}$ is called 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 GUESS FOR THE RADIATION GREEN FUNCTION

We might expect that the fields at a spatial position \vec{r} would again be determined by currents at \vec{r}_* , with a $1/R$ falloff. But we also expect that signals will travel from source point to field point at the finite speed c . So a simple guess for the generalization of Equation 15.18 is that each component of \vec{A} is given by

$$\vec{A}(t, \vec{r}) \stackrel{?}{=} \frac{\mu_0}{4\pi} \int d^3r_* \frac{1}{R} \vec{j}(t - R/c, \vec{r}_*). \quad \text{trial solution, Coulomb gauge} \quad (25.4)$$

In words, we are again proposing that the vector potential at time t gets contributions from each source point. 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, our guess 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

Before proceeding, verify that the proposed trial solution Equation 25.4 really obeys the Coulomb gauge condition $\vec{\nabla} \cdot \vec{A} = 0$. [*Hint*: Adapt the approach used in magnetostatics (Section 15.5.4, page 221).]

The form of our trial solution suggests part of the answer to Hanging Question #H (page 31): 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

²See Equation 2.6 (page 30).

our detector. *Once formed, radiation proceeds autonomously* through space. It reflects only the behavior of currents at the **retarded time**³ $t - R/c$.

25.4 CHECK THE GUESS

We now apply the **wave operator** $\square = \nabla^2 - c^{-2}\partial^2/\partial t^2$ to our trial solution, to see whether we indeed recover $-\mu_0\vec{j}$ (Equation 25.3).⁴ The wave operator involves derivatives with respect to the field point, so throughout this section $\vec{\nabla}_i$ will denote $\partial/\partial\vec{r}_i$ (not $\partial/\partial\vec{r}_{*i}$).

Your Turn 25B

Show that (or review why)

$$\vec{\nabla}R = \hat{R}; \quad \vec{\nabla} \cdot \vec{R} = 3; \quad \vec{\nabla}(R^{-p}) = -pR^{-(p+1)}\hat{R}; \quad \nabla^2(R^{-1}) = -4\pi\delta^{(3)}(\vec{R}).$$

To save writing, let ϕ denote any component of $4\pi\vec{A}/\mu_0$, and \mathcal{J} the corresponding component of \vec{j} . So our proposed Green function solution, Equation 25.4, says

$$\phi(t, \vec{r}) \stackrel{?}{=} \int d^3r_* \frac{1}{R} \mathcal{J}(t - R/c, \vec{r}_*), \quad (25.5)$$

and we wish to show

$$\nabla^2\phi - c^{-2}\frac{\partial^2}{\partial t^2}\phi \stackrel{?}{=} -4\pi\mathcal{J}. \quad (25.6)$$

The gradient of Equation 25.5 is

$$\vec{\nabla}\phi = \int d^3r_* \left[(\vec{\nabla}(R^{-1}))\mathcal{J}(t - R/c, \vec{r}_*) - \frac{1}{cR}(\vec{\nabla}R)\frac{\partial\mathcal{J}}{\partial t}\Big|_{\text{ret}} \right].$$

Here the subscript “ret” means to evaluate at the retarded time $t - R/c$ (after taking any indicated derivatives).

Taking a second derivative gives

$$\begin{aligned} \nabla^2\phi(t, \vec{r}) &= \int d^3r_* \left[(\nabla^2 R^{-1})\mathcal{J}(t - R/c, \vec{r}_*) - c^{-1}(\vec{\nabla}R^{-1}) \cdot (\vec{\nabla}R)\frac{\partial\mathcal{J}}{\partial t}\Big|_{\text{ret}} - c^{-1}\vec{\nabla} \cdot (R^{-1}\hat{R}\frac{\partial\mathcal{J}}{\partial t}\Big|_{\text{ret}}) \right] \\ &= \int d^3r_* \left[-4\pi\delta^{(3)}(\vec{R})\mathcal{J}(t - R/c, \vec{r}_*) - c^{-1}(-R^{-2}\hat{R}) \cdot \hat{R}\frac{\partial\mathcal{J}}{\partial t}\Big|_{\text{ret}} - c^{-1}\vec{\nabla} \cdot (R^{-2}\vec{R}\frac{\partial\mathcal{J}}{\partial t}\Big|_{\text{ret}}) \right]. \end{aligned}$$

The three delta functions eliminate the integral over \vec{r}_* and set $\vec{r}_* = \vec{r}$, so continuing,

$$\begin{aligned} &= -4\pi\mathcal{J}(t, \vec{r}) + \underbrace{\int d^3r_* \left[(cR^2)^{-1}\frac{\partial\mathcal{J}}{\partial t}\Big|_{\text{ret}} + c^{-1}2R^{-3}\hat{R} \cdot \vec{R}\frac{\partial\mathcal{J}}{\partial t}\Big|_{\text{ret}} - (cR^2)^{-1}3\frac{\partial\mathcal{J}}{\partial t}\Big|_{\text{ret}} \right]}_{=0} \\ &\quad + (cR)^{-2}\vec{R} \cdot (\vec{\nabla}R)\frac{\partial^2\mathcal{J}}{\partial t^2}\Big|_{\text{ret}}. \end{aligned}$$

³This traditional terminology may cause confusion; note that $t - R/c$ is always *earlier* than the observation time t .

⁴ \square is sometimes called the d’Alembert operator, or “d’Alembertian.”

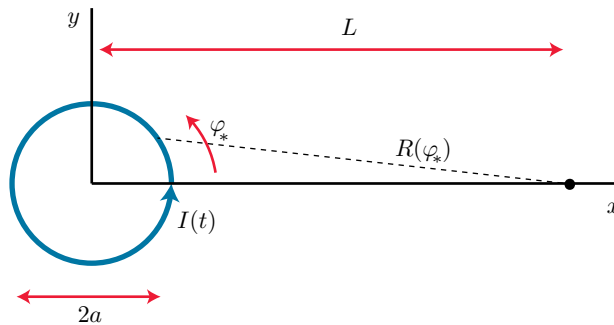


Figure 25.1: **Magnetic dipole antenna.** An arrow indicates the convention that positive current I means net flow in the counterclockwise direction.

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 \mathcal{J} . Reinstating the vector character of \vec{A} and multiplying by $\mu_0/(4\pi)$ proves 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).⁵ As usual, assume that the charge flux is zero everywhere except inside the wire. In the wire, assume a sinusoidal current $I(t) = \bar{I} \cos(\omega t)$ independent of position φ_* . That is, our antenna is an *oscillating magnetic dipole*. Charge never piles up anywhere, so $\rho_q = 0$, and we may use the Green function solution developed in Sections 25.3–25.4.

[T₂] Section 25.5.1' (page 337) 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 distance⁶ like $1/r^3$. Now we want to explore what changes when the current alternates.

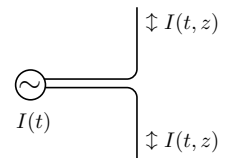


Fig. 43.2 (page 566)

⁵This example appears to be due to G. FitzGerald in 1883. FitzGerald also derived the ω^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 struck out the “im” afterward.

⁶See Your Turn 17A (page 245).

Imagine sitting somewhere far away along the $+x$ axis,⁷ 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 φ_* , which runs from zero (closest point to our observer) to 2π (same point). At any point on the loop, the current points in the azimuthal direction $\pm\hat{\varphi}$. So Equation 25.4 gives⁸

$$\vec{A}(t, \vec{r}) = \frac{\mu_0 \bar{I}}{4\pi} \int_0^{2\pi} (a d\varphi_*) R^{-1} \left[\frac{1}{2} e^{-i\omega(t-R/c)} \hat{\varphi} + \text{c.c.} \right]. \quad (25.7)$$

In this formula, $R = \sqrt{(L - a \cos \varphi_*)^2 + a^2 \sin^2 \varphi_*}$, and $\hat{\varphi}$ is the unit tangent vector to the loop at angular position φ_* .

Our answer can be simplified a lot if we agree to study only the leading-order behavior in powers of a/L (the **small source regime**⁹). Thus, $R^{-1} = L^{-1} + \dots$, where the ellipsis contains only terms that we have already agreed to drop. This factor is independent of φ_* , so it can be moved outside the integral, along with the time dependence:

$$\vec{A} = \frac{\mu_0 \bar{I}}{4\pi} \frac{a}{L} \frac{1}{2} e^{-i\omega t} \int_0^{2\pi} d\varphi_* (-\hat{x} \sin \varphi_* + \hat{y} \cos \varphi_*) \exp \left[i \frac{\omega}{c} L \left(1 - \frac{a}{L} \cos \varphi_* + \dots \right) \right] + \text{c.c.} \quad (25.8)$$

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 $L \rightarrow \infty$. (That is why Equation 25.8 abbreviated them by an ellipsis.)

We therefore find the fields in the small source regime to be

$$\vec{A} \rightarrow \frac{\mu_0 \bar{I}}{4\pi} \frac{a}{L} \frac{1}{2} e^{-i\omega(t-L/c)} \int_0^{2\pi} d\varphi_* (-\hat{x} \sin \varphi_* + \hat{y} \cos \varphi_*) \exp[-i(\omega a/c) \cos \varphi_*] + \text{c.c.} \quad (25.9)$$

The term that points along \hat{x} integrates to zero by a symmetry argument: It is an odd function of φ_* , 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 does not vanish.

⁷By 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 another direction.

⁸Substitute the charge flux in the thin-wire approximation (Equation 15.22, page 223) into Equation 25.4.

⁹In statics, Chapters 3 and 17 called this limit "far field," but in dynamics we reserve that term for a stricter condition. Chapter 43 will explore this and other regimes systematically.

Your Turn 25D

- Suppose that ω is small but nonzero; use a Taylor expansion of the exponential to get an approximate answer for the integral.
- [*Optional:* If you know about stationary phase approximation, use it to get an answer in the opposite limiting case, where ω is large.]
- Ask Wolfram Alpha or some other analytic math assistant about `Integrate[Cos[t]*E^(-I*p*Cos[t]),{t,-Pi,Pi}]`. Graph the answer, look at the limits for large and small p , and compare (a,b).

Because we used restricted Coulomb gauge, the scalar potential is zero. Thus, the electric field is simply $-\frac{\partial}{\partial t}\vec{A}$. The time derivative just introduces a factor of $(-i\omega)$, so

$$\vec{E} \rightarrow (\text{const})\frac{1}{L}e^{-i\omega(t-L/c)}\hat{y} + \text{c.c.} \quad \text{along } x \text{ axis as } L \rightarrow \infty.$$

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 even a static electric dipole.

We also get a prediction that the outgoing wave observed at this point 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} (via 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 $L \rightarrow \infty$ is

$$\vec{B} \rightarrow (\text{const}')\frac{1}{L}e^{-i\omega(t-L/c)}\hat{z} + \text{c.c.},$$

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. In fact, it points along the direction of the magnetic dipole moment whose oscillation gave rise to the wave.

Together, the \vec{E} and \vec{B} far fields form an approximately plane wave moving toward the observer. For a distant observer anywhere in the xy plane, the magnetic field points along $\pm\hat{z}$.

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 $\pi a^2 \bar{I}$), and to the frequency squared.

Media 9 shows the streamlines of \vec{B} (also called magnetic field lines).

Alternating current in a circular loop gives far fields that carry energy all the way to infinity.

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 $\propto L^2 L^{-2}$ — 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.

25.5.4 The loop antenna is directional

See Problem 25.1. 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.¹⁰ We have also found an unexpected Electromagnetic Phenomenon (waves), including specific details (about polarization). 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. Also, so far we have only examined the fields in the xy plane. 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 a deeper understanding, Parts IV and V will uncover an important aspect of Maxwell's 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.

[T2] Antenna theory: Smith, 1997, chap. 8.

¹⁰That's the one involving electrical technology. The next revolution (semiconductors) required quantum mechanics.

T_2

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.

PROBLEMS

25.1 Directionality of antenna

A circular loop antenna emits energy with a specific directional pattern.

A circular loop of wire, carrying an oscillating current, lies in the xy plane (Figure 25.2). Equation 25.7 gives the vector potentials (ψ is zero). In this formula, \vec{r} is the position of the field observation. The angle φ_* 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 φ_* . The distance $R(\varphi_*) = \|\vec{r} - \vec{r}_*\|$. The current in the loop is everywhere $I(t) = \bar{I} \cos \omega t$.

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 335). Can you make a statement that covers both 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 333) 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 $\mathcal{O}(L^{-2})$ 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

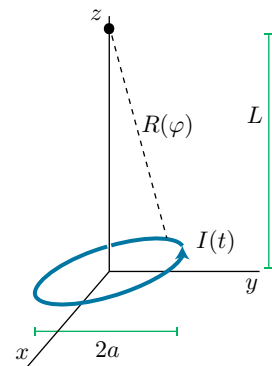


Figure 25.2: A current loop in the xy plane. See Problem 25.1.

was that unlike electric field lines, magnetic field lines must be closed loops (no ends).¹¹ 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. If this detachment really occurs, we’d like to know.

Do: Again consider a circular loop of wire of radius a in the xy plane, carrying a prescribed, harmonically oscillating current $I(t) = \frac{1}{2}[\bar{I}e^{-i\omega t} + \text{c.c.}]$ (see Figure 25.1, page 333). Thus, \bar{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 “merely” involves numerically evaluating a formula obtained in Section 25.5.2:

$$\vec{A}(t, \vec{r}) = \frac{\mu_0 \bar{I}}{4\pi} \int_0^{2\pi} (a d\varphi_*) R^{-1} \left[\frac{1}{2} e^{-i\omega(t-R/c)} \hat{\varphi} + \text{c.c.} \right]. \quad (25.10)$$

In this expression,

- $R = \|\vec{r} - \vec{r}_*(\varphi_*)\|$, where \vec{r}_* is the point on the loop at angular position φ_* , and
- $\hat{\varphi}$ is the unit tangent vector to the loop at angular position φ_* .

In the following, you’ll work out the curl of Equation 25.10, and *then* evaluate it numerically. Actually, it’s enough to find $\vec{B}(t, \vec{r})$ only 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.

Steps:

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 $a/(a/c)$).

- 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; eventually, you’ll evaluate it numerically, but not yet. 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.”

- 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 up¹² the well-known static magnetic dipole field, compare, and if necessary reconcile your result.
- 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

¹¹Chapter 36 will obtain this from $\vec{\nabla} \cdot \vec{B} = 0$.

¹²Or use your result in Your Turn 17A (page 245).

$-5a < z < +5a$. 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 $(x_{(\ell)}, 0)$, that is, points along the x axis at an evenly spaced series of values $\{x_{(\ell)}\}$. (Just make sure none of your choices is $(1, 0)$, because the fields are singular exactly on the wire.)

Note that overall factors like $\mu_0 \bar{I}/(4\pi)$ aren't needed when we want only the streamlines. Your computer 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 initial points appropriately. At time t , ask your software for streamlines that pass through $(x_{(\ell)} + ct, 0)$, where $\{x_{(\ell)}\}$ are the same points you used in (b).]

Show some initiative. Suppose 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 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 Emergence of transversality

First work Problem 25.3. Now compute the longitudinal part of \hat{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.]

¹³See Problem 3.10. For example, Python has a function `plt.streamplot` that accomplishes this. Problem 15.7 discussed the fields created by a stationary current loop of finite size.

¹⁴You 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 ω 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.]

CHAPTER 26

Galilean Relativity

False views, if supported by some evidence, do little harm, as every one 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 wish to set up a 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 butterflies 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 of 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 overthrow 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.

T2 Section 26.1’ (page 354) 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 (or moving in 1D along a frictionless air track in the lab). Newtonian mechanics says that their motions are solutions to the equations

$$\frac{d^2x_{(1)}}{dt^2} = -\frac{k}{m}(x_{(1)} - (x_{(2)} - L)) \quad \frac{d^2x_{(2)}}{dt^2} = -\frac{k}{m}((x_{(2)} - L) - x_{(1)}). \quad (26.1)$$

Although these are familiar equations, let's unpack their content a bit.²

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.³ 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 26.1 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 26.1.

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 Maxwell's 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 our equations:

$$x_{(1)}(t) = C \cos(\omega t) \quad x_{(2)}(t) = L - C \cos(\omega t). \quad (26.2)$$

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 $x_{(2)}$:

$$\tilde{x}_{(1)}(t) = C \cos(\omega t) + A \quad \tilde{x}_{(2)}(t) = L - C \cos(\omega t) + A. \quad (26.3)$$

Such transformations are called **active**, because the new solution is a *physically different motion* from the original. The operation in Equation 26.3 transforms any solution of

²See also page 21.

³Some books use the term "world-line" for this concept.

the equations of motion into *another* solution (and nonsolutions to nonsolutions). We will call such operations **symmetries** of the dynamics.

That is, a symmetry is an operation that can be applied to any trajectory, and that *permutes the solutions* of a set of equations of motion. In addition to the overall translation described by Equation 26.3, 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:

$$x' = x - A \quad t' = t - B. \quad (26.4)$$

Because we are not physically changing the trajectory, this transformation is called **passive**: it just changes the *representation* of a trajectory. Equation 26.3 shifted any trajectory to the right by A , whereas Equation 26.4 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 $d/dt = d/dt'$. 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 26.1), including the numerical values of constants (k , L , and m). We say that the original equations of motion have an **invariance** under the passive transformation Equation 26.4.

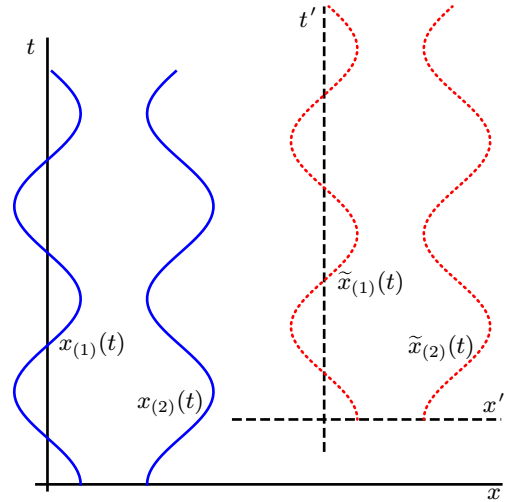
26.4.2 Relation between active and passive

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**.*
- (26.5)

Figure 26.1: Active versus passive. The initial trajectory (Equation 26.2, *solid blue curves*) appears different in the original (*solid black*) and shifted (*dashed black*) coordinate systems. The corresponding actively transformed trajectory (Equation 26.6, *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.

We can illustrate the idea with the solution in Equation 26.2, applying the recipe in Idea 26.5 with the transformation Equation 26.4:

$$\tilde{x}'_{(1)} = C \cos(\omega t') \quad \tilde{x}'_{(2)} = L - C \cos(\omega t').$$

In terms of the original coordinates, we then have

$$\tilde{x}_{(1)} = C \cos(\omega(t - B)) + A \quad \tilde{x}_{(2)} = L - C \cos(\omega(t - B)) + A, \quad (26.6)$$

which is indeed the formula in Equation 26.3, generalized to include time translation. Figure 26.1 illustrates this procedure.

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 $\begin{bmatrix} x \\ y \end{bmatrix}$. Then the *same* point viewed from a *rotated* point of view will be labeled by two different numbers $\begin{bmatrix} x' \\ y' \end{bmatrix}$. We can find the new coordinates by using trigonometry, and the fact that the new coordinate axes are rotated by some angle α relative to the old ones. There's a simple formula expressing this:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos \alpha & \sin \alpha \\ -\sin \alpha & \cos \alpha \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \quad (26.7)$$

To think about this conceptually, imagine digging up all the streets in Manhattan and laying down a new grid of streets rotated counterclockwise relative to the old one by α . 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:⁴ $d = \sqrt{(x'_1 - x'_2)^2 + (y'_1 - y'_2)^2}$. Generic coordinate transformations don't have this property. For example, if we define new coordinates via a **dilation** transform, $\vec{r}' = 2\vec{r}$, the form of the distance function is not quite the same. In short,

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:

Your Turn 26A

- Show that the 3D version of Equation 26.1, 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 $L = 0$).
- Even if $L = 0$, show that a *nonlinear* spring will also spoil dilation invariance.
- Use similar reasoning to establish the rotation invariance of two masses bound by *gravitational* force, and their lack of dilation invariance.
- One may imagine trying to rescue the situation by also allowing dilations in *time*: $\vec{r}' = 2\vec{r}$ and $t' = \zeta t$. Show that in a world with *both* newtonian gravitation and springs, this gambit does not succeed regardless what we choose for ζ .

Newtonian physics does not have any general invariance under dilations.

26.6 GALILEAN GROUP

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

⁴See Section 14.2.

already saw one example (dilation). Similarly, most time-dependent transformations, such as $\vec{r}' = \vec{r} + \vec{a}t^2/2$, introduce new “fictitious forces.”⁵ 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 **G-inertial**, in honor of Galileo. Translations like Equation 26.4 and rotations like Equation 26.7, 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).⁶ 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 **rest frame** to mean “G-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.

[T2] Section 26.6.1' (page 354) extends the discussion to include parity and time reversal invariance.

26.6.2 Boosts connect coordinate systems in relative motion

Returning to one dimension, there’s another important class of symmetry transformations, called **galilean boosts**.

Your Turn 26B

- a. Show that the passive coordinate transformations:

$$x' = x - v_*t, \quad t' = t \quad (26.8)$$

are invariances of the equations of motion. That is, show that re-expressing Equation 26.1 in terms of the new variables yields equations of identical form.

- b. Find the corresponding active transformation of the trajectory Equation 26.2 (see Figure 26.2), and confirm that it does solve Newton’s law (Equation 26.1).

⁵The “Coriolis force” is another example.

⁶Some books restrict these terms to refer only to *inertial* (“good”) coordinate systems, but others don’t.

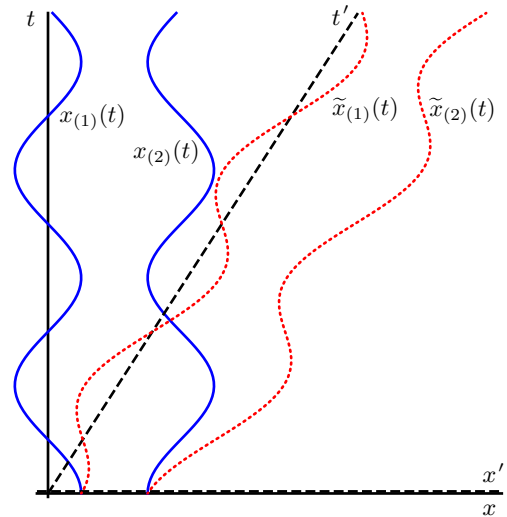


Figure 26.2: Galilean boost. The initial trajectory (Equation 26.2, *solid blue curves*) appears different in the original (*solid black*) and boosted (*dashed black*) coordinate systems. The corresponding actively transformed trajectory (Your Turn 26B, *dotted red*) appears the same in the boosted coordinates 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.

Equation 26.8 describes a new coordinate system, whose axes are moving to the right at speed v_* relative to the original. The minus sign indicates that these moving axes 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 26.8 in matrix form:

$$\begin{bmatrix} t' \\ x' \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -v_* & 1 \end{bmatrix} \begin{bmatrix} t \\ x \end{bmatrix}. \quad \text{galilean boost} \quad (26.9)$$

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 by $v_{*(1)} + v_{*(2)}$, that is, the matrix $\begin{bmatrix} 1 & 0 \\ -(v_{*(1)}+v_{*(2)}) & 1 \end{bmatrix}$.

That **galilean velocity addition formula** 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 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,

$$x'' = x' - v_{*(1)}t' - A_1 = (x - v_{*(2)}t - A_2) - v_{*(1)}t - A_1, \quad t'' = t' = t. \quad (26.10)$$

We see that the overall effect is again the combination of a boost (with speed $v_{*(1)} + v_{*(2)}$ as you found in Your Turn 26C) 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).⁷

Your Turn 26D

- Generalize Equation 26.10 to include time shifts also.
- Show that if we apply any two of these transformations in succession, the result is a single transformation that is also in this family.
- Show that any such transformation has an inverse, which is again in the family.

Mathematicians call a set of transformations with those properties a **group**, hence the name “galilean group.”

26.6.5 The physical significance of invariance

By now, certain questions may be bothering you:

- Why are we spending so much time with balls on springs? Even within newtonian physics, that’s a specialized, and idealized, system.
- 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

All of newtonian physics has the overarching mathematical property of galilean invariance that transcends details of particular springs, clocks, planets, and so on. (26.11)

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.

⁷The galilean group also contains some discrete transformations (spatial and temporal inversion). However, in a dissipative system, the temporal inversions are not invariances.

Idea 26.11 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. (26.12)

In a more lapidary phrase:

Physicists study invariance because it strips away details and lays bare the structural essentials of a dynamical theory. (26.13)

We can now see why Idea 26.11 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:

- *There's an inertial coordinate system where I know what's going on.*
- *But I want to know what's going on in some other inertial system (perhaps one that I set up in my lab).*
- *So I can use the appropriate transformation to go from the first to the second.*

**Relativity
Strategy**

(26.14)

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 26.11–26.14 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 289) 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 $v_{*(1)}$. Imagine mounting that catapult on a train car, bringing it up to speed $v_{*(2)}$ 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 $(v_{*(1)} + v_{*(2)})\hat{x}$.

Let's obtain the result just stated as a consequence of galilean invariance, using Idea 26.14. 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 $v_{*(1)}$.

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 $v_{*(1)}$ with respect to the primed system, in which the catapult is at rest. We also know that the primed system is moving at $v_{*(2)}$ with respect to the ground. Your Turn 26C then says that the doubly primed system is moving uniformly at $v_{*(1)} + v_{*(2)}$ with respect to the ground.

Alternatively, we can write the trajectory of a projectile fired from a catapult *at rest* as the parametric curve $\begin{bmatrix} t \\ x \end{bmatrix} = \begin{bmatrix} \xi \\ v_{*(1)}\xi \end{bmatrix}$, and that of the stationary catapult as $\begin{bmatrix} t \\ x \end{bmatrix} = \begin{bmatrix} \xi \\ 0 \end{bmatrix}$. Now apply an active boost to conclude that there must be another solution, in which the projectile's trajectory is $\begin{bmatrix} t' \\ x' \end{bmatrix} = \begin{bmatrix} \xi \\ v_{*(1)}\xi \end{bmatrix}$ and the catapult's is $\begin{bmatrix} t' \\ x' \end{bmatrix} = \begin{bmatrix} \xi \\ 0 \end{bmatrix}$. Re-expressing these new trajectories in the unprimed coordinate system yields

$$\text{catapult: } \begin{bmatrix} t \\ x \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ v_{*(2)} & 1 \end{bmatrix} \begin{bmatrix} \xi \\ 0 \end{bmatrix} \quad \text{projectile: } \begin{bmatrix} t \\ x \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ v_{*(2)} & 1 \end{bmatrix} \begin{bmatrix} \xi \\ v_{*(1)}\xi \end{bmatrix}.$$

The velocity of the catapult is $\Delta x/\Delta t = v_{*(2)}$ as desired, and that of the projectile is $\Delta x/\Delta t = v_{*(1)} + v_{*(2)}$.

W. deSitter 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)⁻¹ 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

Light from distant objects arrives at Earth with a delay related to distance, but not velocity, of the source.

observed, so light can't be a newtonian particle.⁸

26.7 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 described above.

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 $3 \cdot 10^8$ 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.

⁸More 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 $\approx 2c$ (Alvåger et al., 1964).

T₂

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.

In principle, there must be analogous *gravitational* background radiation, which cannot be so screened, plus relic neutrinos and so on; your system may in principle also emit gravitational radiation. 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 of any laboratory system.

T₂

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 ε_{ijk} and other quantities derived from it would all change sign, a "passive parity (or inversion) transformation." Interestingly, however, the equations of newtonian physics all set two true tensors equal (for example, $m d^2 \vec{r} / dt^2 = -\vec{\nabla} U$), or else they set two pseudotensors equal (for example, $d\vec{L}/dt = \vec{\tau}$). Indeed, we have noted earlier that "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 pseudotensors equal (for example, Faraday's law). Both charge and mass are true scalars ("even under parity").

There is a corresponding active viewpoint: If we replace every particle trajectory $\vec{\Gamma}(t)$ by $-\vec{\Gamma}(t)$ 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 the 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 even or odd parity (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} .

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": $\vec{\Gamma}(t) = \vec{\Gamma}(-t)$. In a nondissipative system (no friction nor diffusion), the new trajectory will solve its 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,

$\vec{E}(t, \vec{r}) = -\vec{E}(-t, \vec{r})$), then find that the Maxwell equations all either set even = even (for example, the electric Gauss law or Lorentz force law) or else set odd = odd (for example, the Faraday law). Both charge and mass are taken to be even under time reversal.

Time reversal is not a symmetry of thermal systems, where the increase of entropy sets an “arrow of time.”

PROBLEMS

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 course on electrodynamics, but more generally it's a course about *where theories come from*. It's good to see abstract things first in a familiar setting.

In the preceding chapter, we started with a vague Principle (of Relativity), but then it turned into precise algebra and calculus (an invariance property). That's a very 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 that we can apply to field theories, including eventually relativistic ones including electrodynamics. Indeed, historically physicists' obsession with symmetry began with electrodynamics.

We'll also extend a framework relevant to other themes of this course, 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 relation, 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 $\rho_m^{(1D)}$ ($\sim \text{kg/m}$). 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 with equilibrium separation Δx , each with mass $\Delta m = \rho_m^{(1D)} \Delta x$ and spring constant $\kappa/\Delta x$. Here κ is a material parameter describing the spring (the stretch modulus, with units of force). Then we think of the spring as a chain of point masses Δm 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

$$f(0) = \frac{\kappa}{\Delta x} (u(-\Delta x) - 2u(0) + u(\Delta x)).$$

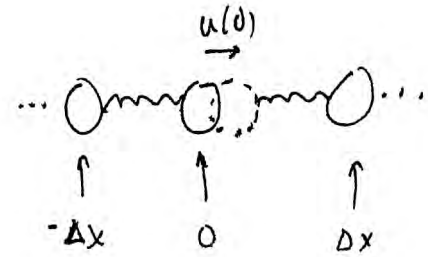


Figure 27.1: Spring modeled as discrete elements.
The mass initially at $x = 0$ has been displaced from equilibrium to the *dashed* position.

For small Δx , Newton's law then becomes

$$\Delta m \frac{\partial^2 u}{\partial t^2} \Big|_{x=0} = \frac{\kappa}{\Delta x} \frac{\partial^2 u}{\partial x^2} \Big|_{x=0} (\Delta x)^2 + \dots, \quad (27.1)$$

where the ellipsis denotes terms that are higher order in Δx . We now take the continuum limit $\Delta x \rightarrow 0$, 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**

$$\frac{\partial^2 u}{\partial t^2} - c_s^2 \frac{\partial^2 u}{\partial x^2} = 0 \quad \text{where } c_s^2 = \kappa / \rho_m^{(1D)}. \quad (27.2)$$

Because this is a partial differential equation, we call it a field theory in one space and one time dimension. (Chapter 18 showed that Maxwell's 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 ω 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 344) and passive (Section 26.4) viewpoints. We consider relabeling events via a Galilean boost:

$$x' = x - v_* t, \quad t' = t. \quad [26.8, \text{page 348}]$$

Active viewpoint

An active transformation replaces a spring configuration u by a different one, \tilde{u} . Following the Idea 26.5 (page 345), 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

$$\tilde{u}_{\pm} = \bar{u} \cos(\omega(-t' \pm x'/c_s)).$$

In the original coordinates, then

$$= \bar{u} \cos(\omega(-t \pm (-v_*t + x)/c_s)).$$

Now manipulate a little to bring this expression closer to the same overall form as before:

Your Turn 27A

a. Obtain

$$\tilde{u}_{\pm}(t, x) = \bar{u} \cos\left(\omega(1 \pm (v_*/c_s))\left(-t \pm \frac{x}{c_s \pm v_*}\right)\right). \quad (27.4)$$

b. Show that \tilde{u}_{\pm} is a traveling wave moving at speed $\pm c_s + v_*$.

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 $\pm c_s$! That may surprise you, so before discussing it let's first rederive it to be sure.

Passive viewpoint

Let's rederive the preceding result by focusing on the wave equation itself, rephrasing it in terms of the new variables

$$x' = x - v_*t, \quad t' = t. \quad [26.8, \text{page 348}]$$

Thus, we define

$$u'(t', x') = u(t, x + v_*t).$$

The change of variables formula from vector calculus lets us rephrase the equation of motion in the new variables:

$$\left[\left(\frac{\partial x'}{\partial x} \frac{\partial}{\partial x'} \right)^2 - c_s^{-2} \left(\frac{\partial x'}{\partial t} \frac{\partial}{\partial x'} + \frac{\partial t'}{\partial t} \frac{\partial}{\partial t'} \right)^2 \right] u' = 0.$$

Simplifying yields

$$\left[\frac{\partial^2}{\partial x'^2} - c_s^{-2} \left(-v_* \frac{\partial}{\partial x'} + \frac{\partial}{\partial t'} \right)^2 \right] u' = 0.$$

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.

27.4 INVARIANCE REGAINED

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 x_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 Δt , the red segment has moved to $x = x_0 + v_m \Delta t$. Hence, its transverse velocity $v_y(t_0, x_0)$ is the limit of

$$v_y(t_0, x_0) = \frac{1}{\Delta t} [u(t_0 + \Delta t, x_0 + v_m \Delta t) - u(t_0, x_0)] = \left(\frac{\partial}{\partial t} + v_m \frac{\partial}{\partial x} \right) u \Big|_{t_0, x_0}.$$

The net transverse force on this segment is still $F_0 \Delta x (\partial^2 u / \partial x^2)$ as before, so during time Δt its transverse momentum p_y changes by $F_0 \Delta x (\partial^2 u / \partial x^2) \Delta t$. That is,

$$p_y(t_0 + \Delta t, x_0 + v_m \Delta t) - p_y(t_0, x_0) = F_0 \Delta x (\partial^2 u / \partial x^2) \Delta t, \quad \text{or} \\ \left[\left(\frac{\partial}{\partial t} + v_m \frac{\partial}{\partial x} \right)^2 - c_s^2 \frac{\partial^2}{\partial x^2} \right] u = 0. \quad (27.5)$$

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.

Your Turn 27B

- a. Substitute a generic traveling wave as a trial solution into Equation 27.5 and show that it only works if the wave moves at speeds $v_m \pm c_s$. (Indeed, 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.)
- 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 (treating 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 Maxwell's equations imply the wave equation, and Section 27.3 showed that the wave equation lacks galilean invariance. Everyone already knew this prior to 1905. Everyone assumed that the cure would be along the lines described in Section 27.4: "Maxwell's 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 RELATIONS 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 one we found for *charge* (Section 8.3, page 113). First note that

$$\text{KE} = \int dx \frac{1}{2} \rho_m^{(1D)} (\partial u / \partial t)^2; \quad \text{PE} = \text{const} + F_0 \int dx \frac{1}{2} (\partial u / \partial x)^2. \quad (27.6)$$

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 $L_{\text{tot}} - \int_0^{L_{\text{tot}}} (dx / \cos \theta(x))$, where θ is the angle relative to straight.¹ 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 (J/m) at t, x is

$$\rho_{\mathcal{E}}^{(1D)}(t, x) = \frac{1}{2} \rho_m^{(1D)} ((\partial u / \partial t)^2 + c_s^2 (\partial u / \partial x)^2).$$

If you pluck just one mass, you'll create some localized potential energy, 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 $j_{\mathcal{E}}^{(1D)}$ (units J/s). To find that flux, note that the rate at which energy gets transported from any mass far away to the one at its right is the rate at which work is done on the right side by the left side. This is the product of velocity (which is transverse) times the transverse component of force, so $j_{\mathcal{E}}^{(1D)} = -F_0 (\partial u / \partial x) (\partial u / \partial t)$.

Your Turn 27C

Use Equation 27.2 to show that for any solution of the wave equation,

$$\frac{\partial \rho_{\mathcal{E}}^{(1D)}}{\partial t} + \frac{\partial j_{\mathcal{E}}^{(1D)}}{\partial x} = 0. \quad \text{continuity equation for energy, newtonian spring}$$

Similarly to the continuity equation for charge (Chapter 8), your result expresses the fact that energy is a locally conserved quantity: 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 114).

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.

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 are both locally conserved

For the solutions given in Equation 27.3, the energy density is

$$\rho_{\mathcal{E}}^{(1D)} = \frac{1}{2} \rho_m^{(1D)} \bar{u}^2 (\omega^2 + c_s^2 (\omega/c_s)^2) \sin^2(\omega t - (\omega/c_s)x). \quad (27.7)$$

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

$$j_{\mathcal{E}}^{(1D)} = -\rho_m^{(1D)} c_s^2 \left(-\frac{\omega}{c_s}\right) \omega \bar{u}^2 \sin^2(\omega t - (\omega/c_s)x)$$

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 relation. 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 relations. 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 $\rho_m^{(1D)}$ ($\sim \text{kg/m}$) and stretch modulus κ ($\sim \text{N}$). 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 systems among which the form does *not* change, and
- Observe that some of those good 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 firmed 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, as there is for waves in air or steel (compression waves). 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.

Vacuum is a unique state.

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*. Maxwell’s 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’s.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.
- And if the Earth itself dragged along the æther, then the observed “aberration

¹See Chapter 27.

²Although 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.

of starlight” wouldn't happen (Chapter 30).

- But if somehow Earth *didn't* drag the æther, then there would be an “æther wind,” and the Michelson–Morley and Fizeau experiments wouldn't have given the results that they did (Chapter 29).

Einstein also thought about an observer who flies over a water surface at the speed of wave propagation. Looking down, that observer sees waves that appear to be *standing still*.³ But there are no static wave solutions to the Maxwell equations, and again Einstein could not see any reasonable way to modify the equations to admit such solutions.

[T2] Section 28.2' (page 369) outlines the interpretation of the Michelson–Morley experiment and gives more details about the æther hypothesis.

28.3 THE NO-ÆTHER HYPOTHESIS

28.3.1 The vacuum is a unique state

So Einstein entertained the bizarre suggestion that Maxwell's equations were actually *correct and complete as written*.

- 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. Perhaps a different invariance that others had already considered 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 said, maybe we need to reexamine the experimental status of Newton's laws.

The preceding discussion 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 somebody smugly pronounce 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 μ_0 and ϵ_0) are *constants*. It has no further state variables beyond \vec{E} and \vec{B} that need to appear in Maxwell's 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”

³Your Turn 27B (page 361).

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 Maxwell's equations, Einstein's clarified a mathematical property (a new invariance) *already hiding* in them.⁴ 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 Maxwell's 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?⁵ 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 Maxwell's 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.⁶

FURTHER READING

Semipopular:

⁴Lorentz had already established this in 1904, but even today it is hard to grasp that from what he wrote.

⁵Public-key encryption could be an apt metaphor for this situation!

⁶It 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.

This fellow, and his gadget, are brilliant: <https://www.youtube.com/watch?v=1rLWVZVWfdY>.

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.

Technical:

Einstein's article: Einstein, 1998; Kennedy, 2012, chap. 4.

History: Michelson & Morley, 1887; Darrigol, 2022.

T2 21st century version of the Michelson-Morley experiment: Müller et al., 2003.

28.2'a 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 the prediction. In more detail, we will test the hypothesis that light is a vibration in a fluid governed by Newtonian mechanics.⁷

Setup

We consider the usual simplified version of the apparatus sketched in Figure 28.2.⁸ Choose coordinates such that the incoming beam travels along the $-\hat{x}$ axis. We will assume that both arms of the interferometer have the same length b . The half-silvered mirror (henceforth “half-mirror”) is oriented at 45 deg to the \hat{x} axis.

Most pre-Einstein physicists were implicitly entertaining something equivalent to 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 at c . This is also the situation 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.⁹
- Interference of light beams split from a common precursor beam is governed by total transit time between the splitting and recombination. Because time is invariant under galilean boosts, we can and will choose to compute 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 and its geometry is unaffected by æther motion. In particular, the apparatus is not changed when it is rotated, or equivalently when the direction of its motion relative to the æther is changed.
- The half-mirror splits a beam 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.¹⁰

⁷Maxwell actually proposed an experiment like Michelson–Morley in 1879, but rejected it as impractical at that time.

⁸Michelson did a preliminary experiment in 1881 with such an apparatus. For the famous experiment, MM created a more elaborate, folded light path to increase their sensitivity.

⁹In *other* G-inertial coordinate systems, light will therefore travel with anisotropic velocity, by the galilean velocity addition formula.

¹⁰We will not explicitly mention the time delay from passing through the glass that composes this

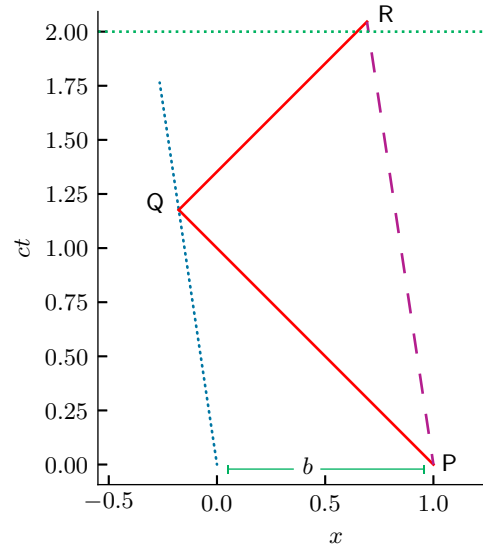


Figure 28.1: [Space-time diagram.] **Light path in the left arm of MM apparatus**, for $b = 1$ m and the exaggerated case $\beta = 0.15$. *Solid red*: Path of light wavefront (at ± 45 deg). *Dashed purple*: Trajectory of half-silvered beam splitter. *Dotted blue*: Trajectory of ordinary mirror on the left.

Again, we will analyze the experiment in one of the special (wind-free) G-inertial systems. For simplicity, suppose that the apparatus is moving along $-\hat{x}$ at speed $v = \beta c$ with respect to the stationary æther, and ask about the total transit times in each arm and how they depend on β . Although we cannot experimentally change the magnitude of β , we can reverse its sign by rotating the apparatus by π .

Everything prior to P is common to both trajectories, so we can neglect that and let time begin at P.

Left arm

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. The incidence is at 0 deg from perpendicular, so by the law of reflection 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 located in a plane of constant y .

Let’s find the transit time for the isolated wavecrest to traverse PQR. Notice that P, Q, and R all sit in the plane $y = 0$. Let P be the point with $\begin{bmatrix} ct \\ x \end{bmatrix} = \begin{bmatrix} 0 \\ b \end{bmatrix}$. Then solve the following two equations to get the intersection of the left mirror trajectory and the ray:

$$\begin{bmatrix} 0 \\ b \end{bmatrix} + \xi \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} ct_Q \\ -\beta ct_Q \end{bmatrix}$$

Eliminating ξ yields that $Q = \begin{bmatrix} b/(1-\beta) \\ -\beta b/(1-\beta) \end{bmatrix}$ (Figure 28.1).

Next, find R, the intersection of the half-mirror with the reflected light trajectory. Solve:

$$\begin{bmatrix} b/(1-\beta)+\eta \\ -\beta b/(1-\beta)+\eta \end{bmatrix} = \begin{bmatrix} ct_R \\ b-\beta ct_R \end{bmatrix}.$$

So the total transit time for PQR is just $t_R = 2(b/c)/(1 - \beta^2)$.

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.

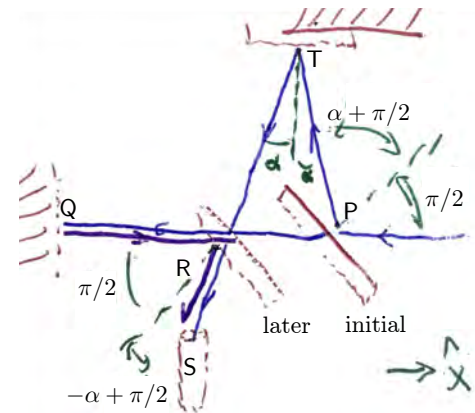


Figure 28.2: [*xy* diagram.] Light paths in both arms of the apparatus. [Not ready yet.]

The time at point R is indeed slightly different from the case of $\beta = 0$ ($ct_R = 2.05$ m for the example in the figure, 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 28.2).¹¹

Begin with the upper arm of the light path. A beam 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, it proceeds at some angle θ to its original direction. Thus, its angle of reflection is $\pi/4 + \alpha$, where α and θ are related by

$$\pi/4 + \alpha = (\pi - \theta) - \pi/4, \quad \text{or} \quad \theta = \pi/2 - \alpha.$$

For a stationary mirror, θ would equal $\pi/2$, 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 $g(-x \cos \theta + y \sin \theta - ct)$. The boundary condition of the wave requires that the outgoing and incoming waves be related near P by $f = -g$ along the surface $\{y = -x - vt\}$. Thus,

$$\begin{aligned} f(-x - ct) &= -g(-x \cos \theta + (-x - \beta ct) \sin \theta - ct) \\ &= -g\left(\left(1 + \beta \sin \theta\right)\left(-x \frac{\cos \theta + \sin \theta}{1 + \beta \sin \theta} - ct\right)\right). \end{aligned}$$

For this to hold at all x and t , we must have that $(\cos \theta + \sin \theta)/(1 + \beta \sin \theta) = 1$, or

$$\sin \alpha + \cos \alpha = 1 + \beta \cos \alpha. \quad (28.1)$$

(In addition, g is Doppler stretched: $g(u) = -f(u/(1 + \beta \cos \alpha))$.)

Expanding Equation 28.1 for small α and β 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 at time b/c (T in Figure 28.2). A similar argument with reversed sign is applicable to the encounter of the left beam with the beamsplitter, and shows that upon this final reflection it, too, emerges heading toward the point where S will be located.¹²

¹¹The key idea for the following argument is discussed in Soni, 1988.

¹²There 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.

Upper arm

Now that we know the light path in the upper arm, we can find the transit time. Because we are in a wind-free coordinate system, the time is just path length divided by c , or

$$\frac{2b/c}{\cos \alpha} \approx (2b/c)(1 + \frac{1}{2}\beta^2).$$

The predicted difference between left and upper arm transit times is then

$$\approx (2b/c)(1 + \beta^2 - (1 + \frac{1}{2}\beta^2)) = b\beta^2/c \neq 0.$$

Consequences

The transit times therefore differ by a β -dependent amount. Michelson and Morley designed their experiment so that the small predicted difference, if present, would be measurable. 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 were done on mountaintops to minimize this possible effect).
- The apparatus is not rigid but instead gets shrunken by the æther wind ("Lorentz-FitzGerald contraction").

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, most scientists were unwilling to entertain the resulting theory because it lacks galilean invariance.

Postscript

The usual "swimmer" analogy is potentially misleading: Anthropomorphizing like this brings 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 looks.

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 343), 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 always related to one another by ordinary Lorentz transformations; those transformations are invariances of the all 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.

T₂

28.3' Poincaré's work

"History has not been kind to [Poincaré]'s contributions. In his *Science and Hypothesis*, first published in 1902, Poincaré boldly declares:

1. 'There is no absolute space, and we only conceive of relative motion; and yet in most cases mechanical facts are enunciated as if there is an absolute space to which they can be referred.
2. There is no absolute time. When we say that two periods are equal, the statement has no meaning and can only acquire a meaning by a convention.
3. Not only have we no direct intuition of the equality of two periods, but we have not even direct intuition of the simultaneity of two events occurring in two different places.
4. Finally, is not our Euclidean geometry in itself only a kind of convention of language?'

These ideas are at the heart of relativity, and it is difficult to believe they did not have a profound effect upon Einstein's thinking. Poincaré was also the first to use the term 'principle of relativity,' which is also stated forthrightly in *Science and Hypothesis*. In a famous 1904 speech at the International Congress of Arts and Sciences in St. Louis, Poincaré even glimpses a new theory in which 'the velocity of light becomes an impassable limit.' But the mathematician did more than make oracular pronouncements; he wrote a pair of technical papers on Lorentz's theory, and in the longer one, completed just before Einstein's own, he has nearly everything his shadowy rival does, and in some respects more. In that paper, Poincaré shows, as Lorentz did, that Maxwell's equations are invariant if the Lorentz transformation is correct; he anticipates Minkowski's combining of space and time, and he virtually derives $E = mc^2$. What Einstein did in those fateful weeks that Poincaré did not was to show that the whole thing results from just the two postulates: the principle of relativity and the constancy of the speed of light." — Rothman, 2003

PROBLEMS

28.1 $\boxed{T_2}$ *Moving mirror*

Fill in the steps leading to Equation 28.1 (page 371).

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 $\vec{r}_{(1)}(t), \vec{r}_{(2)}(t), \dots$, then the modified trajectories

$$\tilde{\vec{r}}_{(1)}(t) = \vec{r}_{(1)}(t) + \vec{v}_*t, \quad \tilde{\vec{r}}_{(2)}(t) = \vec{r}_{(2)}(t) + \vec{v}_*t, \dots$$

will also solve the same equations. Here \vec{v}_* is one overall constant vector.

The equivalent passive view relabels all the events in spacetime according to

$$\begin{bmatrix} ct' \\ x' \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -v_*/c & 1 \end{bmatrix} \begin{bmatrix} ct \\ x \end{bmatrix}, \quad [26.9, \text{page 349}]$$

or a similar formula in three spatial dimensions. Chapter 26 showed that if we take some equations of newtonian physics (two masses joined by a spring¹) and re-express them in terms of the primed coordinates, the new versions have the same algebraic

¹Or 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 particle is moving at constant velocity \vec{v}_0 . Now introduce a new coordinate system related to the first by a galilean boost with velocity \vec{v}_ . The wave or particle will be observed in the second system to be moving at constant velocity $\vec{v}_0 - \vec{v}_*$.* (29.1)

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 periodically 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 independent 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 motion 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 (Section 27.4).
- Galilean invariance is what guaranteed the Principle of Relativity, which is experimentally validated.

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

²Section 26.6.6 (page 352).

³Section 27.4 (page 360).

⁴We 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 Michelson–Morley experiment, but there were other such experiments at the time, and all (eventually) came out null. One appeared in the very first volume of *Physical Review*: Franklin & Nichols, 1894.

neglecting all the delicious complexity brought by the vector character of electromagnetic fields.⁶ 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 $\vec{r}' = 2\vec{r}$, 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 good 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 non-galilean invariance connecting 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 $c \approx 3 \cdot 10^8$ 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, just 14 years after Maxwell. 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 26.9, page 349) 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 a line at 45 deg to the horizontal. Such a trajectory is drawn as a solid diagonal line in the figure below:⁷

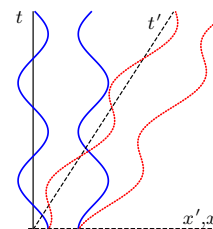
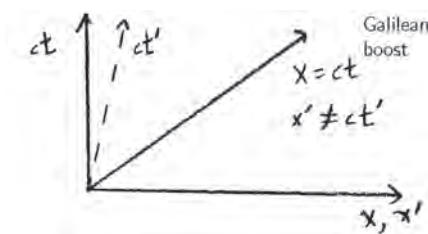


Fig. 26.2 (page 349)

⁶This oversimplification will be remedied in Chapters 32 and 33.

⁷See also Figure 26.2 (page 349).

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 (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 another 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 Figure 29.1b.

29.4 THE WAVE EQUATION IS INVARIANT UNDER PROVISIONAL LORENTZ TRANSFORMATIONS

29.4.1 Coordinate transformation

Figure 29.1b represents the following linear transformation of coordinates:

$$\begin{bmatrix} ct' \\ x' \end{bmatrix} = \gamma \begin{bmatrix} 1 & -\beta \\ -\beta & 1 \end{bmatrix} \begin{bmatrix} ct \\ x \end{bmatrix}. \quad \text{provisional Lorentz boost transformation} \quad (29.2)$$

Here β and γ 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 not all 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

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: $\phi_{\pm}(t, x) = \cos(\frac{\omega}{c}(-ct \pm x))$. Following Idea 26.5 (page 345), we apply an active transformation, that is, construct different functions $\tilde{\phi}_{\pm}$ defined by $\tilde{\phi}_{\pm} = \cos(\frac{\omega}{c}(-ct' \pm x'))$, or

$$\tilde{\phi}_{\pm}(t, x) = \cos(\frac{\omega}{c}(-\gamma(ct - \beta x) \pm \gamma(-\beta ct + x))),$$

⁸Chapter 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.

⁹Section 29.7 will expand on this point.

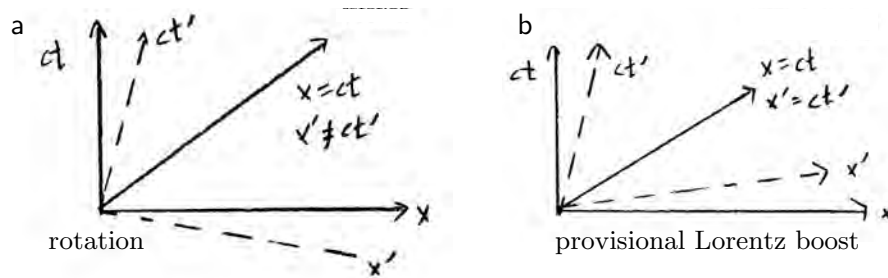


Figure 29.1: Linear transformations on spacetime. (a) Rotation of axes. (b) Bending the axes by opposite angles.

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 switch to 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 358, 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.

Indeed, the wave equation is invariant under a family of transformations that take a coordinate system and boost it into uniform straight-line 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 352, we suppose that when the catapult is at rest, it fires a projectile into uniform motion with velocity v_* . The corresponding trajectory can again be written in parametric form as $\begin{bmatrix} ct \\ x \end{bmatrix} = \begin{bmatrix} \xi \\ v_*\xi/c \end{bmatrix}$, and that of the stationary catapult

itself as $\begin{bmatrix} ct \\ x \end{bmatrix} = \begin{bmatrix} \xi \\ 0 \end{bmatrix}$. We then apply an active boost transformation with velocity $v_{*(2)} = \beta c$ 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 $\begin{bmatrix} ct' \\ x' \end{bmatrix} = \begin{bmatrix} \xi \\ 0 \end{bmatrix}$, and the projectile's is $\begin{bmatrix} ct' \\ x' \end{bmatrix} = \begin{bmatrix} \xi \\ v_* \xi / c \end{bmatrix}$. Now express these in terms of the original coordinates:

$$\text{catapult: } \gamma \begin{bmatrix} 1 & -\beta \\ -\beta & 1 \end{bmatrix} \begin{bmatrix} ct \\ x \end{bmatrix} = \begin{bmatrix} \xi \\ 0 \end{bmatrix}; \quad \text{projectile: } \gamma \begin{bmatrix} 1 & -\beta \\ -\beta & 1 \end{bmatrix} \begin{bmatrix} ct \\ x \end{bmatrix} = \begin{bmatrix} \xi \\ v_* \xi \end{bmatrix}.$$

Multiply both sides by the inverse matrix:

$$\text{catapult: } \begin{bmatrix} ct \\ x \end{bmatrix} = \frac{1}{\gamma(1-\beta^2)} \begin{bmatrix} 1 & \beta \\ \beta & 1 \end{bmatrix} \begin{bmatrix} \xi \\ 0 \end{bmatrix}; \quad \text{projectile: } \begin{bmatrix} ct \\ x \end{bmatrix} = \frac{1}{\gamma(1-\beta^2)} \begin{bmatrix} 1 & \beta \\ \beta & 1 \end{bmatrix} \begin{bmatrix} \xi \\ v_* \xi \end{bmatrix}. \quad (29.3)$$

Thus, both catapult and projectile are in uniform motion with respect to the ground: The velocity of the catapult is $c\Delta x/\Delta(ct) = \beta c$ as desired, whereas that of the projectile is

$$v_{\text{lab}} = c\Delta x/\Delta(ct) = (\beta c + v_*)/(1 + \beta v_*/c). \quad (29.4)$$

Equation 29.4 is a disturbing result. It surely doesn't look like the galilean formula $v_{\text{lab}} = \beta c \pm v_*$. But suppose that $|\beta| \ll 1$ and $|v_*| \ll c$; 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 $v_* \rightarrow \pm c$, our formula boils down to $v_{\text{lab}} \rightarrow \pm c$. *A trajectory that moves at speed c in (t', x') has the same property in (t, 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

Finally, you should think about the limit $\beta \rightarrow 1$, holding v_* 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.

¹⁰Section 26.4.2 (page 345).

¹¹This 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 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.

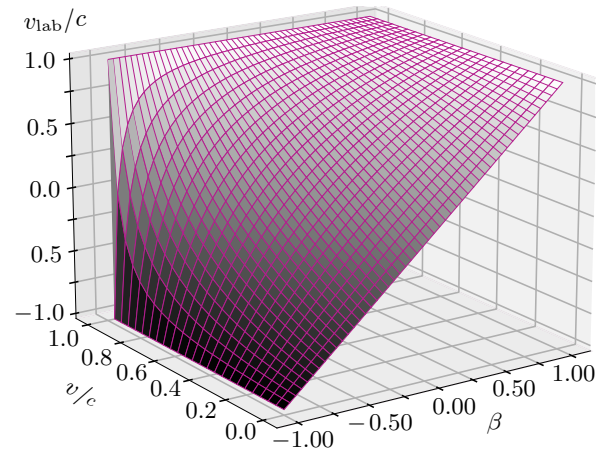


Figure 29.2: The velocity addition formula (Equation 29.4) never gives a result larger than c .

Your Turn 29C

- Suppose that a boost with (γ_1, β_1) is followed by another with (γ_2, β_2) . Show that the combined transformation is again of the form Equation 29.2, with new values for γ and β .
- 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. Later, when we finish specifying the relation between γ and β in Section 30.3, this group will be called the “Lorentz group.”

29.6 A NONNULL, FALSIFIABLE PREDICTION

Is this just speculation? We should think about some real 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 the aberration of starlight and an experiment first done by M. Fizeau.¹² We’ll discuss the second of these now, and the first in Chapter 30.

It is sometimes said that the other, 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’a, page 369). Null experiments are subject to the criticism that zero is a very special value. There may

¹²Fizeau’s experiment was first done in 1859, then redone with greater precision by A. Michelson and E. Morley in 1886 (a little-known result published a year before the famous MM experiment). Figure 29.3 below shows their data. 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.

be various explanations for why you got zero (maybe your sensitivity wasn't as good as you thought).¹³

The speed of light in flowing water is different from that in still water, but in a quantitatively different way from the newtonian expectation.

It's more convincing when two theories make two quantitative, 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 old astronomical measurement.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. 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.¹⁵ Unlike in vacuum, he found that the motion of the water can slow down or speed up the light, depending on its motion.

Let's apply the Relativity Strategy to this problem (Idea 26.14, page 351):

- 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 βc 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 βc and light flashes at $v_{\text{lab}} = (\pm cn^{-1} + \beta c)/(1 + \beta(\pm cn^{-1})/c)$.

In experiments, we can never get the water flowing anywhere near the speed of light. So $|\beta| \ll 1$, and we can make a simplified approximate formula:¹⁶

$$v_{\text{lab}} \approx c(\beta \pm n^{-1})(1 \mp \beta/n) \approx \pm c/n + \beta c(1 - n^{-2}). \quad (29.5)$$

At last, we have a testable prediction. The hypothesis that the full equations of electromagnetism have galilean invariance predicts $v_{\text{lab}} = c(\pm n^{-1} + \beta)$, which *differs from Equation 29.5*. If we plot v_{lab} (speed of light in water, measured in the lab's coordinate system) versus β , then the two competing theories make different predictions for the slope of the data.

That is, both theories make firm, nonnull predictions, with no fudge factors (no fit parameters).¹⁷ That is, both are highly *falsifiable*, if you've got enough accuracy

¹³To get a convincing 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 usually portrayed, and it certainly was not viewed as such by contemporaries.

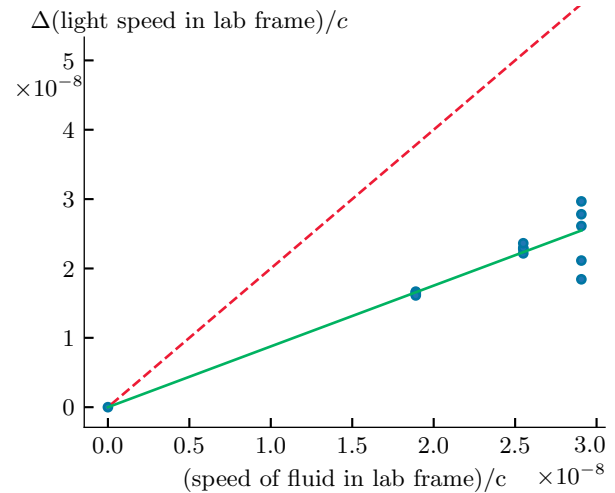
¹⁴Light travels a tiny bit slower in air than in interplanetary space.

¹⁵Today one uses a chunk of quartz on the rim of a rapidly spinning disk to eliminate turbulence that occurs in water.

¹⁶The first published derivation of this formula seems to be by Max von Laue (1907).

¹⁷The value of n was independently measured from experiments on refraction. So it's a parameter, but not a fit parameter. $\boxed{\mathcal{T}2}$ Our derivation neglects the effect of dispersion in the medium; see L'erche, 1977.

Figure 29.3: [Experimental data.] **Some results from Michelson and Morley’s lesser-known experiment.** A beam of light was split into two. One beam traversed a 6 m pipe parallel to the flow of water; the other traveled the same distance with water flowing the opposite direction. Interferometry was used to compare the velocities in each pipe (*vertical axis*); thus, the curve must pass through 0 when the water is at rest (*leftmost dot*). Other dots show the data from a total of 13 trials spanning three different nonzero fluid velocities. The *solid line* shows the prediction of Equation 29.5 with $n = 1.78$. For comparison, the *dashed line* shows the prediction based on the hypothesis of galilean invariance. [Data from Michelson & Morley, 1886.]



to measure the effect at all. Figure 29.3 shows the data from Michelson and Morley’s improved version of Fizeau’s experiment. To get enough accuracy, Fizeau and successors used interferometry. The figure shows significant scatter, but the data certainly rule out the prediction of the galilean invariance theory (slope 1, dashed line),¹⁸ and they don’t rule out Einstein’s prediction.

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; in the figure, these differences have been divided by two to show the change relative to still water. Because the graph shows only a small range of values for (water speed)/ c , the Einstein prediction appears to be nearly a straight line; the data agree. (At water speed approaching c , Figure 29.2, page 381 predicts that the curve would level off, but we’re nowhere near that regime.)

Equation 29.5 has another key feature. Suppose that we remove the water, that is, we set $n \rightarrow 1$. Then we find that $v \rightarrow \pm c$. That was after all our starting point: The speed of light in vacuum must always equal c .

29.7 PLUS ULTRA

1. Our provisional Lorentz boost (Equation 29.2) has the disturbing feature that $t' \neq t$. A lot of people got (and still get) philosophically confused: “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.

¹⁸After 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.

- 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 know ways to attach sets of four numbers to events.¹⁹ 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 transformations that include (at least some of) the ones given in Equation 29.2.
- It is true that these 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? (Einstein couldn’t find any.) Why must the good coordinate systems all agree about the value of t ? (Einstein couldn’t see why they must.)

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 (like the rest of physics) is invariant under provisional Lorentz transformations.²⁰ The kinematic approach followed above is much simpler²¹ than solving Maxwell’s equations for light moving through a medium of water molecules!

3. Although Lorentz invariance looks promising, we are far from being done. We wish to prove that the full Maxwell equations also have exact Lorentz invariance. 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:

Galison, 2003; Pais, 1982.

Intermediate:

Lahaye et al., 2012; Zhang, 1997.

Technical:

Fizeau, 1859.

¹⁹One way Einstein suggested to set up such a coordinate system is to use an array of identical clocks and synchronize them using light flashes.

²⁰Later, when we complete our specification of the Lorentz transformations, the derivation will still hold, because γ drops out of this particular prediction.

²¹Other scientists came close to relativity before Einstein. Today we regard their work as mostly unreadable, because they got bogged down in detailed dynamical hypotheses.

Einstein's recollection that two key experiments were "enough": Shankland, 1964; document with control number 1 168, Einstein Archive. Available in facsimile at the Einstein Archives Online as www.alberteinstein.info/db/ViewImage.do?DocumentID=34187&Page=1. Will, 2006a = arxiv.org/abs/gr-qc/0504085; Will, 2006b.

PROBLEMS

29.1

Confirm that the provisional Lorentz transformation (Equation 29.2, page 378) really implements the sketch Figure 29.1b and find the angles by which the ct and x axes are bent.

29.2 [Not ready yet.]

CHAPTER 30

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 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

The property of moving at speed $3 \cdot 10^8$ m/s (or not) 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, however, we'll refine our provisional form of our proposed transformations to get their final form.

¹Section 41.2 will discuss this statement in detail, but we already see the idea in Figure 29.1b (page 379): 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.

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 ones 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 346). Consider a situation that’s not just fields in empty space, specifically the Coulomb repulsion of two identical, charged particles:²

$$m \frac{\partial^2}{\partial t^2} \vec{r}_{(1)} = \frac{q^2 \vec{R}}{4\pi\epsilon_0 R^3}, \text{ and so on, where } \vec{R} = \vec{r}_{(1)} - \vec{r}_{(2)}. \quad (30.1)$$

Here the two point charges, labeled by $\alpha = 1$ or 2 , are assumed to be identical (the same charge q and mass m).

Now consider the provisional Lorentz boost with $\gamma \neq 1$ but $\beta = 0$, that is, $\vec{r}' = \gamma\vec{r}$ and $t' = \gamma t$. 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 γ 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 be a whole family of different electrons with continuously varying charges and/or

²We combine the equations in this way so that we don’t have to worry about how \vec{E} transforms; it’s been eliminated. Although this formula will later need relativistic corrections, it’s certainly valid for slowly-moving particles.

³It’s true that the new equation has the same form except for the value of $q^2/(m\epsilon_0)$, but that’s not good 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.

⁴Atomic 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.

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 course 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 of those transformations in succession does *not* amount to any single boost with $\gamma = 1$. That is, the $\gamma = 1$ transformations do not “close into a group.”


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, then we can explore the physical hypothesis that this reduced set of transformations are invariances of all of Nature.

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 γ is not independent of the boost velocity $\vec{\beta}$, but instead is a scalar

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.

⁶The logic here parallels our rescue of galilean invariance by acknowledging a new dynamical variable in Section 27.4 (page 360).

⁷See Chapters 32–34.

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 γ won't depend on which direction $\vec{\beta}$ points. We also expect that a boost by $\vec{\beta}c$, followed by a boost by $-\vec{\beta}c$, should amount to no boost at all (think about jogging backwards at speed v_* inside a train car that itself is moving at $+v_*$ relative to Earth). Thus, we require

$$\gamma \begin{bmatrix} 1 & \beta \\ \beta & 1 \end{bmatrix} \gamma \begin{bmatrix} 1 & -\beta \\ -\beta & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

This fixes $\gamma = (1 - \beta^2)^{-1/2}$, or⁸

$$\begin{bmatrix} ct' \\ x' \end{bmatrix} = (1 - \beta^2)^{-1/2} \begin{bmatrix} 1 & -\beta \\ -\beta & 1 \end{bmatrix} \begin{bmatrix} ct \\ x \end{bmatrix}. \quad \text{Lorentz boost, 1D} \quad (30.2)$$

In particular, if $\beta = 0$ then $\gamma = 1$, and so pure dilations are *not* part of this subset of transformations, as desired. From now on, γ will always mean this particular function of β .

For very small β , we see that $\gamma \rightarrow 1$ and the transformations Equation 30.2 reduce to

$$t' \approx t - (v_*/c^2)x \approx t, \quad x' \approx x - v_*t \quad \text{where } v_* = \beta c.$$

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.

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 379):

$$\left[-\frac{\partial^2}{\partial(ct)^2} + \frac{\partial^2}{\partial x^2} \right] u \quad \text{becomes} \quad \gamma^2(1 - \beta^2) \left[-\frac{\partial^2}{\partial(ct')^2} + \frac{\partial^2}{\partial x'^2} \right] u. \quad (30.3)$$

We see that among the provisional Lorentz boosts, the subgroup of true invariances are those that leave the wave operator *completely* form-invariant—not just a multiple of itself. The following chapters will show that indeed, these transformations are invariances of the full Maxwell/Lorentz system. That is, *the coordinate systems in which electrodynamics takes the simplest form are related 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.⁹

⁸Lorentz actually showed in 1904 that these transformations were invariances of the full Maxwell equations. However, Lorentz viewed this invariance as mathematical curiosity about the Maxwell equations—not an invariance of all of physics—and certainly not as justification to eliminate the æther.

⁹The latter were called galilean, or “G-inertial” systems in Section 26.6.1.

30.3.2 Rapidity parameter

The preceding section characterized true Lorentz transformations as those that leave something (the form of the wave operator) unchanged. Two such transformations in succession will *also* have that property, so right away we see that the Lorentz transformations must close into a group.

It's algebraically messy to prove that statement directly, but there is a remarkable reformulation that makes it easy. 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 \mathbf{S} is by the statements $\mathbf{S}^t \mathbf{S} = \mathbf{1}$ and $\det \mathbf{S} = 1$ (special orthogonal matrices). But equivalently, rotations are those linear maps of coordinates that leave the algebraic form of the pythagorean formula invariant:¹⁰ $x^2 + y^2 = (x')^2 + (y')^2$. Chapter 14 used this property to show that the Laplace operator is invariant under rotations. Rotations close into a group: For example, in 2D α_1 followed by α_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

$$c^2 \Delta \tau^2 = (c \Delta t)^2 - (\Delta x)^2, \quad (30.4)$$

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 30.4 invariant.

Because Equation 30.4 looks similar to the rotation case (except for the minus sign), we may hope that the appropriate symmetries will also look similar. Indeed,

$$\begin{bmatrix} \cosh \Upsilon & -\sinh \Upsilon \\ -\sinh \Upsilon & \cosh \Upsilon \end{bmatrix} \quad (30.5)$$

does the job, for any Υ . Some books call Υ the **rapidity parameter**.

Your Turn 30C

- Confirm that any transformation of the form Equation 30.5 is a special case of the provisional Lorentz boosts, with $\gamma = \cosh \Upsilon$ and $\beta = \tanh \Upsilon$...
- ... and that moreover, these transformations also satisfy the condition to be in the subgroup of true Lorentz transformations.
- Show that conversely, any Lorentz boost can be written in the form Equation 30.5.

Thus, once again, we have found a 1-parameter subset of the provisional 1D Lorentz boosts that closes into a group, excludes dilations, but does include boosts.

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, $\det \mathbf{S}$ may also equal -1 .

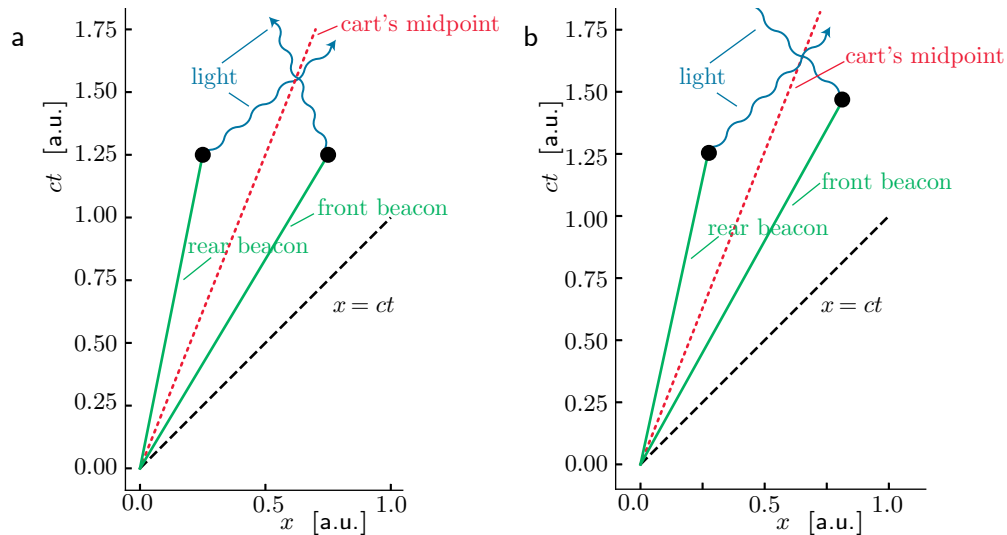


Figure 30.1: Thought experiment described in Section 30.4. The beacons emit flashes of light at the events shown as *dots*. (a) Newtonian kinematics. The coordinate axes refer to the G-inertial system in which the lab is at rest. Light flashes follow trajectories of unequal slope determined by the galilean velocity addition formula, Your Turn 26C (page 349). The flashes arrive back at the cart’s center simultaneously, showing that this experiment cannot detect the cart’s motion. (b) Relativistic kinematics. The coordinate axes refer to an E-inertial system in which the lab is at rest. This time, the flashes follow trajectories of slopes ± 45 degrees. Nevertheless, they arrive at the cart’s center simultaneously, due to the different lab times of emission. See text and Problem 30.2.

Compared to Section 30.3.1, however, the derivation just given has the advantage of revealing a *geometric* interpretation: Lorentz transformations are the analogs of rigid rotations in a weird new kind of geometry. Either way, we now have a candidate physical hypothesis about the invariances of Nature and can get to work testing it.

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 we place the following apparatus in a cart that moves at uniform velocity v with respect to the lab. The cart is rigid: Its length is always 0.5 meter, when measured in an inertial coordinate system in which it’s at rest. At some moment, we set two beacons at the center of the cart (as seen in its rest frame). 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 30.1 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 fixed by transforming the duration T to the lab’s E-inertial coordinate system (ct, x) .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 blue lines are the trajectories of the light flashes. Their slopes are $\pm 45^\circ$ in any E-inertial coordinate system, for example, the lab.

We see that, contrary to the claim in quotes above, the flashes coincide at the center of the cart, regardless of the value of u . Therefore we cannot use that observed coincidence to claim that u 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 30.4:

$$c^2 \Delta\tau^2 = (c\Delta t)^2 - \|\Delta\vec{r}\|^2, \quad (30.6)$$

and hence of the wave operator $\nabla^2 - \partial^2/\partial(ct)^2$, and hence of the wave equation itself. Thus, there are three independent kinds of 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 space and time translations gives a *ten*-parameter invariance group sometimes called the **Poincaré group**.¹²

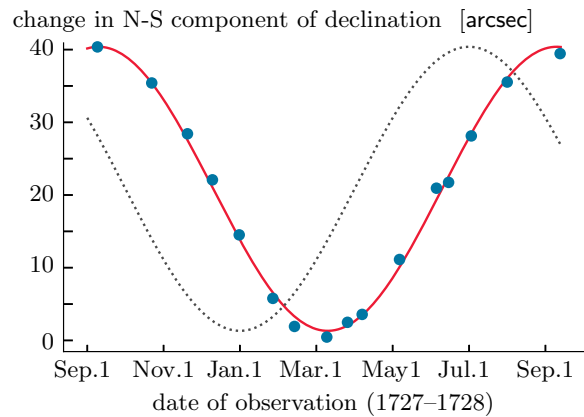
We can now put all our ideas together and formulate a successor to Idea 29.1 (page 376):

Suppose that we have a coordinate system on spacetime in which a wave or particle is moving at constant velocity \vec{v}_0 . Now introduce a new coordinate system related to the first by a Lorentz boost with velocity \vec{v}_ that is parallel to \vec{v}_0 . The wave or particle will be observed in the second system to be moving at constant velocity $c(-v_* + v_0)/(1 - v_*v_0/c^2)$ in the chosen direction.* (30.7)

¹¹This step in turn came from applying the Relativity Strategy (Equation 26.14, page 351).

¹²If we take the three-parameter group of spatial rotations and adjoin the three rigid spatial translations, the resulting six-parameter group is called the “euclidean group.”

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 cosine 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.)



If the two velocities are not parallel, then the formula is not as simple. However, the next section shows that in at least one important case it is still straightforward.

Your Turn 30D

The preceding discussion implies that only one of the following matrices is a Lorentz transformation:

$$\begin{bmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}; \quad \gamma \begin{bmatrix} 1 & -\beta & 0 & 0 \\ -\beta & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Show that.

30.6 MORE KEY EXPERIMENTS: ABERRATION OF STARLIGHT AND DOPPLER SHIFT

30.6.1 Light-speed trajectories bend while remaining at light 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.¹³ 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 arcsec.

Aberration of starlight.

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 measurements by James Bradley around 1726. As outlined in the epigraph to this chapter, Bradley was mystified to find there were indeed tiny annual variations in the relative positions of stars, but

¹³Actually, the period is one “sidereal year.” Refraction effects can be minimized by observing stars close to the zenith.

with the wrong annual phase to be explained by parallax (Figure 30.2).¹⁴ Bradley even found an explanation for this phenomenon, based on the hypothesis that light was a stream of 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, and in particular the distance to the source.
- 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 26.14, page 351). Consider a spacetime curve (particle or wavecrest trajectory) specified in parametric form by

$$\begin{bmatrix} ct \\ x \\ y \end{bmatrix} = \begin{bmatrix} \xi \\ \xi \\ 0 \end{bmatrix}.$$

This formula specifies a chain of events depending on a parameter ξ , 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.¹⁵ Applying a Lorentz boost transformation along \hat{y} yields the same trajectory as viewed in another E-inertial coordinate system:

$$\begin{bmatrix} ct' \\ x' \\ y' \end{bmatrix} = \begin{bmatrix} \gamma & 0 & -\gamma\beta \\ 0 & 1 & 0 \\ -\gamma\beta & 0 & \gamma \end{bmatrix} \begin{bmatrix} \xi \\ \xi \\ 0 \end{bmatrix} = \begin{bmatrix} \gamma\xi \\ \xi \\ -\gamma\beta\xi \end{bmatrix}. \quad (30.8)$$

Your Turn 30E

- Show that the new trajectory’s speed is $\sqrt{(\Delta x')^2 + (\Delta y')^2}/(\Delta t') = \sqrt{(\Delta\xi)^2 + \gamma^2\beta^2(\Delta\xi^2)}/(\gamma\Delta\xi/c)$.
- Confirm that this equals c , as it must.
- But the new trajectory is no longer directed along \hat{x}' . Show that instead, it makes an angle θ with the \hat{x}' axis, where $\tan\theta = \Delta y'/\Delta x' = -\gamma\beta$.

¹⁴In Bradley’s time, the distance to any star was unknown. Today we know that parallax would have been unobservably small with his instruments. But *no* hypothesized distance could reconcile his measurements with parallax.

¹⁵We’ll suppress the z coordinate to shorten the formulas. It’s there, but it’s not doing anything interesting.

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 deg). The new angle depends both on that original angle, and on β , so *the relative positions of the stars are different* according to the boosted (Earth-bound) observer. The effect is small because Earth’s velocity change over the course of a year is much smaller than c ; nevertheless, the effect was measurable in the 17th century.

Your Turn 30F

- 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.
- From that, estimate the maximum magnitude of stellar aberration.
- 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)?
- 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 fudge factors (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 its direction is arbitrary. Again boost along the y direction.

Your Turn 30G

- Show again that the apparent direction of \vec{k} changes, and find the change in its magnitude (as well as the change in ω).
- The passive 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).
- Specialize your result to the case with \vec{k} is parallel to the boost. Interpret your result in terms of a “longitudinal **Doppler shift**.” [*Hint*: Recall Section 29.4.2.]
- 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.”
- 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 magnitude from your prediction in (b) above.¹⁶ And newtonian kinematics predicts *zero* transverse shift, unlike your answer to (c)—a testable prediction.¹⁷

Doppler shifts.

Quantitative confirmation that the Doppler effect follows the relativistic formula, and excludes the galilean formula, had to wait for the Ives–Stilwell experiment (1938). Much more accurate experiments have been done right into the 21st century.¹⁸

30.7 AN ENORMOUS GENERALIZATION

30.7.1 Lorentz invariance must apply to all of physics

Let’s step back. Section 29.2.2 offered the paradox that the wave equation implied by Maxwell’s equations doesn’t have galilean invariance, so it was not clear that Maxwell is compatible with the Principle of Relativity. But we have now seen that the wave equation, with no modifications or additions, *is* invariant under a family of passive transformations that relate coordinate systems moving at constant velocity with respect to one another. We still need to do some work to upgrade this result to a corresponding statement about the full Maxwell equations, but looking ahead, we can state Einstein’s proposed resolution to the problem of Section 29.2.2 by saying¹⁹

Maxwell’s equations hardwire in the Principle of Relativity by using equations of motion that are invariant under Lorentz transformations—not galilean transformations.

Einstein took an extraordinary additional step.²⁰ 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 Maxwell’s 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:

*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.*

¹⁶Compare Equation 27.4, page 359.

¹⁷Hay, Schiffer, Cranshaw, and Egelstaff, *PRL* 4:165 1960.

¹⁸Some experiments were based on an ultrasensitive measure of wave frequency (Mössbauer effect). Other experiments used single atom emitters moving at high speeds.

¹⁹Compare our galilean statement (Equation 26.12, page 351).

²⁰Extraordinary but strangely familiar: Section 26.6.5 (page 350) 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.

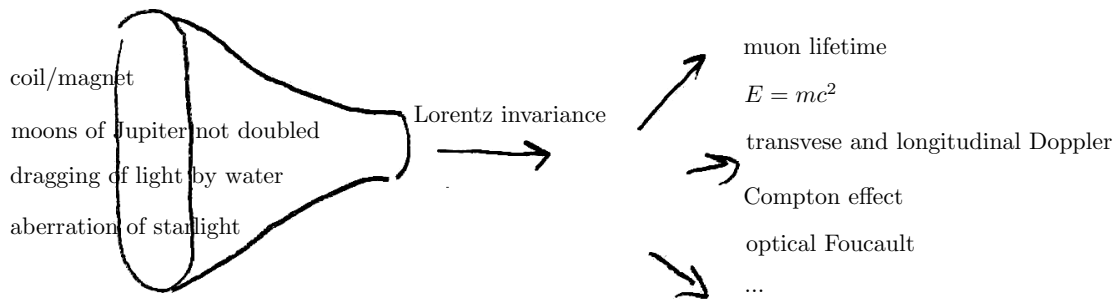


Figure 30.3: Evidence → hypothesis → predictions.

That’s quite a leap. We can’t have it both ways. Einstein’s hypothesis was that

*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 has to be changed**, not Maxwell.*

Or, paralleling Idea 26.11:

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. (30.9)

(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 30.3), 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 26.14, page 351). For example, we’ve seen how to understand Fizeau’s experiment, the aberration of starlight, and all kinds of Doppler shift, by using that approach.²¹

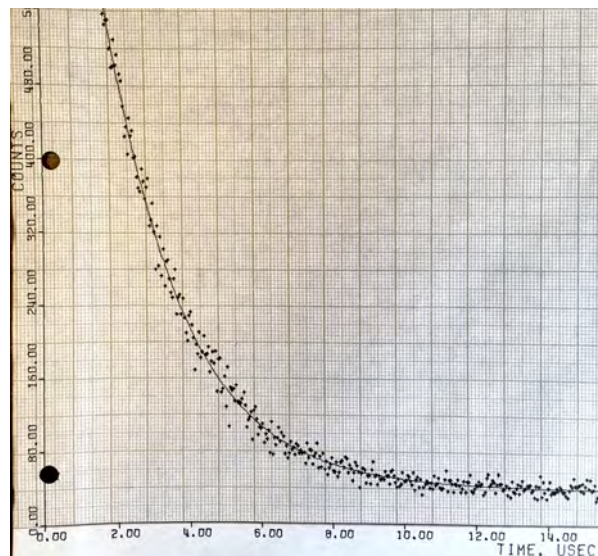
Note that when we hypothesize that “all laws of physics are invariant under Lorentz transformations,” we mean *all*, including quantum physics. Here are more examples:

Muon lifetime.

- 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 $2.2 \mu\text{s}$ (Figure 30.4). Then we can predict that a muon moving rapidly relative to the lab’s

²¹For 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. . . .”

Figure 30.4: [Experimental data.] **Lifetime measurement for muons nearly at rest.** Distribution of waiting times between detection of a muon entering a chamber and its subsequent disintegration after being captured in that chamber. The curve shows a fit to an exponential function plus a constant background. The background arose from accidental near-coincident events involving two different charged particle detections. [Data courtesy Janice Enagonio.]



E-inertial coordinate system will *also* live $2.2 \mu\text{s}$ in an E-inertial coordinate system in which the muon is at rest (a rest frame). Transforming this duration into the laboratory coordinate system via Equation 30.2 (page 390) 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 $\gamma(2.2 \mu\text{s})$, during which the muon travels $\gamma\beta c(2.2 \mu\text{s})$, 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,²² 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 later quantitatively verified.
- The Doppler shift formula also lets us deduce the motion of distant galaxies relative to us:²³ 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

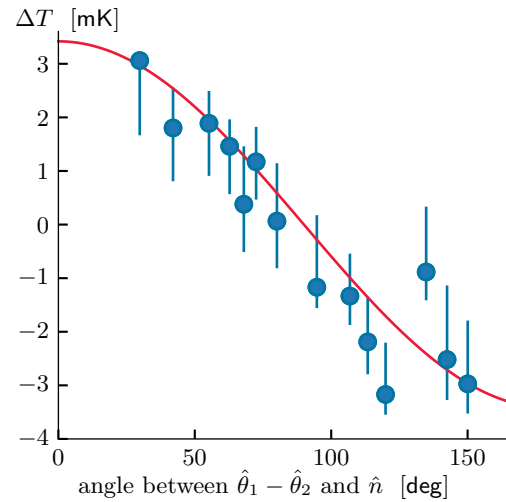
Mössbauer effect.

Doppler spectral line shifts.

²²You'll examine this phenomenon in Problem 31.2.

²³In 1868, William and Margaret Huggins detected a Doppler shift in the spectrum of Sirius, the birth of this indispensable astronomical method.

Figure 30.5: G. Smoot and coauthors' historic original data documenting the dipole anisotropy of CMBR. To cancel atmospheric emission, two receivers observed in two directions $\hat{\theta}_{1,2}$ on the sky and the apparent temperatures of their black body spectra were subtracted. Thus, we expect a ΔT proportional to $\hat{n} \cdot \hat{\theta}_1 - \hat{n} \cdot \hat{\theta}_2$, 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 the direction of maximal temperature was 10.8 hours right ascension and five degrees declination. [From Smoot et al., 1977.]



Cosmic microwave background dipole anisotropy.

that outgoing wave to our lab's E-inertial coordinate system gives its apparent frequency when we observe it with a spectrometer.

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 30.5).²⁴ This effect was observed shortly after the discovery of the cosmic microwave background radiation.²⁵ 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.

- Strong and weak nuclear forces, which are not electrodynamic in origin, lead to particle reactions that conserve energy and momentum. But we'll soon see 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 incredible 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 a miracle, the *basic epistemological miracle of physics*. It's what gives physical law a different character from the rules governing other branches of science.

Again: 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*.

²⁴Your result in Your Turn 30G includes the full angular dependence of this shift.

²⁵You'll explore the CMBR dipole anisotropy in Problem 30.6. Prediction: Peebles and Wilkinson, Phys. Rev. **174**(1968)2168. Observation: Figure 30.5.

T₂ Section 30.7.2' (page 403) 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; that nagging experiment with the magnet and coil seems to imply that \vec{E} *mixes with* \vec{B} under a boost (Hanging Question #A, page 12). So we need to augment our Lorentz transformations on spacetime by making a proposal for what exactly 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—after first inventing some powerful notation to help us (“high-tech relativity”), based on the close relation of Equation 30.5 to rotations.

2. Our logic may still feel a bit ad hoc, but here we were still just 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 in Chapters 32–33.

3. First, however, Chapter 31 will explore more generic (kinematic) consequences of Lorentz invariance, and their experimental signatures.

T₂

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**

$$\begin{bmatrix} u \\ v \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} ct \\ x \end{bmatrix} \quad \text{so} \quad \begin{bmatrix} ct \\ x \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}. \quad (30.10)$$

Then the general solution to the wave equation takes the simple form $f(u) + g(v)$ 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 $\partial^2/\partial u\partial v$. In terms of the transformed coordinates, this is $(AB)(\partial^2/\partial u'\partial v')$. 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 $A = B \neq 1$; 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.

$$\begin{bmatrix} ct' \\ x' \end{bmatrix} = \frac{1}{2} \begin{bmatrix} A+A^{-1} & A-A^{-1} \\ A-A^{-1} & A+A^{-1} \end{bmatrix} \begin{bmatrix} ct \\ x \end{bmatrix}. \quad (30.11)$$

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 30.5.

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 invariant under Lorentz transformations (it acquires a factor of $A/A = 1$). Indeed, it is just the invariant interval between two events. If those 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 $A_{\text{tot}} = A_1A_2$. To interpret this result, invert the relations between A and (β, γ) to find

$$A = \gamma(1 + \beta) \quad \text{or} \quad A^{-1} = \gamma(1 - \beta).$$

Thus, $A_{\text{tot}} = \gamma_1(1 + \beta_1)\gamma_2(1 + \beta_2)$ gives

$$\beta_{\text{tot}} = \frac{(1 + \beta_1)(1 + \beta_2) - (1 - \beta_1)(1 - \beta_2)}{(1 + \beta_1)(1 + \beta_2) + (1 - \beta_1)(1 - \beta_2)} = \frac{\beta_1 + \beta_2}{1 + \beta_1\beta_2}.$$

and we recover Equation 29.4 (page 380).

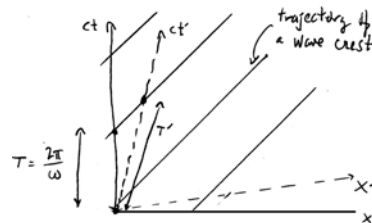
30.3'd Relation to rapidity

Equation 30.11 is the same as Equation 30.5 (page 391) with $\Upsilon = \ln A$. This is helpful, because in light-cone coordinates the composition law is simply $A_{\text{tot}} = A_1 A_2$ (show that). So $\Upsilon_{\text{tot}} = \ln(A_1 A_2) = \Upsilon_1 + \Upsilon_2$, which agrees with your result in Your Turn 30B (page 391).

T₂

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.

T₂

30.7.2' More about muon lifetime

The muon had not yet been discovered in 1905, so the result in Figure 30.4 (page 399) was not available to Einstein. We now call the relevant physical law “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 $2.2 \mu\text{s}$. 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 when 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.

PROBLEMS

30.1 *Rapidity*

Continue Your Turn 30C (page 391):

- a. Section 30.3.1 argued that because the transformations Equation 30.5 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 30.5 to show that a boost with Υ_1 , followed by one with Υ_2 , is equivalent to a single boost with $\Upsilon_{\text{tot}} = \Upsilon_1 + \Upsilon_2$.
- 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 380).

30.2 *Cart before the horse*

Figure 30.1 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 $u = \pm 0.4c$ in the cart's rest frame. For concreteness, suppose that in its rest frame, the cart's total length 0.5 m 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 Δx in the lab coordinate system, because the wire is neutral. The quantities I , Δq and v are all positive. Thus, there is current in the $+\hat{x}$ direction. We are imagining a continuum limit where $\Delta q \rightarrow 0$ holding fixed the linear charge density $\Delta q/\Delta x$.

- a. Write an expression for v in terms of I , Δq , and Δx .

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 Lorentz-invariant (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 [Not ready yet.]

30.5 **T2** 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.

A ring of mirrors sets up an “optical Foucault pendulum.”

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 clockwise or counterclockwise (when viewed on a line directed toward the center of Earth.) The apparatus is much smaller than Earth: $L = 1$ m. It is located at north latitude α , that is, the polar angle measured from the north pole is $\theta = \pi/2 - \alpha$.

- a. You know the angular frequency ω 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 ϵ . 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.

- 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 26.14).]
- Repeat for a light beam circulating counterclockwise.
- Subtract the two preceding results and express your answer in terms of θ , L , ω , R_{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.
- Evaluate your answer for an apparatus located at north latitude $\alpha = \pi/4$. Which transit time is faster: the clockwise or the counterclockwise route?
- Compare your answer to the period of visible light. Is this a measurable effect?

30.6 CMBR anisotropy

- Generalize Your Turn 30G to three spatial dimensions. That is, start with a plane wave with angular frequency ω 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 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 wavenumbers drawn from the isotropic probability distribution

$$\wp(\vec{k})d^3k = C f_0(\|\vec{k}\|/\tau)d^3k. \quad (30.12)$$

In this formula, \wp is a probability density function. Recall what that means: In a little box of \vec{k} space with volume d^3k , we have $M\wp(\vec{k})d^3k$ component plane waves, where M is some big constant. The constant τ is related to the temperature of the radiation (that is, $(2.7\text{ K})k_{\text{B}}/(\hbar c)$), 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 E-inertial but moving at speed v_* 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.

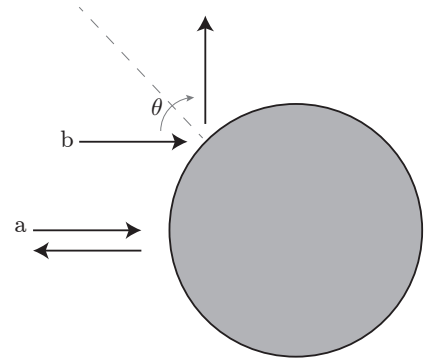


Figure 30.6: See Problem 30.8.

- b. Find the probability density $\wp'(\vec{k}')$ of \vec{k}' vectors. (You can forget about polarization.) Show that the distribution in the primed system, restricted to any particular direction \hat{k}' , again has the Planck form, but with a direction-dependent effective temperature $\tau_{\text{eff}}(\hat{k}')$, which you are to find.

[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 Lorentz I

[Not ready yet.]

30.8 Disco discovery

In this problem, use classical (not quantum) physics.

An electromagnetic plane wave has wavelength λ 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 $\tilde{\lambda}$ as measured in the lab system, in terms of λ , v , and physical constants. [Hint: Apply the Relativity Strategy (Equation 26.14, page 351). 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 θ with the \hat{z} axis in the lab system. [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 θ .]

[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 domain of classical EM (coherent states of many photons, for example, bouncing radio off a satellite) your result is accurate.

Radar speed trap. (ii) Bouncing a radar beam off a speeding car and measuring the beat frequency between outgoing and returning signals is another real-world application.]

30.9 *Velocity addition*

[Not ready yet.]

CHAPTER 31

Relativistic Momentum and Energy of Particles

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

There is another outstanding kinematic consequence of the hypothesis that all of physics, not just electrodynamics, is Lorentz invariant. It concerns energy and momentum. However, the experiments confirming it (and later ending WW2), came long after Einstein's initial discovery, which was based on—Electromagnetic Phenomena.

Electromagnetic phenomenon: Huge amounts of energy can be liberated in nuclear reactions.

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 Maxwell's 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, all initially mutually noninteracting, which come together and interact during a finite time interval (a “collision”), and suppose that eventually some other “particles” emerge that are again noninteracting. Thus, 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 two particles with masses $m_{1,2}$ and velocities $\vec{v}_{(1,2)}$ are initially noninteracting, then a “collision” occurs, and two other particles with $m_{3,4}$ and $\vec{v}_{(3,4)}$ emerge, eventually separating so that they are again noninteracting. In first-year physics, we start with Newton's laws and prove that

$$\vec{p}_{(1)}^N + \vec{p}_{(2)}^N = \vec{p}_{(3)}^N + \vec{p}_{(4)}^N, \quad \text{where } \vec{p}_{(\ell)}^N = m_{\ell}\vec{v}_{(\ell)} \quad (\text{newtonian}). \quad (31.1)$$

The quantity $\vec{p}_{(\ell)}$ is called the **newtonian momentum** of particle ℓ .

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 rotation-invariant (scalar), so the $m_\ell \vec{v}_\ell$ are also 3-vectors:

$$\vec{p}_{(\ell)}^N = S \vec{p}_{(\ell)}^{N'} \quad (31.2)$$

When we express each term of Equation 31.1 in terms of a rotated coordinate system, then, the matrix S is a common factor:

$$S(\vec{p}_{(1)}^{N'} + \vec{p}_{(2)}^{N'} - \vec{p}_{(3)}^{N'} - \vec{p}_{(4)}^{N'}) = 0. \quad (31.3)$$

Multiplying both sides of this equation by S^{-1} 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:

$$m_1 + m_2 = m_3 + m_4 \quad (\text{newtonian}). \quad (31.4)$$

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}' is a complicated, nonlinear function of \vec{v} (Equation 29.4, page 380). Indeed, given a set of four momenta $\vec{p}_{(\ell)}^N$ that 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 exploring. 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:¹

$$\mathcal{E}_{(1)}^N + \mathcal{E}_{(2)}^N = \mathcal{E}_{(3)}^N + \mathcal{E}_{(4)}^N, \quad \text{where } \mathcal{E}_{(\ell)}^N = \frac{1}{2} m_\ell \|\vec{v}_\ell\|^2. \quad (\text{newtonian}) \quad (31.5)$$

This formula is rotation invariant by an even easier argument than before: Each term is *separately* invariant.

¹Newton himself didn't use conservation of energy. Although 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 energy as a distinct concept, and she disseminated that view in her translation and commentaries on Newton.

Ex. Check that Equation 31.5 is galilean invariant.

Solution:

$$\begin{aligned}\sum_{\ell} \mathcal{E}_{(\ell)}^{N'} &= \sum_{\ell} \frac{1}{2} m_{\ell} \|\vec{v}_{(\ell)} - \vec{v}_{*}\|^2 = \sum_{\ell} (\mathcal{E}_{(\ell)}^N - m_{\ell} \vec{v}_{(\ell)} \cdot \vec{v}_{*} + \frac{1}{2} m_{\ell} \|\vec{v}_{*}\|^2) \\ &= \left(\sum_{\ell} \mathcal{E}_{(\ell)}^N \right) + (\vec{v}_{*} \cdot \sum_{\ell} \vec{p}_{(\ell)}^N) + \|\vec{v}_{*}\|^2 \sum_{\ell} m_{\ell}.\end{aligned}$$

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 $\mathcal{E}^{N'}$ 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

Once again, Einstein realized that *there is some freedom in how we interpret the conservation laws*. Maybe $\vec{p}^N = m\vec{v}$ and $\mathcal{E}^N = \frac{1}{2}m\|\vec{v}\|^2$ are not the right formulas, and some other formula *would* give conserved quantities.

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**.² 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.³ 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.

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

²In 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.

³[\[T2\]](#) For example, we might guess the correct lagrangian function, then apply Noether’s theorem to it (Chapter 40).

denominator by something that *didn't* transform, then we'd be in a simple situation like that for rotations: t is invariant under rotations, so d/dt doesn't alter the rotational properties of \vec{r} , so velocity transforms linearly, leading us to Equation 31.2.

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 30.6, page 393) is unchanged under Lorentz transformations, so $\Delta\tau = \Delta\tau'$. 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 $d\tau$ is actually positive for any two distinct events on its trajectory. That means that we can integrate $d\tau$ forward along the trajectory to obtain a parameter for the trajectory, called **proper time** τ . That is,⁴ we may consider the time and the spatial position along the particle's trajectory both to be functions of τ .

Working in 1D for simplicity, define

$$\check{p} = m \frac{dx}{d\tau}, \quad \text{relativistic momentum} \quad (31.6)$$

which is a new function defined along the trajectory. In this formula, m is a constant with dimensions of mass, an invariant property of the particle. We'll call it "the mass" of the particle.⁵

We also introduce an analogous quantity

$$K = m \frac{d(ct)}{d\tau}. \quad (31.7)$$

The point of these definitions is that then the pair

$$\begin{bmatrix} K \\ \check{p} \end{bmatrix} = m \frac{d}{d\tau} \begin{bmatrix} ct \\ x \end{bmatrix}$$

has the same simple transformation under Lorentz boosts as do ct and x :

$$\begin{bmatrix} K \\ \check{p} \end{bmatrix} = m \frac{d}{d\tau} \Lambda \begin{bmatrix} ct' \\ x' \end{bmatrix} = \Lambda \left(m \frac{d}{d\tau} \begin{bmatrix} ct' \\ x' \end{bmatrix} \right) = \Lambda \left(m \frac{d}{d\tau'} \begin{bmatrix} ct' \\ x' \end{bmatrix} \right) = \Lambda \begin{bmatrix} K' \\ \check{p}' \end{bmatrix}. \quad (31.8)$$

Here Λ is a 2×2 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 βc relative to the original one. In short, K and \check{p} form a pair with a simple, linear transformation rule—the *same rule* as the one for ct and x .

We now propose two new conservation laws:

$$\check{p}_{(1)} + \check{p}_{(2)} - \check{p}_{(3)} - \check{p}_{(4)} = 0 \quad \text{and} \quad (31.9)$$

$$K_{(1)} + K_{(2)} - K_{(3)} - K_{(4)} = 0, \quad (31.10)$$

⁴See also Problem 31.1. We have previously used ξ to denote a generic parameter along a trajectory; τ 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 394).

⁵Old 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.

which differ from the discredited newtonian versions. Exactly as in the discussion of rotation, we know at once that Equations 31.9–31.10 are automatically Lorentz invariant. Proof: Equation 31.8 is analogous to Equation 31.2, and we can repeat the argument based on Equation 31.3.

That’s remarkable: We still haven’t proposed any detailed dynamical laws for collisions (possibly involving nuclear forces and so on), and yet we *still found a proposal for the corrected form of the momentum* that leads to an acceptable conservation law. Indeed, for a slowly moving particle \vec{p} becomes equal to Newton’s momentum. To see this, note that

$$d\tau = \sqrt{dt^2 - (vdt/c)^2} = dt\sqrt{1 - (v/c)^2} = \gamma^{-1}dt, \quad (31.11)$$

and $\gamma \rightarrow 1$ for a slowly moving particle. Thus, $m(dx/d\tau) \rightarrow m(dx/dt) = p^N$.

What about the new quantity K ? To identify its meaning, note that Equation 31.11 gives $K = mc\gamma \approx mc(1 + \frac{1}{2}(v/c)^2 + \dots)$. So Equation 31.10 multiplied by c says

$$(m_1 + m_2 - m_3 - m_4)c^2 + \mathcal{E}_1^N + \mathcal{E}_2^N - \mathcal{E}_3^N - \mathcal{E}_4^N \approx 0. \quad (31.12)$$

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

$$\boxed{\check{\mathcal{E}} = cK = mc \frac{d(ct)}{d\tau}. \quad \text{relativistic energy}} \quad (31.13)$$

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 $\vec{p} = m \frac{d\vec{r}}{d\tau}$, will henceforth refer only to the relativistic formula, and \mathcal{E} will always mean $mc \frac{d(ct)}{d\tau}$. There won’t be any ambiguity, because from now on we won’t use the newtonian quantities at all. Reinstating the other spatial components gives our proposed conservation law as an equality of 4D vectors:⁶

$$\begin{bmatrix} \mathcal{E}_{(1)}/c \\ \vec{p}_{(1)} \end{bmatrix} + \begin{bmatrix} \mathcal{E}_{(2)}/c \\ \vec{p}_{(2)} \end{bmatrix} - \begin{bmatrix} \mathcal{E}_{(3)}/c \\ \vec{p}_{(3)} \end{bmatrix} - \begin{bmatrix} \mathcal{E}_{(4)}/c \\ \vec{p}_{(4)} \end{bmatrix} = 0. \quad (31.14)$$

31.3.2 What has/has not been shown

We have shown that proposed conservation laws involving two 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.⁷

⁶Chapters 32–33 will christen such quantities **four-vectors**.

⁷Remarkably, today’s Standard Model’s interactions all look a *lot* like electrodynamics.

Instead, following “Einstein thinking,” we will shelve 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 specific context of electrodynamics. Specifically, we 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 Maxwell’s equations and the Lorentz force law.

31.3.3 A geopolitical consequence

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 proves the conservation of energy, Equation 31.5, assuming separate conservation of mass. But we only obtained a single combined law, Equation 31.12 in the newtonian limit. Einstein realized that there was *no fundamental reason why masses must be unchanged*, nor even for total mass to be conserved, 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.⁸ He immediately grasped that even a fraction of a percent change in mass could account for the enormous energies that seemed to come from nowhere in radioactive decay.⁹ Just two years later, Einstein wrote: “Bodies whose energy content is variable to a high degree, for example radium salts,” may perhaps be used to test his prediction about the mass-energy equivalence. Then in a laconic, eerily prescient remark in 1907, he wrote “It is possible that radioactive processes may become known in which a considerably larger percentage of the mass of the initial atom is converted into radiations. . . than is the case for radium.”

The first artificially induced nuclear reactions confirmed Einstein’s rule.

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 ${}^7\text{Li}$ 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 $2.56 \cdot 10^{-29}$ kg, consistent¹⁰ with their estimate of the total gain in kinetic energy of 17.2 MeV.

That is definitely a practical result. 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 at the time how either of these transformations could be done in practice. But within a few decades the outlines began to form. All three belligerents in the second World War embarked on urgent crash programs to develop such weapons. An entire world vanished forever on 16 July, 1945.

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 Rainville et al., 2005.

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 $\beta \rightarrow 1$. In this limit, we expect Newton's formulas to be badly inaccurate. In this situation, Equations 31.6, 31.7 and 31.13 give

$$\frac{p}{\mathcal{E}} = \frac{m \, dx/d\tau}{cm \, d(ct)/d\tau} = c^{-2} \frac{dx}{dt} \rightarrow \frac{1}{c},$$

or

$$\mathcal{E} \approx pc. \quad (31.15)$$

This is precisely the relation that we found earlier for energy and momentum *fluxes* of a classical plane wave solution!¹¹ 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.¹² What was wrong was the expectation that newtonian formulas should apply to things moving at speed c .

You may object that as $\beta \rightarrow 1$, our formula for $\gamma \rightarrow \infty$, 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 $\beta \rightarrow 1$ and $m \rightarrow 0$ in just such a way that $p \rightarrow \text{constant}$:

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. It's no accident that when Einstein was working on his light-quantum hypothesis, he was *also* working out special relativity.¹³

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.

When an electron scatters a photon, the outgoing photon has different frequency from the incoming one.

You may still be bothered, however: *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_{box} . Now imagine filling the

¹¹Equation 20.9 in Section 20.3 (page 289). Section 56.4 will derive this relation for photons by quantizing the field.

¹²See Section 20.3 (page 289).

¹³Section 29.5 already disposed of another objection.

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 $\mathcal{E} = pc$.

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.¹⁴ And now we see that \mathcal{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:

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). Three-tensors have definite, linear transformations under rotations. Every physical quantity in newtonian physics belongs to (is a component of) some 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). Four-tensors 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 extraordinarily 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.

¹⁴Section 30.3.2 (page 391) pointed this out.

PROBLEMS

31.1 Proper time

Section 31.3.1 claimed that the trajectory of any material particle (that is, not a photon) admits a convenient parameterization by an invariant quantity called proper time. 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 ξ . To be physical, the trajectory must always be moving with speed less than c , or in other words $\|\mathrm{d}\vec{r}/\mathrm{d}\xi\|^2 < (c\mathrm{d}t/\mathrm{d}\xi)^2$ everywhere. Show how to obtain a new parameter τ (an increasing function of ξ) that gives the property

$$-(c\mathrm{d}t)^2 + \|\mathrm{d}\vec{r}\|^2 = -c^2(\mathrm{d}\tau)^2.$$

31.2 Recoil

The unstable nuclide $^{60}\mathrm{Co}$ decays in two steps to an excited state $^{57}\mathrm{Fe}^*$, 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 $^{57}\mathrm{Fe}$ to get excited by an incoming photon is equally narrow.

An isolated $^{57}\mathrm{Fe}^*$ 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.

- Find the recoil kinetic energy if the iron nucleus is isolated. The mass of an $^{57}\mathrm{Fe}$ nucleus is 56.9 Da . A convenient definition of the dalton is $1\text{ Da} = 931.5\text{ MeV}/c^2$.
- What is the corresponding fractional reduction of the energy of the photon?
- 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.¹⁵ 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.

- 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

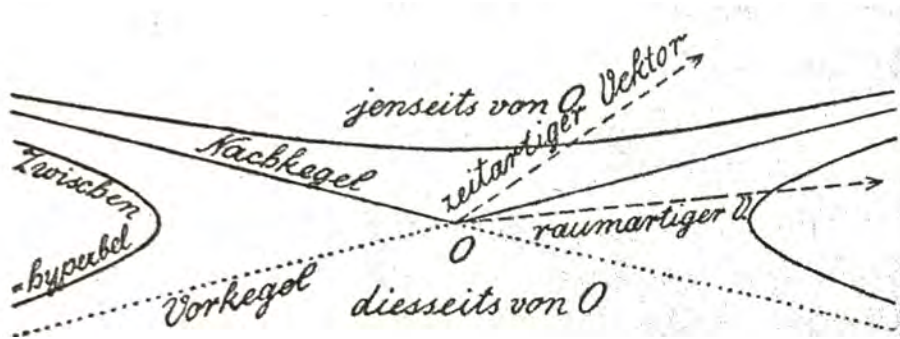
¹⁵In quantum language, there is a significant probability of creating zero phonons (no lattice vibrations) in the transition.

without recoil will be Doppler shifted. How large may v be before those photons can no longer be resonantly absorbed without recoil?

[*Remark:* 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.]

PART V

Relativity: High Tech



Minkowski's diagram for distinguishing different kinds of invariant intervals (*Nachkegel* for the future light cone, *Vorkegel* for the past light cone, *jenseits von O* for the future of O *diesseits von O* for the past of O , *Zwischen hyperbel* "between hyperbola," *zeitartiger Vektor* for time-like vector, and *raumartiger V.* for space-like vector). ["Raum und Zeit" *Physikalische Zeitschrift* **10**:104–111 (1909).]

CHAPTER 32

Four-Vectors

In the fall of 1943 [Julian Boyd asked] Einstein to give the manuscript of the June paper to the Book and Authors War Bond Committee as a contribution to the sale of war bonds. 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.” Sorry if up till now, profs have been withholding this vital information from you on the dubious premise that you’re “not ready yet.” You’re ready.

The next few chapters will rediscover some results already seen in the preceding Parts III–IV. Why repeat?

- 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.
- But the high-tech approach is abstract. 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.

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 system of Maxwell’s 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 character 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, that we can then stick together (following some grammatical Rules) to build equations that are guaranteed to be Lorentz invariant. Then we’ll see if 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 ($d/d(ct)$ and $\vec{\nabla}$) transform differently from coordinates (ct and \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.

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*.¹ 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?

¹Hanging Question #A (page 12). See also Problem 30.3.

- 2 If so, what does that say about the transformation properties of \vec{E} and \vec{B} ?
- 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 corrections will be needed at all.

You could, hypothetically, jump ahead to Equations 33.3 and 33.5 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. 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. Let's review some material introduced in Chapters 13–14.

32.3.1 Rotations preserve the form of the metric

Here are some things we've already discussed. The components of a 3-vector \vec{r} , referred to 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 , where²

$$\vec{r}'_a = S_{ai}\vec{r}_i \quad (\text{and } t = t'). \quad (32.1)$$

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, \dots from the middle of the alphabet for one coordinate system, but a, b, \dots 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 pythagorean formula always has the same form:³

$$\|\vec{r}\|^2 = \vec{r}_i\vec{r}_i = (\|\vec{r}'\|)^2 = \vec{r}'_a\vec{r}'_a = S_{ai}\vec{r}_i S_{aj}\vec{r}_j. \quad (32.2)$$

²This 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.

³This 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 write 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

$$[\vec{r}']^t[\vec{r}'] = [\vec{r}]^t[\mathbf{S}^t\mathbf{S}][\vec{r}]. \quad (32.3)$$

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 $[\vec{r}]^t[\vec{r}]$, for any \vec{r} , only if \mathbf{S} has the property⁴

$$\boxed{[\mathbf{S}^t\mathbf{S}] = \mathbf{1}. \quad \text{transformation between cartesian systems}} \quad (32.4)$$

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 \mathbf{S} . Therefore we expect a family of solutions with $9 - 6 = 3$ parameters—for example, the Euler angles used to specify a rotation.

Here is some mathematical terminology you may hear on the street. A real matrix that satisfies Equation 32.4 is called **orthogonal**. If two matrices \mathbf{S} and \mathbf{T} are both orthogonal, then so is the product \mathbf{ST} (and also \mathbf{S}^{-1}). Thus, orthogonal matrices close into a **group**, a notion previously introduced in Section 30.2. The group of orthogonal 3×3 matrices is sometimes called $O(3)$. Taking the determinant of both sides of Equation 32.4 shows⁵ that $\det \mathbf{S} = \pm 1$.

Rotation matrices have the additional property that $\det \mathbf{S} = +1$; orthogonal matrices without this property include inversion, $\vec{r} = -\vec{r}$, and reflection through a plane. Again, if \mathbf{S} and \mathbf{T} both meet this extra condition, then so will their product, and so will \mathbf{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 specify, and hence 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

$$S_{ai}\vec{v}_i + S_{ai}\vec{w}_i = S_{ai}(\vec{v}_i + \vec{w}_i).$$

Similarly, if we multiply a 3-vector by, say, 2, the result is again a 3-vector.

Now consider Newton's law for an isotropic, harmonic oscillator with viscous friction:

$$m(d^2\vec{r}/dt^2) = -k\vec{r} - \zeta(d\vec{r}/dt). \quad (32.5)$$

⁴This is also Equation 14.4.

⁵This is also Equation 14.5.

Multiply everything from the left by S :

$$S \cdot (m(d^2\vec{r}/dt^2) + k\vec{r} + \zeta(d\vec{r}/dt)) = 0.$$

Now push the constant matrix S inside the derivatives:

$$m(d^2\vec{r}'/dt^2) + k\vec{r}' + \zeta(d\vec{r}'/dt) = 0.$$

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 ζ were all unaffected by the rotation: They are **scalars**, also called **3-tensors of rank zero**. Of these, m , k , and ζ 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).⁶ There is a coordinate system for which every allowed motion is a solution to the equation

$$m(d^2\vec{r}/dt^2) = -\vec{K} \cdot \vec{r}. \quad (32.6)$$

Here \vec{K} is a vector-valued linear function of vectors. Chapter 13 called such a function a tensor, and Section 13.3 (page 190) explained how to represent it via an array of components. Multiply everything from the left by a rotation matrix:

$$\begin{aligned} S \cdot m(d^2\vec{r}/dt^2) &= -S \cdot \vec{K} \cdot \vec{r} = 0 \\ m[d^2\vec{r}'/dt^2] &= -[S\vec{K}S^{-1}][S\vec{r}] = -[S\vec{K}S^{-1}][\vec{r}']. \end{aligned}$$

The new version has the same form as the original equation, albeit with a modified spring constant matrix:

$$\vec{K}'_{ab} = S_{ai}(S^{-1t})_{bj}\vec{K}_{ij}.$$

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 $S^{-1} = S^t$ (Equation 32.4):⁷

$$\vec{K}'_{ab} = S_{ai}S_{bj}\vec{K}_{ij}. \quad (32.7)$$

Any nine-component quantity whose entries transform in this way can specify, and hence represent, a **3-tensor of rank 2**. We say that one copy of S “acts on” each index of \vec{K} .

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 .⁸ More generally, we can define 3-tensors of any rank p : They are represented by collections of 3^p components, with a transformation

⁶Example 3 on page 193 introduced this example.

⁷This is also Equation 14.2.

⁸Thus, the electric quadrupole tensor and the moment of inertia tensor are physical quantities specified by 3-tensors of rank 2.

law involving p copies of \mathbf{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

$$\vec{K} = A\hat{x} \otimes \hat{x} + B\hat{y} \otimes \hat{y} + C\hat{z} \otimes \hat{z},$$

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 as we used for the dyad product (which is the case $p = q = 1$).

32.3.4 Symmetric and antisymmetric 3-tensors

A spring constant tensor has the property that $\vec{K}_{ij} = \vec{K}_{ji}$, or in matrix language $[\vec{K}] = [\vec{K}]^t$. The quadrupole moment tensor from Chapter 3, and the moment of inertia tensor, also have this “symmetric” property.

Your Turn 32A

- Show that if a tensor is symmetric in one cartesian coordinate system, the same will be true after transformation via Equation 32.7.
- Show that the sum of two symmetric tensors of the same rank is itself 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 $\vec{\omega}$ or the magnetic dipole moment tensor $\vec{\Gamma}$.⁹

If a three-tensor \vec{T} is not symmetric (or antisymmetric), 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 \vec{T} ,

$$\vec{T}^{[S]} = \frac{1}{2}(\vec{T} + \vec{T}^t), \quad \vec{T}^{[A]} = \frac{1}{2}(\vec{T} - \vec{T}^t) \quad (32.8)$$

respectively.¹⁰ Then $\vec{T} = \vec{T}^{[S]} + \vec{T}^{[A]}$.

⁹See Sections 15.2 (page 214) and 17.2 (page 243).

¹⁰There are corresponding operations on rank-three three-tensors as well, involving sums over all six permutations of the three indices (Equation 15.10, page 218). However, the totally antisymmetric and totally symmetric parts of a general rank-three tensor do not exhaust its information as in the rank two case.

32.3.5 3D contraction is another invariant operation

The dot product of two vectors may be thought of as the trace of their dyad product, which is a 3-scalar. More generally, the trace of a rank-two tensor is a scalar (Section 13.3.1). Similarly, the dot product, or “contraction,” on the right side of Equation 32.6 “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, \dots, 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. 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,

$$T'_{a_1, \dots, b, \dots, a_p} = S_{a_1 i_1} \cdots \underbrace{S_{b, i_K}} \cdots \underbrace{S_{b, i_L}} \cdots T_{i_1, \dots, i_p}.$$

The factors in braces, summed over b , yield $[S^t S]_{i_K, i_L}$, which is the identity matrix. The result is therefore equal to the contraction of T_{i_1, \dots, 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 for each contraction.

32.4 OTHER ROTATIONALLY INVARIANT SYSTEMS IN MECHANICS

32.4.1 Newtonian gravitation

Here is another example, mentioned in Section 26.5 (page 346): 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:

$$m(d^2 \vec{r}/dt^2) = -\frac{GMm}{r^3} \vec{r}. \quad (32.9)$$

To analyze this equation’s symmetry, begin with the denominator. It involves the function $r = \sqrt{\|\vec{r}\|^2}$, which we saw in Section 32.3 is invariant under rotations about the origin. So the right-hand side of Equation 32.9 is a scalar constant $-GMm$, times a scalar function r^{-3} , times the 3-vector \vec{r} . All together, it’s therefore a 3-vector. Setting it equal to the left side then yields a rotationally-invariant equation, just as in the isotropic harmonic oscillator.

Your Turn 32B

- Equation 32.9 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.
- Show that expanding the scope of the system in this way (acknowledging that \vec{r}_{Sun} is a dynamical variable) also restores explicit translation invariance. This property was hidden in Equation 32.9, 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:¹¹

$$\nabla^2 \phi_N = 4\pi G \rho_m. \quad (32.10)$$

First notice that the chain rule from calculus gives

$$\vec{\nabla}_i \equiv \frac{\partial}{\partial \vec{r}_i} = \frac{\partial \vec{r}'_a}{\partial \vec{r}_i} \frac{\partial}{\partial \vec{r}'_a} = S_{ai} \frac{\partial}{\partial \vec{r}'_a} = (S^t)_{ia} \vec{\nabla}'_a, \text{ or} \quad (32.11)$$

$$\vec{\nabla}'_a = S_{ai} \vec{\nabla}_i. \quad (32.12)$$

We again used the fact that $[S]^{-1t} = [S]$.

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.¹²

From Equation 32.12, you can prove that $\vec{\nabla}_i \vec{\nabla}_i = \vec{\nabla}'_a \vec{\nabla}'_a$, and hence that Equation 32.10 is rotationally invariant if we take G to be a scalar constant, and the mass density ρ_m and the gravitational potential ϕ_N to be scalar fields.

To practice and extend the concepts, consider the velocity vector field of a fluid, $\vec{v}(\vec{r})$. We may be interested in whether the velocity is uniform in space.

Your Turn 32C

Show that

- $\vec{\nabla} \otimes \vec{v}$ is a rank-two tensor field; and
- $\vec{\nabla} \cdot \vec{v}$ is a scalar field, that is, an ordinary function.
- What can we say about the **strain rate tensor**, whose components are $\vec{\nabla}_i \vec{v}_j + \vec{\nabla}_j \vec{v}_i$?

In short, $\vec{\nabla}$ raises the rank of any tensor field by one.

32.5 SUMMARY: THE RULES IN 3D

It's time to state explicitly something that is only implicit in most physics books. Because it's crucial to the general comprehensibility of Physics, we'll dignify it with

¹¹See Chapter 1.

¹²See Section 34.2.1 (page 454). **[T2]** This step also fails if we use curvilinear coordinates, even in 3D euclidean space, but 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.)

the name **Tensor Principle**:¹³

*Physical quantities arrange themselves into 3-tensors (or 3-tensor fields), in some cases constrained by symmetry or anti-symmetry. Physical laws are rotationally invariant, and moreover can be written in **manifestly** invariant form by exploiting simple Rules about tensors.*

3D Tensor
Principle

(32.13)

Idea 32.13 used the standard terminology “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 $\det S = +1$.
- b A 3-tensor field of rank p is the same idea, but each entry is a function of \vec{r} .
- c Permuting the indices on the components of a tensor yields another tensor of the same rank [Section 32.3.4].
- d The sums of corresponding components of two tensors with the same rank, for example, $\vec{A}_{ij} + \vec{B}_{ij}$, yield the components of a new tensor of that same rank [Section 32.3.3]. The correspondence may be nontrivial, for example, $\vec{A}_{ij} + \vec{B}_{ji}$ makes sense. However, every term must have exactly the same set of loose indices; $\vec{A}_i + \vec{B}_j$ 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 $\vec{r} \otimes \vec{r}$ is a rank 2 tensor [Section 32.3.3].
- f **Contraction** (dot product) 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 $\vec{\nabla}$ increases the rank of a tensor field by 1 [Section 32.4.2].
- h A physics equation of the form $A_{i_1, i_2, \dots} = 0$, 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 32.9.
- i The volume element d^3r transforms to d^3r' under rotations¹⁴ because the jacobian matrix has $|\det S| = 1$. Thus, we may convert any tensor field to a constant 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

¹³ **[T2]** Quantum mechanics amends this claim slightly to allow an additional class of quantities called “spinors” (Section 34.11', page 471), but with the proviso that spinor fields are not directly observable.

¹⁴For this reason, we place no arrow over the r .

true tensors; for example, Section 15.2 (page 214) re-expressed magnetic field as an antisymmetric rank-2 tensor field $\vec{\omega}$.

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 E-inertial coordinate system to another, is an obvious property of the equations of motion. The Lorentz transformations modify both 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:

$$\underline{X}^\mu = \left[\begin{array}{c} ct \\ x \\ y \\ z \end{array} \right]^\mu . \quad (32.14)$$

Here μ is an index that runs over the four values $0, \dots, 3$. Note the conventions:

- *Time is regarded as coordinate number zero*, or more precisely, $\underline{X}^0 = ct$.
- *The index indicating which coordinate we're discussing is placed in the upper position*, not lower as we always do in three dimensions. Thus, \underline{X}^1 is the quantity we've been calling x or \vec{r}_1 up till now, and so on.¹⁵ (Lower 4D indices will be given a different meaning later.)
- 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 \underline{X}^1 is a single number, still it's *not* a scalar; it is a component of a 4-vector.¹⁶

As with 3-vectors (Equation 32.3), we will sometimes write $[\underline{X}]$ as an abbreviation for the components \underline{X}^μ regarded as a column (that is, we suppress the explicit index μ) and use the rules of matrix multiplication to imply summations. Then $[\underline{X}]^t$ is the corresponding row vector.

¹⁵How 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 likely denotes a component (and also that an exponent would not make sense).

¹⁶Unlike 3D, when we introduce 4-tensors of higher rank we will use the same underscore for all of them.

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):

$$\underline{X}'^\alpha = \Lambda^\alpha_\mu \underline{X}^\mu. \quad (32.15)$$

More generally, any four-component quantity whose entries transform in the same way as \underline{X} when we change from one E-inertial coordinate system to another can specify, and hence represent, a **4-vector**.

As in 3D, summation over repeated indices (here μ) is implied. As with 3D rotations, the entries of Λ are all *constants*, and so may be pushed past derivatives.¹⁷ For extra clarity, we will often use coordinate indices μ, ν, \dots from the middle of the Greek alphabet for one coordinate system, but α, β, \dots from the start of the alphabet for the alternative coordinate system. Notice a typographic convention: The second index μ on Λ appears to the right of the α 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.

It's convenient to introduce an abbreviation: The **metric** $\underline{g}_{\mu\nu}$ is the matrix of constants¹⁸

$$[\underline{g}] = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \quad (32.16)$$

Analogously to the condition for a rotation (Equation 32.4), let's consider those special matrices Λ with the property that

$$[\Lambda^t \underline{g} \Lambda] = [\underline{g}]. \quad \text{defining property of Lorentz transformation} \quad (32.17)$$

This property is slightly different from Equation 32.4 (page 423), because $[\underline{g}]$ is not the identity matrix.

Your Turn 32D

Confirm that the Lorentz transformations found in Chapter 30 obey Equation 32.17, for example,

$$\text{Boost along } \hat{x}: [\Lambda] = \begin{bmatrix} \gamma & -\gamma\beta & & \\ -\gamma\beta & \gamma & & \\ & & 1 & \\ & & & 1 \end{bmatrix}. \quad \text{Rotation about } \hat{z}: [\Lambda] = \begin{bmatrix} 1 & & & \\ \cos\theta & \sin\theta & & \\ -\sin\theta & \cos\theta & & \\ & & & 1 \end{bmatrix}. \quad (32.18)$$

Any other set of four quantities that transform under Lorentz transformation in the same way are also said to constitute a 4-vector. Later, we'll connect our original method of discovering Lorentz transformations to Equation 32.17 (Section 34.2.2).

Lorentz transformations close into a **group**, sometimes called $O(3,1)$:

¹⁷Also as in 3D, Λ 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.

¹⁸ $[\underline{g}]$ is the same set of numerical values in any inertial coordinate system. It may not be obvious that \underline{g} defined in this way is a 4-tensor, but Section 34.3.3 will show that that is true, just as in three dimensions the collection of nine constants δ_{ij} is a "tensor from Heaven" (Section 14.2.3).

Your Turn 32E

Show that, if Λ_1 and Λ_2 both satisfy Equation 32.17, then so does the product $\Lambda_1\Lambda_2$ (and also the inverse $(\Lambda_1)^{-1}$).

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 disconnected, 3-parameter submanifolds. One of those—the one containing the identity matrix—is the rotation subgroup $SO(3)$.¹⁹ In fact, there are exactly two 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:

$$\underline{g}_{00} = -1 = \Lambda^\mu_{\ 0} \underline{g}_{\mu\nu} \Lambda^\nu_{\ 0} \quad \text{so} \quad (\Lambda^0_{\ 0})^2 = 1 + \sum_{i=1}^3 (\Lambda^i_{\ 0})^2 \geq 1.$$

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, sometimes called $O^+(3,1)$. It in turn contains the still smaller group of transformations that are *both* proper and orthochronous, sometimes called **restricted Lorentz** or $SO^+(3,1)$.

In fact, 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 4×4 matrices equal, so it's ten independent constraints on the sixteen numbers $[\Lambda]$. 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.

Your Turn 32F

For future use, check that Equation 32.17 implies these useful identities:

$$\begin{aligned} [\Lambda^t \underline{g} \Lambda \underline{g}] &= \mathbf{1}, & [\Lambda^t \underline{g}] &= [\Lambda \underline{g}]^{-1}, & [\underline{g} \Lambda] &= [\Lambda^{-1t} \underline{g}] \\ [\underline{g} \Lambda^t \underline{g} \Lambda] &= \mathbf{1}, & [\underline{g} \Lambda^t] &= [\underline{g} \Lambda]^{-1}, & [\underline{g} \Lambda^{-1t}] &= [\Lambda \underline{g}]. \end{aligned} \quad (32.19)$$

[*Hint*: First notice that $[\underline{g}]^2 = \mathbf{1}$ and $[\underline{g}]^t = [\underline{g}]$.]

¹⁹The other sector consists of combined rotation/reflections.

32.6.3 The invariant interval

Lorentz transformations are therefore 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 and is invariant under Lorentz transformation. Consider a particle trajectory as a curve in spacetime. For any two nearby points on that curve, the invariant interval (Equation 30.6, page 393) can be rewritten as

$$c^2 \Delta\tau^2 = -(\Delta\underline{X})^\mu g_{\mu\nu} (\Delta\underline{X})^\nu. \quad (32.20)$$

To show that the invariant interval really is form-invariant under Lorentz transformations, write²⁰

$$(c\Delta\tau')^2 = -[\Lambda\Delta\underline{X}]^t [g] [\Lambda\Delta\underline{X}] = -[\Delta\underline{X}]^t [\Lambda^t g \Lambda] [\Delta\underline{X}] = -[\Delta\underline{X}]^t [g] [\Delta\underline{X}] = (c\Delta\tau)^2. \quad (32.21)$$

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 dimension \mathbb{L} , so $\Delta\tau$ 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 apt²¹ because that coordinate system would also be the rest frame of an inertial observer who runs 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 $\sum_\mu \underline{Y}^\mu \underline{Y}^\mu$, 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 formula²²

$$\|\underline{Y}\|^2 = \underline{Y}^\mu g_{\mu\nu} \underline{Y}^\nu. \quad (32.22)$$

This quantity really does equal $(\underline{Y}')^\alpha g_{\alpha\beta} (\underline{Y}')^\beta$, so it's a 4-scalar. (The proof is the same as in Equation 32.21.) 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 $\underline{Y}^\mu g_{\mu\nu} \underline{Z}^\nu$. It's also a 4-scalar, analogous to the dot product in 3D.

A big difference with ordinary geometry, however, is that we can have $\|\underline{Y}\|^2 = 0$ even if \underline{Y} itself is not zero. A 4-vector with this property is called **lightlike**, because any two points on a light ray's trajectory have such a separation.²³ More generally, if $\Delta\underline{X}$ is the spacetime separation between two events, then we call the three cases

²⁰This step is analogous to Equation 32.2. This analogy is the reason that g is again called the "metric."

²¹In French, "propre" can mean "one's own."

²²Note 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 Y with itself.

²³Some books use the synonym **null** for lightlike.

$\|\Delta\underline{X}\|^2 < 0$, $= 0$, 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.

The locus of all events that are lightlike-separated from \mathbf{P} is called \mathbf{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 $t < t_{\mathbf{P}}$ is the “past light cone of \mathbf{P} ”; the other part is the “future light cone of \mathbf{P} .”

With this terminology, we can partially clear up the ambiguity concerning the sign of $\Delta\tau$: 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:²⁴ $\underline{X} = \underline{\Gamma}(\tau)$. Because the invariant interval is a 4-scalar (Equation 32.21), the operation $d/d\tau$ does not alter the transformation properties of whatever it hits. Thus, the four functions

$$\underline{U}^\mu(\tau) = \frac{d\underline{\Gamma}^\mu}{d\tau} \quad (32.23)$$

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 ξ , then compute $\underline{U} = (d\underline{\Gamma}/d\xi)/(d\tau/d\xi)$.

Your Turn 32G

- a. Show that the 4-velocity always obeys the identity

$$\|\underline{U}(\tau)\|^2 = -c^2. \quad (32.24)$$

- b. Construct a 3D analogy: A curve may be parameterized by arc length s . Then $d\underline{\Gamma}/ds$ 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} :

$$[\underline{\Gamma}(\xi)] = \begin{bmatrix} \xi \\ \beta\xi \\ 0 \\ 0 \end{bmatrix}; \quad \frac{d}{d\xi}[\underline{\Gamma}] = \begin{bmatrix} 1 \\ \beta \\ 0 \\ 0 \end{bmatrix}.$$

²⁴Recall Section 31.3.1 (page 411). In particular, we choose the sign of τ such that larger values correspond to later events in time. 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. [T2] Section 34.7'c will point out that this sign choice complicates the notion of time reversal invariance.

Equation 32.20 gives $d\tau = c^{-1}\sqrt{1-\beta^2}d\xi = (c\gamma)^{-1}d\xi$, where²⁵ $\gamma = (1-\beta^2)^{-1/2}$, and so

$$[\underline{U}] = (d\underline{X}/d\xi)/(d\tau/d\xi) = \begin{bmatrix} \gamma c \\ \gamma\beta c \\ 0 \\ 0 \end{bmatrix}. \quad (32.25)$$

Your Turn 32H

Confirm that Equation 32.24 holds, starting from Equation 32.25.

32.6.6 Summary and first 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 \underline{X} of an event has components \underline{X}^μ with upper index, and hence so does its derivative \underline{U} . The constant matrix \underline{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 \underline{g} matrix. (32.26)

For example, if you forget the $[\underline{g}]$ factor in Equation 32.20, the rule (32.26) 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 $\underline{k}^\mu \underline{g}_{\mu\nu} \underline{X}^\nu$, where the **4-wavevector** is defined as

$$\underline{k}^\mu = \begin{bmatrix} \omega/c \\ \vec{k} \end{bmatrix}^\mu = \begin{bmatrix} \omega/c \\ k_x \\ k_y \\ k_z \end{bmatrix}^\mu. \quad (32.27)$$

The virtue of this reformulation is that it tells how \underline{k} must transform. The invariance of 4D contraction says that

$$\underline{k}'^\alpha \underline{g}_{\alpha\beta} \underline{X}'^\beta = \underline{k}^\mu \underline{g}_{\mu\nu} \underline{X}^\nu, \text{ where } \underline{k}'^\alpha = \Lambda^\alpha_\mu \underline{k}^\mu. \quad (32.28)$$

Thus, the same wave, viewed in the new coordinate system, has a phase function of the same form (that is, linear in \underline{X}) but with modified 4-wavevector. Indeed, \underline{k} transforms as a 4-vector.

Your Turn 32I

- 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).
- Also show that the fact that electromagnetic waves travel at speed c can also be expressed by the even more compact statement $\|\underline{k}\|^2 = 0$.

²⁵We previously obtained this in Equation 31.11 (page 413).

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 momentum²⁶ as

$$\underline{p} = m\underline{U}. \quad \text{four-momentum} \tag{32.29}$$

Thus, \underline{p}^0 is a particle’s energy/ c and \vec{p} is its momentum. The mass m is a 4-scalar, a single number characterizing the particle. Because \underline{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 32.25 and 32.29 gives that $\underline{p}^0 = mc\gamma$ and $\underline{p}^i = m\gamma\vec{v}_i$.

With this definition, Einstein’s proposed conservation law says

$$\sum_{\ell} \underline{p}^{(\ell,\text{in})} = \sum_{\ell} \underline{p}^{(\ell,\text{out})}. \tag{32.30}$$

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 sides transform the same way (as 4-vectors), so Equation 32.30 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.

32.7.1 Beyond $\mathcal{E} = mc^2$

Equations 32.29 and 32.24 imply a relationship between the momentum, energy, and mass of any particle:

$$\|\underline{p}\|^2 = -(mc)^2 \quad \text{or} \quad -(\underline{p}^0)^2 + \underline{p}^i \underline{p}^i = -(mc)^2. \tag{32.31}$$

Our identifications of \underline{p}^0 as a particle’s total \mathcal{E}/c , and the spatial components \underline{p}^i as its momentum, \vec{p}_i , yield the relation

$$\mathcal{E}^2 = (\|\vec{p}\|c)^2 + (mc^2)^2. \tag{32.32}$$

Particle energy and momentum are related, but in a non-newtonian way.

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 32.25 to get $\mathcal{E} \approx mc^2 + \frac{p^2}{2m} + \dots$, approximately a constant plus the newtonian formula, recovering Equation 31.12 (page 413).

²⁶See Section 31.3.1 (page 411).

32.7.2 The 4-momentum of a massless particle is a null 4-vector

There is another interesting limiting case. For a particle moving *fast*, so that $\|\vec{p}c\| \gg mc^2$, we recover $\mathcal{E} \approx \|\vec{p}\|c$ (Equation 31.15, page 415). For the case $m = 0$, this relation is true regardless of the value of momentum:²⁷

$$\mathcal{E} = \|\vec{p}\|c. \quad \text{massless particle} \quad (32.33)$$

32.7.3 de Broglie's prediction for electron wavelength was dictated by Lorentz invariance

We also saw earlier that angular frequency and wavevector can be combined into a quantity that transforms as a 4-vector (Equation 32.28). In order for Einstein's light-quantum proposal $E = \hbar\omega$ to be part of a Lorentz-invariant physical theory, then, it must be one component of a bigger law:

$\underline{p} = \hbar\underline{k}. \quad \text{Einstein relations}$

(32.34)

Indeed, this is no surprise: Certainly light does also have a wavenumber \vec{k} . We already found that Maxwell's equations require $\|\underline{k}\|^2 = 0$, and so Equation 32.34 implies $\|\underline{p}\|^2 = 0$ as well. But that relation is just a concise version of Equation 32.33.

The de Broglie wavelength is related to momentum.

When de Broglie proposed that electrons and other “material” particles also had a dual nature, he didn't need to look hard for a Lorentz-invariant rule describing the wave: *Equation 32.34 is again the only suitable proposal*. Indeed, after substituting the electron mass into Equation 32.32, then Equation 32.34 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.”²⁸

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 Creation. 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 Creation.

Particle creation/destruction; cosmic ray showers; e^+e^- annihilation.

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

²⁷Section 56.4 will derive this relation for photons by quantizing the field.

²⁸See Section 31.3.1 (page 411).

have sufficient energy, then more participants can exit than entered, created from nothing but that energy.²⁹ 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 creation ex nihilo solved that puzzle, and the much more perplexing puzzle of where the electrons emitted in nuclear beta decay were located prior to the reaction.³⁰ Later, as particle accelerators became available, creation ex nihilo was observed even for massive particles, first electrons and then everything else. Even without constructing an accelerator, we can see showers of cosmic rays created in the upper atmosphere from a single energetic incoming particle.

Conversely, an electron and positron can mutually *annihilate*, the key process underlying positron emission tomography (PET). The energy equivalent of their combined masses emerges as light.

FURTHER READING

Note that many authors use a different convention that takes \underline{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, treats \underline{X}^0 as an *imaginary* quantity equal to *ict*. 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.)

Historic: de Broglie, 1923a; de Broglie, 1923b.

PROBLEMS

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 \underline{X}^1 axis of an inertial coordinate system \underline{X}^μ 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.)

²⁹Chapter 56 will develop a framework to describe this phenomenon.

³⁰Confinement 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.

Your trajectory can be written in parametric form: $\underline{\Gamma}(\tau)$, where τ is the time you perceive on the ship.³¹ Section 32.6.5 (page 433) defined four-velocity as $\underline{U} = d\underline{\Gamma}/d\tau$. It will be convenient to define the dimensionless variable $w = \underline{U}^1/c$ and substitute cw for \underline{U}^1 . Equation 32.24 gave a relation that also lets us express \underline{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 just 1.5 times your normal Earth weight.³² Now translate that requirement into a differential equation for $d\underline{U}/d\tau$, as follows.

Consider one moment τ_* along your journey. There is an inertial coordinate system \underline{X}'^{α} in which you are momentarily at rest at τ_* . This is the system obtained by boosting the unprimed system by β_*c where

$$\beta_* = \underline{U}^1(\tau_*)/\underline{U}^0(\tau_*).$$

In it, your velocity at τ_* equals zero, and hence your velocity *near* τ_* 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 τ_* , measured in the primed coordinate system, should equal your weight on Earth, times 1.5, divided by your mass. Call that quantity $a_0 = 1.5(9.8 \text{ m/s}^2)$. Thus, we demand of the trajectory that

$$\left. \frac{d}{d\tau} \left[\frac{\underline{U}'^1}{c^{-1}\underline{U}'^0} \right] \right|_{t'_*} = a_0. \quad (32.35)$$

Now apply the Relativity Strategy,³³ that is, translate Equation 32.35 to the Earth-bound inertial coordinate system. Remember that (i) the Lorentz boost connecting the primed and unprimed systems depends on τ_* , but not on τ (it's not an accelerating system). (ii) Factors like \underline{U}^μ that do depend on τ may be evaluated at τ_* , but only after the time derivatives (if any) have been evaluated.

a. Express Equation 32.35 in terms of the one unknown function $w(\tau)$ and its derivative(s). Specifically show that

$$dw/d\tau|_* = \frac{a_0}{c} \sqrt{1 + w_*^2}. \quad (32.36)$$

b. Equation 32.35, and hence 32.36, must hold at every τ_* along the acceleration part of the trip; that is, it is a differential equation. Solve it for $w(\tau)$ with appropriate initial condition.

c. Integrate your answer to (b) to find the actual trajectory $\underline{\Gamma}(\tau)$.

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 τ_{mid} , 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.

³¹Of course, you won't see the Sun rise and set, but you could measure τ by the growth of your fingernails, or the number of heartbeats, or a clock you carry with you.

³²Don't worry about how the ship is propelled, fuel requirements, and so on!

³³Idea 26.14 (page 351).

- d. Revise your sketch to show the entire journey.
- e. Suppose that your total elapsed time is $2\tau_{\text{mid}} = 1 \text{ year}$. Find the total distance ΔX^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 we, who stayed behind, expect the scouts to return home to us?
- h. Following (g), take the total distance ΔX^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.]

An accelerated round trip alters elapsed time in an objectively measurable way.

CHAPTER 33

The Faraday Tensor

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*; 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 high.

Physical idea: The Lorentz force law must be interpreted 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^p components that transform analogously to Equation 32.7 (page 424) as we change between E-inertial coordinate systems on spacetime can specify, and hence represent, a **4-tensor of rank**¹ $\binom{p}{0}$. For example, if

$$\underline{F}{}^{\alpha\beta} = \Lambda^\alpha{}_\mu \Lambda^\beta{}_\nu \underline{F}{}^{\mu\nu}. \quad (33.1)$$

then \underline{F} has rank $\binom{2}{0}$. (A 4-vector has rank $\binom{1}{0}$.)

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

¹Chapter 34 will justify the elaborate rank notation by extending the definition to rank $\binom{p}{q}$. $\boxed{\mathcal{T}2}$ As in three dimensions, we can instead define tensors without committing to any coordinate system by regarding them as multilinear functions of vectors. This viewpoint works even on curved spacetimes that have no E-inertial coordinates.

²Recall Section 13.3.

33.2.2 An extended Tensor Principle

The preceding construction suggests an upgraded Tensor Principle:³

*Physical quantities all arrange themselves into 4-tensors (or 4-tensor fields), in some cases constrained by symmetry or anti-symmetry. Physical laws are Lorentz invariant, and moreover can be written in **manifestly** invariant form by exploiting simple Rules about tensors.*

4D Tensor
Principle

(33.2)

If we restrict to rotations only, 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 we found that some quantities obey it:

- The mass m of a point particle is a single, Lorentz-invariant quantity—a 4-scalar. Later, we’ll also refer to m as a “4-tensor of rank- $\binom{0}{0}$,” 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 invariant interval between neighboring events is a 4-scalar as well.
- The time and location of an event have been fused into \underline{X} , which we have called a 4-vector. We’ll also refer to it as a 4-tensor of rank $\binom{1}{0}$ 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 \underline{p} , which we saw indeed transforms the same way as \underline{X} and hence is also a 4-vector.⁴
- The angular frequency and wavenumber of a plane wave have been fused into \underline{k} , which again is a 4-vector.⁵

The next section will explore whether the electric and magnetic fields also follow the Tensor Principle. First let’s review how we have already begun to see that 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 434) that the wave equation boils down to $\|\underline{k}\|^2 = 0$, a manifestly invariant condition on \underline{k} . (Section 34.2.2 will return to the wave equation itself.)

Momentum conservation

Chapter 31 gave us a taste of “Einstein thinking”:

³Compare Section 32.5 (page 427). $\boxed{T2}$ Quantum mechanics amends this claim slightly to allow an additional class of quantities called “spinors” (Section 34.11’, page 471), but with the proviso that spinor fields are not directly observable.

⁴See Equation 32.29 (page 435).

⁵See Equation 32.28 (page 434).

- We still expect four conservation laws, even if they're not exactly Newton's.
- 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.
- What could that four-vector be? Newton says that both energy and momentum are proportional to an invariant constant, m , intrinsic to the body in question. The expression $\underline{p} = m\underline{U}$ is a four-vector related to velocity with appropriate units, and hence so is $\underline{p}_{\text{tot}} = \sum_{\ell} \underline{p}(\ell)$.
- The statement $\underline{p}_{\text{tot},\text{in}} = \underline{p}_{\text{tot},\text{out}}$ is therefore a manifestly invariant candidate law.
- The four quantities $\underline{p}_{\text{tot}}^{\mu}$ look like Newton's momentum and (a constant plus) energy, in the case of slowly moving bodies whose masses do not change.
- So our candidate law is plausible. We then found some experimental confirmation for it.

Next steps

"Einstein thinking" proved to be powerful, and quickly came to dominate in the search for other new laws. Next, we'll apply it to rediscover the Lorentz force law.

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:⁶

$$(\text{time rate of change of momentum}) = q(\text{linear function of velocity}),$$

where q is a constant of proportionality intrinsic to a test body.

We can write a formula of this sort involving 4-tensors:

$$\boxed{\frac{d\underline{p}}{d\tau} = q\underline{F}(\underline{U}(\tau)). \quad \text{Lorentz force law, 4-vector}} \quad (33.3)$$

Here q is a 4-scalar constant and $\underline{p}(\tau)$ is its 4-momentum at proper time τ . \underline{F} is a linear function that takes a 4-vector and returns a 4-vector. In three dimensions, such a machine is specified by a rank-two tensor: For example,⁷ the anisotropic spring system studied in Section 32.3.3 had a restoring force given by $-\vec{K} \cdot \vec{r}$. Similarly, a 4-tensor of rank $\binom{2}{0}$ can be used to specify a linear function via

$$\underline{U} \rightarrow \underline{F}(\underline{U}) \quad \text{where} \quad \underline{F}(\underline{U})^{\mu} = \underline{F}^{\mu\nu} g_{\nu\lambda} \underline{U}^{\lambda}. \quad (33.4)$$

We will call $\underline{F}^{\mu\nu}$ the **Faraday tensor**. More precisely, the components $\underline{F}^{\mu\nu}(\underline{X})$ 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, \underline{F} is a 4-tensor *field* of rank $\binom{2}{0}$.

⁶Equations 0.5, page 3 and 15.1 page 214.

⁷Other examples we studied included electric polarizability and the moment of inertia, which are 3-tensors defining linear, vector-valued functions of 3-vectors (Section 13.3.1, page 190).

Your Turn 33A

Show that including the \underline{g} factor in Equation 33.4 guarantees that $\underline{F}(\underline{U})$ is a 4-vector. [Hint: Adapt the derivation that led from Equation 32.17 to 32.21.]

Hence, multiplying $\underline{F}(\underline{U})$ by the 4-scalar q and setting the result equal to the 4-vector $d\underline{p}/d\tau$ constructs an invariant equation of motion (Equation 33.3).

At first, however, 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 \underline{F} appearing in Equation 33.3 seems to have $4 \times 4 = 16$ entries!

To make progress, note that \underline{F} is not entirely free. Equation 33.3 says that it specifies the rate of change in \underline{U} , but \underline{U} cannot change in an arbitrary way: Section 32.6.5 pointed out that always $\|\underline{U}\|^2 = -c^2$, a constant. Thus,

$$\frac{d}{d\tau} (\underline{U}^\mu \underline{g}_{\mu\nu} \underline{U}^\nu) = 0.$$

Using the product rule gives $2\underline{U}^\mu \underline{g}_{\mu\nu} \frac{d\underline{U}^\nu}{d\tau} = 0$. Equations 33.3 and 33.4 then imply

$$(\underline{U}^\mu \underline{g}_{\mu\nu}) \underline{F}^{\nu\lambda} (\underline{g}_{\lambda\xi} \underline{U}^\xi) = 0 \quad \text{for any } \underline{U}.$$

That is, \underline{F} must always give us zero when contracted on each of its indices with the same thing. To guarantee that, we must demand that⁸ \underline{F} be an *antisymmetric* 4-tensor of rank $\binom{2}{0}$. This extra condition is itself Lorentz-invariant.⁹

An antisymmetric 4×4 matrix has just *six* independent entries—just what we need to contain the electric and magnetic fields.

33.3.2 Relate to traditional form

We can give those six entries any names we like. Here are some suggestive names for them:

$$\underline{F}^{\mu\nu} = \left[\begin{array}{c|c} 0 & [\vec{E}]^t/c \\ \hline -[\vec{E}/c] & 2\vec{\omega} \end{array} \right]^{\mu\nu} = \frac{1}{c} \begin{bmatrix} 0 & \vec{E}_x & \vec{E}_y & \vec{E}_z \\ -\vec{E}_x & 0 & \vec{B}_z & -\vec{B}_y \\ -\vec{E}_y & -\vec{B}_z & 0 & \vec{B}_x \\ -\vec{E}_z & \vec{B}_y & -\vec{B}_x & 0 \end{bmatrix}^{\mu\nu}. \quad (33.5)$$

Here the magnetic field tensor $\vec{\omega}$ is defined by Equations 15.2 or 15.3 (page 214) and $\vec{B} = c\vec{\omega}$. Equation 33.5 can be summarized by¹⁰

$$\underline{F}^{0i} = -\underline{F}^{i0} = \vec{E}_i/c \quad \text{and} \quad \underline{F}^{ij} = \varepsilon_{ijk} \vec{B}_k, \quad i, j, k = 1, 2, 3. \quad \text{Faraday tensor} \quad (33.6)$$

⁸The logic is the same as when we interpreted the magnetic field as an antisymmetric 3-tensor (Section 15.2, page 214), because that machine eats a particle's velocity and always yields a force perpendicular to \vec{v} .

⁹The 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 425). 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, up).

¹⁰The identifications in Equation 33.6 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 33.5) 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

$$\frac{d}{d\tau} p^1 = q(\underline{F}^{10} \underline{g}_{00} \underline{U}^0 + \underline{F}^{12} \underline{g}_{22} \underline{U}^2 + \underline{F}^{13} \underline{g}_{33} \underline{U}^3).$$

Use Equations 31.11, 32.25, and 33.5 to find

$$\gamma \frac{d}{dt} p^1 = q(-(\vec{E}_1/c)(-1)(c\gamma) + \vec{B}_3(+1)\gamma \vec{v}_2 - \vec{B}_2(+1)\gamma \vec{v}_3).$$

Canceling the γ 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,

The Lorentz force law, formulated using relativistic momentum, can be compactly stated in 4-tensor form as Equation 33.3. The electric and magnetic fields enter as the components of an antisymmetric rank- $\binom{2}{0}$ 4-tensor via Equation 33.5 or 33.6.

Your Turn 33B

Work out the 0-component of Equation 33.3 in terms of \vec{E} and \vec{B} , and interpret it.

33.3.3 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 say some of those things explicitly one more time.

We imagine some apparatus, with coils, charged plates, whatever, that creates some 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 $\underline{F}^{\mu\nu}$ 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 *measure* the Faraday tensor.

What gave us the right to just declare that $\underline{F}^{01} = \vec{E}_1/c$ and so on? Remember, *names are arbitrary*. We could give all six entries different letters of the alphabet if we wished (as indeed Einstein did). Equation 33.5 just assigns names that clarify the connection to our previous form of the Lorentz force law. What’s important is that we *consistently* use the same names everywhere (for example, rename \vec{E}_1 as $c\underline{F}^{01}$ both in the Lorentz force law and 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.

33.3.4 On beauty

Any physicist will tell you that Equation 33.3 is “beautiful.” What is beauty?

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.

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 coordinate system, and deduce the six functions $\underline{F}^{\mu\nu}$, and your friend observes with a right-handed system, both of you deduce $\underline{F}'^{\alpha\beta}$, then your \underline{F} ’s will be related by the inversion matrix. You will both agree that particle motion is described by Equation 33.3, with the same value of q . We need never introduce “pseudo-tensor” quantities like \vec{B} .

T2 Section 33.3.4' (page 450) muses more on beauty in physics.

33.3.5 Better than beauty: an experimental consequence for cyclotron motion

We have drifted far out into Theoryland. Are there any Electromagnetic Phenomena that can ratify our proposed modification of the Lorentz force law?

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. We can test this modification: When charged particles orbit in a uniform magnetic field (**cyclotron motion**), the naïve form of the Lorentz force law predicts that the orbital period will be independent of energy. The corrected form predicts deviations from this behavior as the particles’ speed approaches c . Not only is this effect seen experimentally; it also imposes an important practical limitation on the design of cyclotron accelerators.¹¹

The constancy of cyclotron frequency breaks down at high velocity.

¹¹You’ll work out more details in Problem 33.3. After numerous false starts, the relativistic prediction was verified in 1914 by G. Neumann.

Your Turn 33C

Work out the correction.

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 constant 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 classic examples, which already 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 the components of \underline{F}' will be given by the matrix product $[\underline{\Lambda} \underline{F} \underline{\Lambda}^t]$, or

$$\begin{bmatrix} \gamma & -\gamma\beta & & \\ -\gamma\beta & \gamma & & \\ & & 1 & \\ & & & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & \vec{B}_3 & 0 \\ 0 & -\vec{B}_3 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \gamma & -\gamma\beta & & \\ -\gamma\beta & \gamma & & \\ & & 1 & \\ & & & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & -\gamma\beta\vec{B}_3 & 0 \\ 0 & 0 & \gamma\vec{B}_3 & 0 \\ \gamma\beta\vec{B}_3 & -\gamma\vec{B}_3 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \quad (33.7)$$

The final expression is again antisymmetric, as it must be. Comparing to Equation 33.6, we read off the primed fields:

$$\vec{B}'_3 = \gamma\vec{B}_3; \quad \vec{E}'_2 = -\gamma\beta c\vec{B}_3. \quad (33.8)$$

The second of Equations 33.8 illustrates the mixing of electric and magnetic fields upon Lorentz boosts anticipated in the Prologue.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 33.8 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 *exactly the same explanation* for the resulting current as when a moving magnet approaches a stationary coil: In both cases an electric field drives charges, resolving the paradox in Hanging Question #A.

That is, \vec{E} and \vec{B} have no separate identities. They are just bits of some bigger, unified object, the Faraday tensor. The situation $\vec{E} = 0$ is *not* a Lorentz-invariant property of a system;¹³ in our example, it was true in our original coordinate system but not in the boosted one.

¹²See Section 0.4.1 (page 11).

¹³Unless *all* fields equal zero, or we boost along the direction of \vec{B} .

In his first relativity paper, Einstein somehow managed to find the right transformations in the ugly, mysterious form Equation 33.8, and show that they were exact invariances of the Lorentz force law and Maxwell's 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 spinors, Yang–Mills theory) and beyond (supersymmetry. . .).¹⁴ 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 $c\beta\hat{x}$. Rather than solve the Maxwell equations with a tricky moving boundary condition, we can apply the Relativity Strategy:¹⁵ First solve them 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.¹⁶ 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.

Your Turn 33D

Apply the appropriate Lorentz transformation to find that at $z = 0$,

$$\vec{E}_x = \frac{\gamma(x - \beta ct)}{(\gamma^2(x - \beta ct)^2 + y^2)^{3/2}} \frac{q}{4\pi\epsilon_0} \quad (33.9)$$

$$\vec{E}_y = \frac{\gamma y}{(\gamma^2(x - \beta ct)^2 + y^2)^{3/2}} \frac{q}{4\pi\epsilon_0}. \quad (33.10)$$

Although you obtained the result in Your Turn 33D by using relativity, they could instead be obtained directly from the (relativistically invariant) Maxwell equations.¹⁷ They are complicated formulas, but note first the ratio

$$\frac{\vec{E}_x(t, \vec{r})}{\vec{E}_y(t, \vec{r})} = \frac{x - \beta ct}{y} = \frac{x - x_*(t)}{y - y_*} \quad \text{where} \quad \begin{bmatrix} x_*(t) \\ y_* \end{bmatrix} = \begin{bmatrix} \beta ct \\ 0 \end{bmatrix}.$$

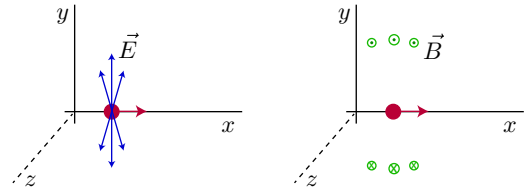
¹⁴There are vistas in Sections 7.5.3, 34.10, 34.11', Chapter 58, and Section 40.4'b (page 534).

¹⁵Idea 33.1 (page 440). Heaviside actually solved this problem without relativity in 1888, but only by a more strenuous effort.

¹⁶Remember that charge is a 4-scalar quantity.

¹⁷Remarkably, Heaviside did this in 1889.

Figure 33.1: Fields in the xy plane created by a point charge in uniform motion. *Left:* The electric fields point away from the particle location at the time of observation. *Right:* The magnetic fields encircle the particle location at the time of observation. Both fields are bunched toward the plane perpendicular to the velocity. See also Problem 33.4.



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:

$$\|\vec{E}\| = r^{-2} \gamma (1 + (\gamma^2 - 1) \cos^2 \theta)^{-3/2} \frac{q}{4\pi\epsilon_0}.$$

This is isotropic when the velocity is small (and thus $\gamma \rightarrow 1$). But for large γ , it is peaked around $\theta = \pi/2$ (the equatorial plane). In short,¹⁸

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 distance to that position squared.

Your Turn 33E

Do a similar calculation to find the \vec{B} field, describe it in words, and relate to Figure 33.1.

33.5 SUMMARY

Table 33.1 shows our constructions so far, highlighting the parallels and contrasts between 3D and 4D concepts.

¹⁸You'll display the field graphically in Problem 33.4.

Table 33.1: Constructions so far, highlighting the parallels and contrasts between 3D and 4D concepts and notation.

3D	4D	ref. (also see Appendix B)
$\vec{r}'_a = S_{ai}\vec{r}_i$ where $S^t = S^{-1}$	$\underline{X}'^\alpha = \Lambda^\alpha_\mu \underline{X}^\mu$ where $[\Lambda]^t[\underline{g}][\Lambda] = \underline{g}$	Eqns. 32.1, 32.4; Eqns. 32.15, 32.17
$\vec{T}'_{ab} = S_{ai}S_{bj}\vec{T}_{ij}$	$\underline{T}'^{\alpha\beta} = \Lambda^\alpha_\mu \Lambda^\beta_\nu \underline{T}^{\mu\nu}$	Eqn. 32.7; Eqn. 33.1
$\ \vec{r}'\ ^2 = \vec{r}_i\vec{r}_i = (\ \vec{r}'\ ')^2 = \vec{r}'_a\vec{r}'_a$	$\ \underline{X}'\ ^2 = \underline{X}'^\mu \underline{g}_{\mu\nu} \underline{X}'^\nu = (\ \underline{X}'\ ')^2 = \underline{X}'^\alpha \underline{g}_{\alpha\beta} \underline{X}'^\beta$	Eqn. 32.2; Eqn. 32.22
$\vec{\nabla}'_a = (([S]^t)^{-1})_{ai}\vec{\nabla}_i = S_{ai}\vec{\nabla}_i$, where $\vec{\nabla}_i = \partial/\partial\vec{r}_i$	$\underline{\partial}'_\alpha = (([\Lambda]^t)^{-1})^\mu_\alpha \underline{\partial}_\mu = [\underline{g}\Lambda\underline{g}]^\mu_\alpha \underline{\partial}_\mu$, where $\underline{\partial}_\mu = \partial/\partial\underline{X}^\mu$	Eqn. 32.12; Eqn. 34.2
Arc length s is a parameter along a curve with $\ \underline{d}\vec{\Gamma}/ds\ ^2 = 1$.	Proper time τ is a parameter along a trajectory with $\ \underline{d}\underline{\Gamma}/d\tau\ ^2 = -c^2$.	Sect. 21.4.4, Your Turn 32G; Eqn. 32.23
Thus, $d\vec{\Gamma}/ds$ is the unit tangent vector along a curve.	Thus, 4-velocity $\underline{U} = d\underline{\Gamma}/d\tau$ obeys $\ \underline{U}\ ^2 = -c^2$.	Sect. 21.4.4, Your Turn 32G; Eqn. 32.24
$\vec{v} \cdot \vec{w} = \vec{v}_i\vec{w}_i = \vec{v}'_a\vec{w}'_a$	$\underline{V}^\mu \underline{g}_{\mu\nu} \underline{W}^\nu = \underline{V}'^\alpha \underline{g}_{\alpha\beta} \underline{W}'^\beta$	Sect. 14.4.1; Sect. 32.6.4
Symmetry and antisymmetry are rotation invariant properties that a tensor may have.	Symmetry and antisymmetry are rotation invariant properties that a tensor may have.	Sect. 32.3.4; Sect. 33.3.1
Contraction reduces rank by 2, for example, $\text{Tr } \vec{T} = \vec{T}_{ii} = \vec{T}'_{aa}$.	Contraction reduces rank by 2, for example, $\underline{T}^{\mu\nu} \underline{g}_{\mu\nu} = \underline{T}'^{\alpha\beta} \underline{g}_{\alpha\beta}$.	Sects. 13.3.1, 32.3.5; Sect. 32.6.4
A rank-2 tensor may be used to express any linear, vector-valued function of a vector via $\vec{g}(\vec{u}) = \vec{\omega} \cdot \vec{u}$.	A rank-2 tensor may be used to express any linear, vector-valued function of a vector via $\underline{G}(\underline{Y})^\mu = \underline{F}^{\mu\nu} \underline{g}_{\nu\lambda} \underline{Y}^\lambda$.	Sect. 13.3.1, Eqn. 15.1; Eqn. 33.4

33.6 PLUS ULTRA

It is hard to overstate the importance of symmetry analysis in physics. *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 *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.

T2 Section 33.6' (page 450) gives some more hints about the Standard Model.

FURTHER READING

Intermediate:

4-tensors as functions of vectors: Thorne & Blandford, 2017, chap. 1.

Historical, on cyclotron motion: Cushing, 1981.

T_2

33.3.4' More on beauty

Is an idea 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.

 T_2

33.6'a Bigger symmetry groups

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 $SU(2) \times U(1)$ respectively), and *all 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. General relativity has proven to be yet more subtle.

PROBLEMS

33.1 *Too much of a good thing?*

This question follows up on Your Turn 33B. Section 33.3.1 proposed the equation of motion

$$\frac{dp^\mu}{d\tau} = qF^\mu{}_\nu U^\nu$$

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 $\vec{\Gamma}(t)$ 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 \vec{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 33C. A proton is released into a region of uniform magnetic field (that is, \vec{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 $\|\vec{B}\|$. Comment on the small- and large- r limits of your answer (at fixed $\|\vec{B}\|$).

33.4 *Uniformly moving charge*

Use a computer to evaluate Equations 33.9–33.10 at time $t = 0$ and display that vector field. That is, find \vec{E} everywhere in space, at one moment of time, for a charged particle in uniform motion along the x axis at speed βc . The formulas show that \vec{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 \vec{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 $\|\vec{E}(x, y)\|$. If it’s not informative, try instead $\log \|\vec{E}(x, y)\|$. (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 $(x^2 + y^2)\|\vec{E}(x, y)\|$. (In this version you won’t need the log trick.)
- Also make a heat map for the new function.

In each graphic, make sure 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 result—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 surface charge density induced on the sphere to lowest order in v/c .

CHAPTER 34

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. Those rules do appear simple and natural in 4-tensor language, but Nature cares little for our aesthetic judgements. Experiments give more compelling foundations to a theory. Section 33.4.1 did find that the transformation we assigned to \vec{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 completely specified the transformation of 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 invariant *under the same field transformations*. 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 the notion of charge flux 4-vector, and add it as a source term in our invariant form of Maxwell's 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 $\underline{\partial}_\mu$ to mean $\partial/\partial X^\mu$, similarly to the notation $\vec{\nabla}_i$ for $\partial/\partial \vec{r}_i$. Then proceeding as in Equation 32.11 (page 427) gives

$$\underline{\partial}_\mu = \frac{\partial}{\partial X^\mu} = \frac{\partial X'^\alpha}{\partial X^\mu} \frac{\partial}{\partial X'^\alpha} = \Lambda^\alpha_\mu \underline{\partial}'_\alpha = [\Lambda^t]_\mu^\alpha \underline{\partial}'_\alpha, \quad \text{or} \quad (34.1)$$

$$\underline{\partial}'_\alpha = [\Lambda^{-1t}]^\mu_\alpha \underline{\partial}_\mu. \quad (34.2)$$

For example, applying both sides of Equation 34.2 to a scalar field tells us that the 4-gradient $\underline{\partial}_\mu \phi$ of a scalar function is a set of four functions with the transformation rule Equation 34.2. The new wrinkle is that this rule is *different from* the one we started with ($X'^\alpha = \Lambda^\alpha_\mu X^\mu$).¹ This issue did not arise in three dimensions, because for rotation matrices $S^{-1t} = S$. But $\Lambda^{-1t} \neq \Lambda$ in general.

Thus, there are *two fundamental tensor types* in relativity: the ones previously called 4-vectors (or 4-tensors of rank $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$), and new ones that transform like Equation 34.2, which are called **4-covectors** (or 4-tensors of rank $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$). The rank notation is motivated by the fact that the 4-gradient $\underline{\partial}_\mu \phi$ has one index in the *lower* position.²

This doubling of index types causes us surprisingly little trouble in practice, however. Suppose that \underline{W}_μ is a collection of four numbers that constitute a 4-covector. Define $\underline{g}^{\mu\nu}$ to be a 4×4 matrix of constants that is the inverse³ of the matrix $\underline{g}_{\mu\nu}$. We now show that the four quantities $\underline{g}^{\mu\nu} \underline{W}_\nu$ are the components of a 4-vector. That is, there is a standard way to interconvert between 4-vectors and 4-covectors; if we like, we can do all of our work using only 4-vectors.

To understand the claimed result, suppose that we have a 4-vector field \underline{A}^μ and a 4-scalar field ϕ . The directional derivative of ϕ along \underline{A} is geometrically defined: At any point, move along \underline{A} and see how ϕ is changing. Thus, we expect that the expression $\underline{A}^\mu \underline{\partial}_\mu \phi$ should be invariant, a new scalar field. Indeed, it can be rewritten as

$$\sum_\alpha (\Lambda^{-1})^\mu_\alpha \underline{A}'^\alpha \frac{\partial X'^\beta}{\partial X^\mu} \frac{\partial \phi}{\partial X'^\beta} = \underline{A}'^\alpha (\Lambda^{-1})^\mu_\alpha \Lambda^\beta_\mu \frac{\partial \phi}{\partial X'^\beta} = \underline{A}'^\alpha \frac{\partial \phi}{\partial X'^\alpha}.$$

Now insert the identity matrix, in the form $[g][g]$, to find that $(\underline{A}^\mu g_{\mu\nu})(g^{\nu\lambda} \underline{\partial}_\lambda \phi)$ is also invariant. But this expression is the invariant inner product of \underline{A} with $g^{\nu\lambda} \underline{\partial}_\lambda \phi$, so we conclude that

$$g^{\nu\lambda} \underline{\partial}_\lambda \phi \text{ are the components of a 4-vector.}$$

The same argument applies to any 4-covector, because by definition they all transform like a gradient:

¹Equation 32.15 (page 430).

²Some books use the term “contravariant vector” for what others abbreviate as “vector,” and “covariant vector” for what others abbreviate as “covector.” But physicists generally can never remember which is co- and which is contra-. To avoid confusion, we will usually use the notation $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$, which says just what it means, namely “one index up, none down,” and so on for other cases.

³In fact, these are two names for the *same* matrix, because $[g]^{-1} = [g]$, but we nevertheless use different notation for the two different uses, in part because they won't be the same in general relativity, nor even in special relativity with curvilinear coordinates.

Your Turn 34A

Derive the last result by substituting the transformation of a general 4-covector \underline{W} into the same formula:

$$\underline{g}^{\alpha\beta}\underline{W}'_{\beta} = [\underline{g}\underline{W}']^{\alpha} = [\underline{g}\Lambda^{-1t}\underline{W}]^{\alpha}.$$

Next, use one of the identities in Equation 32.19 (page 431) to rewrite this expression as

$$= [\Lambda\underline{g}\underline{W}]^{\alpha} = \Lambda^{\alpha}_{\mu}[\underline{g}\underline{W}]^{\mu}.$$

It's traditional to name the four new quantities \underline{W}^{μ} , to emphasize that:

- They are very closely related to \underline{W}_{μ} , and so deserve to be called by the same letter of the alphabet, but
- Unlike \underline{W}_{μ} , they transform like \underline{X}^{μ} (or any other rank- $\binom{1}{0}$ tensor).

The process of constructing a 4-vector from a 4-covector by contraction with \underline{g} is called **index raising**.⁴ Because $[\underline{g}]^2 = \mathbb{1}$, we can invert this operation by *another* multiplication by \underline{g} :

$$\underline{W}_{\mu} = \underline{g}_{\mu\nu}\underline{W}^{\nu}. \quad \text{index lowering}$$

It's not hard to find an invariant product for two covectors: Simply convert each to a 4-vector and use the usual product:

$$(\underline{g}^{\mu\nu}\underline{W}_{\nu})\underline{g}_{\mu\lambda}(\underline{g}^{\lambda\sigma}\underline{V}_{\sigma}) = [\underline{W}^t\underline{g}\underline{g}\underline{V}] = \underline{W}_{\nu}\underline{g}^{\nu\sigma}\underline{V}_{\sigma}.$$

It's easier still to find the invariant product of a covector and a vector:

$$(\underline{g}^{\mu\nu}\underline{W}_{\nu})\underline{g}_{\mu\lambda}\underline{U}^{\lambda} = [\underline{W}]^t[\underline{g}\underline{U}] = \underline{W}_{\nu}\underline{U}^{\nu}.$$

No \underline{g} factor at all is needed in this case.

Ex. Will this elaboration go on forever? Do we need yet another class of objects to describe the transformation of $\partial/\partial\underline{X}_{\mu}$?

Solution: Notice that

$$\frac{\partial}{\partial\underline{X}_{\mu}} = \frac{\partial\underline{X}^{\nu}}{\partial\underline{X}_{\mu}} \frac{\partial}{\partial\underline{X}^{\nu}}$$

The jacobian factor is

$$\frac{\partial}{\partial\underline{X}_{\mu}}(\underline{g}^{\nu\lambda}\underline{X}_{\lambda}) = \underline{g}^{\nu\mu},$$

so $\partial/\partial\underline{X}_{\mu} = \underline{\partial}^{\mu}$, which transforms as a 4-vector—not anything new. Just the two index positions already defined are enough for our needs. The mnemonic is to say that “lower index in the denominator is an upper index,” just as we previously had “upper index in the denominator is a lower index.”

⁴A mathematician might call this operation “taking the adjoint with respect to the inner product \underline{g} .”

34.2.2 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. To define it, it's first convenient to define $\underline{\partial}^\mu$ by raising the index on $\underline{\partial}_\mu$. Then we can construct the Lorentz-invariant operator

$$\square = \underline{\partial}^\mu \underline{\partial}_\mu.$$

It's called the **wave operator**, d'Alembert operator, or **dalembertian**.

Your Turn 34B

Show that \square is the same wave operator that we have been writing all along (Section 25.4, page 332), and whose invariances led us to discover the Lorentz transformations in the first place.

But now we can 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

$$\square \underline{F}^{\mu\nu} \stackrel{?}{=} 0$$

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 Rank

We can now extend the concept introduced in Section 33.2 (page 440): Any quantity with 4^{p+q} components that transform with p copies of Λ and q copies of $(\Lambda^t)^{-1}$ as we change between E-inertial coordinate systems on spacetime can specify, and hence represent, a **4-tensor of rank** $\binom{p}{q}$. 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,

- The gradient of a scalar function has rank $\binom{0}{1}$;
- The Faraday tensor has rank $\binom{2}{0}$;
- The quantities $\underline{F}^{\mu\nu} \underline{g}_{\nu\lambda}$ constitute a 4-tensor of rank $\binom{1}{1}$; and so on.

34.3.2 Symmetry

Let $A^{\mu_1 \dots \mu_p}{}_{\nu_1 \dots \nu_q}$ be a 4-tensor of rank $\binom{p}{q}$.

Your Turn 34C

Show that:

- a. If the components of a tensor \underline{A} are antisymmetric under permutation of some or all of its upper indices in one inertial coordinate system, then \underline{A} will have that same property in any other such system (and similarly for lower indices).⁵ 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.⁶

Your Turn 34D

- a. The metric as we first introduced it, $\underline{g}_{\mu\nu}$, has two lower indices. Prove that this matrix indeed gives the components of a 4-tensor of rank $\binom{0}{2}$, as implied by the notation. [*Hint*: Use an identity from Equation 32.19 (page 431).]
- b. Section 34.2.1 defined the related symbol $\underline{g}^{\mu\nu}$ as the inverse matrix to $\underline{g}_{\mu\nu}$ (and hence, numerically equal to it). Prove that this matrix indeed gives the components of a constant 4-tensor of rank $\binom{2}{0}$, as implied by the notation.

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 to the ones in Section 32.5 (page 427), 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,⁷ where:

- a' A 4-tensor of rank $\binom{p}{q}$ can be represented in a particular inertial coordinate system by a collection of 4^{p+q} numbers, indexed by p upper and q lower indices, with Lorentz transformation law that “acts on” each index in a way appropriate its up/down status (for example, Equations 33.1 and 34.2).

⁵In particular, the statement that a tensor is totally antisymmetric is a Lorentz-invariant property, as we saw in an example already (Equation 33.7, page 446).

⁶See Chapter 14.

⁷See Idea 33.2 (page 441).

- b' A 4-tensor *field* is the same idea, but each entry is a function of \underline{X} .
- c' Permuting a set of indices on the components of a tensor, all in the same position (all up or down) yields another tensor of the same rank (Your Turn 34C).
- d' The sums of corresponding components of two tensors with the same rank yield the components of a new tensor of that same rank.
- e' The collection of all products of the components of a rank- $\binom{p}{q}$ and a rank- $\binom{p'}{q'}$ tensor itself constitutes a rank- $\binom{p+p'}{q+q'}$ tensor. It's called the **tensor product** and sometimes denoted $\underline{A} \otimes \underline{B}$, a generalization of the dyad product.
- f'1 Only contract indices in up/down pairs. Such a contraction is invariant; that is, the result is again a tensor, with reduced rank $\binom{p-1}{q-1}$.
- f'2 Whenever we are tempted to contract two upper indices, we must first lower one of them (introduce a factor of the metric). Index lowering is an invariant operation that changes the rank from $\binom{p}{q}$ to $\binom{p-1}{q+1}$. Then we contract the resulting new lower index with the other upper index, bringing the rank down to $\binom{p-2}{q}$ as desired.
- f'3 Similarly, to contract two lower indices we must first raise one of them. Index raising and lowering are each others' inverse operations.
- g' The derivative operator $\underline{\partial}$ increases the rank of a tensor field by $\binom{0}{1}$ (see Section 34.2.1).
- h' A physics equation of the form $A = B$, where both A and B are tensors (or tensor fields) of the same rank, 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 determinant⁸ ± 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, $\underline{F}^\mu{}_\nu$ and $\underline{F}_{\mu\nu}$; 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 you mean, because they differ by some crucial minus signs. Matrix notation is extremely concise, but for that very reason we will generally 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 look pretty complicated; they have some apparently ad hoc minus signs; we found that \vec{E} and \vec{B} have complicated transformation rules under Lorentz transformations. To see through the derivation, let's 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

⁸Take the determinant of both sides of Equation 32.17 (page 430). For more details, see Section 34.9.3. There is no underscore on the X because d^4X is a 4-scalar.

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- $\binom{2}{0}$ tensor field \underline{F} . In addition, four of them involve charges and currents, while the other four do not. There are also two scalar constants ϵ_0 and μ_0 , or equivalently μ_0 and $c = (\epsilon_0\mu_0)^{-1/2}$.
- What could the equations be? If they take the form (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. 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 (begun in the preceding chapter).

We could implement the first bullet with the candidate equation

$$\partial_\nu \underline{F}^{\mu\sigma} \stackrel{?}{=} 0, \quad (34.3)$$

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 \underline{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:

$$\partial_\nu \underline{F}^{\mu\nu} \stackrel{?}{=} 0. \quad (\text{in vacuum}) \quad (34.4)$$

The Rules say this formula is still Lorentz-invariant, but now it's just *four* equations, because there's one loose index.

Your Turn 34E

- Rephrase Equation 34.4 in terms of the traditional \vec{E} and \vec{B} by using the dictionary in Equation 33.5 (page 443). 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.
- There are three ways to contract two indices in Equation 34.3, and 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 34F

Show that the totally antisymmetric part of Equation 34.3 is⁹

$$\partial_\mu \underline{F}_{\nu\lambda} + \partial_\nu \underline{F}_{\lambda\mu} + \partial_\lambda \underline{F}_{\mu\nu} = 0. \quad (34.5)$$

The Rules say that the left side of Equation 34.5 is a tensor, so the statement that these quantities equal zero is Lorentz invariant, and hence a candidate for a law of Nature.

Equation 34.5 may appear to be $4^3 = 64$ equations, because it has three loose indices. Really, however, most of these equations are vacuous or redundant, because a totally antisymmetric 4-tensor of rank $\binom{0}{3}$ has only four independent components.¹⁰

Your Turn 34G

Write down all four independent components of Equation 34.5. You'll need the expressions obtained by index lowering the identifications we found in Equation 33.5 (page 443):

$$\underline{F}_{\mu\nu} = \begin{bmatrix} 0 & -\vec{E}_1/c & -\vec{E}_2/c & -\vec{E}_3/c \\ \vec{E}_1/c & 0 & \vec{B}_3 & -\vec{B}_2 \\ \vec{E}_2/c & -\vec{B}_3 & 0 & \vec{B}_1 \\ \vec{E}_3/c & \vec{B}_2 & -\vec{B}_1 & 0 \end{bmatrix}_{\mu\nu}. \quad (34.6)$$

Once again, you'll find *precisely* the magnetic Gauss law and Faraday's law—so they, too, were secretly Lorentz-invariant all along.

34.6 THE CHARGE FLUX 4-VECTOR

To complete our job, we need to upgrade Equation 34.4 to include charges and currents. (Equation 34.5 is already complete, because the magnetic Gauss law and Faraday's law don't involve charges nor currents.)

34.6.1 The 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 τ : $\underline{X}_{(\ell)}^\mu = \underline{\Gamma}_{(\ell)}^\mu(\tau)$. Each carries a scalar constant q_ℓ (its charge). As always, we choose an inertial coordinate system on spacetime.

To define charge density at some point \underline{X}_* (an "event"), set up a small spatial volume element $\Delta^3 \underline{X}_\perp$, that is,

$$\begin{aligned} ct &= \underline{X}_*^0 = \text{const}, \\ \underline{X}_*^1 &< x < \underline{X}_*^1 + \Delta \underline{X}^1, \\ \underline{X}_*^2 &< y < \underline{X}_*^2 + \Delta \underline{X}^2, \\ \underline{X}_*^3 &< z < \underline{X}_*^3 + \Delta \underline{X}^3. \end{aligned}$$

⁹As 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 raising the lower one or (as done above) by lowering the upper ones.

¹⁰See Problem 34.1.

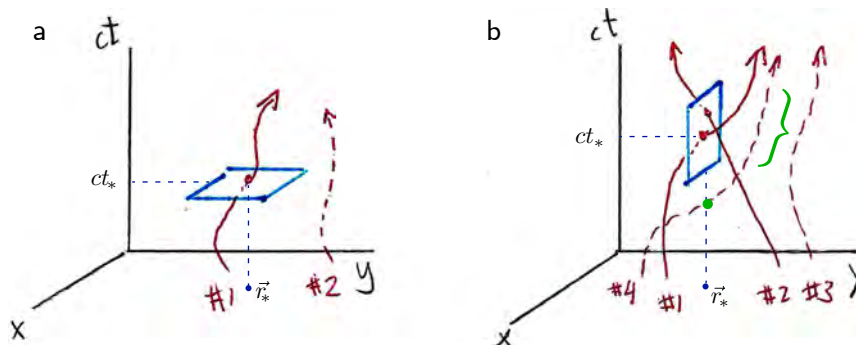


Figure 34.1: Unified construction of (a) charge density and (b) charge flux, an extension of the one in Figure 8.1b (page 113). For artistic reasons, these spacetime diagrams don't show the z direction; the *blue box* is actually a solid 3D region in each panel. Dashed lines indicate charged particle 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_* . Similarly, in (b), trajectory #4 does pass through the selected range of ct and x (*green bracket*), and it also crosses y_* (*green dot*), but no point along it does both. In contrast, *red dots* denote nonzero contributions to the charge density (in (a)) or flux (in (b)).

In Figure 34.1a, the blue rectangle shown represents $\Delta^3 \underline{X}_\perp$. Now we add up all the charges on lines crossing this element from past to future, divide by volume $\Delta^3 \underline{X}_\perp$, multiply by c , and call the result $\underline{J}^0(\underline{X}_*)$. For example, trajectory #1 contributes $cq_1/\Delta^3 \underline{X}_\perp$, whereas trajectory #2, which misses the volume element, contributes nothing.

Note that the quantity \underline{J}^0 just defined has units coul/(s m²). In fact, \underline{J}^0 is the quantity we've previously called $c\rho_q$.

Next, define charge flux at \underline{X}_* by setting up a new small volume element (Figure 34.1b), again called $\Delta^3 \underline{X}_\perp$:

$$\begin{aligned} \underline{X}_*^0 &< ct < \underline{X}_*^0 + \Delta \underline{X}^0, \\ \underline{X}_*^1 &< x < \underline{X}_*^1 + \Delta \underline{X}^1, \\ y &= \underline{X}_*^2 = \text{const}, \\ \underline{X}_*^3 &< z < \underline{X}_*^3 + \Delta \underline{X}^3. \end{aligned} \quad (34.7)$$

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 $\Delta^3 \underline{X}_\perp$, multiply by c , and call the result $\underline{J}^2(\underline{X}_*)$. Thus, in the sketch trajectory #1 contributes $cq_1/\Delta^3 \underline{X}_\perp$, #2 contributes $-cq_2/\Delta^3 \underline{X}_\perp$, and #3–4 contribute nothing.

Define the other two components \underline{J}^1 and \underline{J}^3 similarly. Thus, all four components of \underline{J} have the same units. In fact, \underline{J}^i are the three quantities called the charge flux¹¹ \vec{j}_i in Section 8.3 (page 113). The advantage of the present formulation is that it treats all four components in the same way. In any inertial coordinate system,

$$\begin{aligned} \underline{J}^\mu &= \text{net amount of charge crossing the surface } \{\underline{X}^\mu = \text{constant}\}, \\ &\text{from smaller to larger } \underline{X}^\mu, \text{ per } d^3 X_\perp, \text{ times } c. \end{aligned} \quad (34.8)$$

¹¹And that some books instead call the “current density.”

34.6.2 \underline{J} is a 4-vector

The Tensor Principle claims that all physical quantities can be packaged into 4-tensors.¹² Does \underline{J}^μ defined in the preceding section fit?

Chapter 8 considered a small hypercube and showed that, because charge is locally conserved, we must have

$$\frac{\partial}{\partial t} \rho_{\text{q}} + \vec{\nabla} \cdot \vec{J} = 0. \quad [8.4, \text{page } 114]$$

We now can recognize that this continuity equation can be written more elegantly as

$$\frac{\partial \underline{J}^\mu}{\partial \underline{X}^\mu} = 0, \quad (34.9)$$

or more concisely still as

$$\underline{\partial}_\mu \underline{J}^\mu = 0. \quad \text{Continuity} \quad (34.10)$$

Our derivation of Equation 8.4 was valid in *any* coordinate system, so in particular the form of Equation 34.10 is the same in any inertial system. We also know that $\underline{\partial}_\mu$ form a covector and the index contraction is a Lorentz-invariant operation. Thus, the four quantities

$$\underline{J}(\underline{X}) = \begin{bmatrix} c\rho_{\text{q}}(t, \vec{r}) \\ \vec{j}(t, \vec{r}) \end{bmatrix} \quad (34.11)$$

must themselves transform as a rank- $\binom{1}{0}$ field: the **charge flux 4-vector** field.¹³

34.7 COMPLETE, INVARIANT MAXWELL EQUATIONS

We are now ready to add charges and currents to Equation 34.4. Once again, there's really *no freedom!* The left side of Equation 34.4 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:

$$\underline{\partial}_\nu \underline{F}^{\mu\nu} = \mu_0 \underline{J}^\mu \quad \text{and} \quad \underline{\partial}_\mu \underline{F}_{\nu\lambda} + \underline{\partial}_\nu \underline{F}_{\lambda\mu} + \underline{\partial}_\lambda \underline{F}_{\mu\nu} = 0. \quad \text{Maxwell equations}$$

(34.12)

Your Turn 34H

Extend Your Turn 34G to confirm that the Equation 34.12 really gives the full Maxwell equations as we have been using them.

The eight beautiful¹⁴ new equations, Equations 34.4–34.5, have turned out to be exactly the Maxwell equations we have been using since the Prologue! But their complete

¹²Idea 33.2 (page 441).

¹³See also Section 34.9.3 for a more explicit proof.

¹⁴“Surprising yet inevitable” (Section 33.3.4, page 445).

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 general principles. Moreover, no Levi-Civita tensor appears in Equations 34.12; thus, they are also manifestly invariant under inversions, unlike the traditional formulation in terms of \vec{E} and \vec{B} .¹⁵

T2 Section 34.7' (page 469) 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 34.12, together with the Poincaré lemma,¹⁶ implies that we can always write the Faraday tensor in terms of a **four-vector potential**.¹⁷

$$\underline{F}^{\mu\nu} = \partial^\mu \underline{A}^\nu - \partial^\nu \underline{A}^\mu. \quad (34.13)$$

The Rules imply that \underline{A} must be a four-vector field in order for \underline{F} to be an antisymmetric four-tensor field as desired.

Your Turn 34I

Work out the corresponding \vec{E} and \vec{B} , and show that Equation 34.13 reproduces Equation 18.26 (page 268) when we make the assignments

$$\underline{A}^\mu = \begin{bmatrix} \psi/c \\ \vec{A} \end{bmatrix}^\mu.$$

Thus, the potentials we found long ago also adhere to the 4D Tensor Principle. SI units for the 4-vector potential are $[\underline{A}] \sim \text{kg m}/(\text{coul s})$.

Gauge invariance is the observation that the Faraday tensor doesn't change when we replace \underline{A}^μ by

$$\tilde{\underline{A}}^\mu = \underline{A}^\mu + \partial^\mu \Xi. \quad (34.14)$$

¹⁵This addresses Hanging Question #E. Nor is any choice of right hand buried in the recipe that converted particle trajectories into \underline{J}^μ (Equation 34.8), nor in the one that let us operationally define (measure) \underline{F} (the Lorentz force law, Equation 33.3).

¹⁶See Chapter 15, where we noted that our result holds in any number of dimensions.

¹⁷Chapter 18 already derived this, but in a way that required a tricky insight.

Your Turn 34J

- Prove that last statement starting from Equation 34.13 and connect to Section 18.8.2 (page 268).
- Show that when we substitute Equation 34.13 into Maxwell's equations, one set is vacuous (always automatically satisfied).
- Show that the remaining Maxwell equations become

$$-\square \underline{A}^\nu + \partial_\mu \partial^\nu \underline{A}^\mu = \mu_0 \underline{J}^\nu. \quad (34.15)$$

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 34.12).

T2 Section 34.8.1' (page 470) discusses the counting in more detail, and introduces an extended notion of gauge field.

34.8.2 Particle in uniform motion revisited

For a first look at the benefits of using potentials, we can return to the problem of a charged particle in uniform motion, already solved in Section 33.4.2. Again restrict to the xy plane and suppress the z direction from our notation. Also over the next few lines we'll temporarily drop the tiresome $q/(4\pi\epsilon_0)$ factor. Denote the moving coordinate system with a prime. Then the 4-vector potential seen in the moving coordinate system is just that of a point charge at rest:

$$\underline{A}' = \begin{bmatrix} 1/(cr') \\ \mathbf{0} \end{bmatrix}.$$

So

$$\underline{A} = \Lambda^{-1} \underline{A}' = \frac{1}{c} \begin{bmatrix} \gamma & \beta\gamma & & \\ \beta\gamma & \gamma & & \\ & & 1 & \\ & & & 1 \end{bmatrix} \begin{bmatrix} f(t, x, y) \\ 0 \\ 0 \\ 0 \end{bmatrix} = \frac{1}{c} \begin{bmatrix} \gamma f \\ \gamma\beta f \\ 0 \\ 0 \end{bmatrix}, \quad (34.16)$$

where $f(t, x, y) = 1/r' = (\gamma^2(x - \beta ct)^2 + y^2)^{-1/2}$.

We can now compute the Faraday tensor as usual. For example,

$$\underline{F}^{01} = c^{-1} \vec{E}_x = \partial^0 \underline{A}^1 - \partial^1 \underline{A}^0 = -\frac{\partial}{\partial ct} (c^{-1} \gamma\beta f) - \frac{\partial}{\partial x} (c^{-1} \gamma f) = \frac{\gamma f^3}{c} (x - \beta ct).$$

Reinstating the dropped factor $q/(4\pi\epsilon_0)$ gives again the results found in Your Turn 33D and Your Turn 33E (page 448). However, sometimes \underline{A} is all that's needed, and we see it was easier to obtain than the electric and magnetic fields.

34.9 MORE ABOUT \underline{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, for other purposes it's good to know that another formulation is equivalent to the geometric one.

34.9.1 The delta function composed with an ordinary function is still a delta function

First we need to review a key fact about the delta function.¹⁸ Think of it as a bump, $\delta(x; \sigma) = (2\pi\sigma)^{-1/2} e^{-x^2/(2\sigma^2)}$ with σ very small. So

$$\int_{-\epsilon}^{\epsilon} dx \delta(x; \sigma) \rightarrow 1$$

if we hold ϵ fixed to any positive value and take $\sigma \rightarrow 0$.

Now define a new function $f(x; \sigma) = \delta(2x; \sigma)$ and compute the integral, changing variables to $y = 2x$:

$$\int_{-\epsilon}^{\epsilon} dx f(x; \sigma) = \int_{-2\epsilon}^{2\epsilon} \frac{dy}{2} (2\pi\sigma)^{-1/2} e^{-y^2/(2\sigma^2)} \rightarrow \frac{1}{2}.$$

Again the limit is taken holding ϵ fixed to any positive value and $\sigma \rightarrow 0$. In the same limit, the integral would have been zero had we chosen any range not centered on $x = 0$.

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:

$$\delta(ax) = \frac{1}{|a|}\delta(x) \quad \text{for any constant } a. \quad (34.17)$$

More generally, if $h(x)$ is any smooth function that vanishes at an isolated point x_* , then

$$\delta(h(x)) = \left| \frac{dh}{dx} \Big|_{x_*} \right|^{-1} \delta(x - x_*). \quad (34.18)$$

(Equation 34.17 corresponds to $h(x) = \pm ax$.) If $h(x) = 0$ at several points, then we get the sum of one term for each such point.

34.9.2 \underline{J} may alternatively be formulated in terms of individual trajectories

Here is another set of quantities that may also seem reasonable as a candidate for the current. We will propose it, then show that it's the same as \underline{J} .

Define four functions on spacetime by putting bumps all along each trajectory $\underline{\Gamma}(\ell)$:

$$\underline{J}_{\text{alt}}^{\mu}(\underline{X}) = \sum_{\ell} \int_{-\infty}^{\infty} cd\tau q_{\ell} \underline{U}_{(\ell)}^{\mu}(\tau) \delta^{(4)}(\underline{X} - \underline{\Gamma}_{(\ell)}(\tau)). \quad (34.19)$$

We now want to show that $\underline{J}_{\text{alt}}$ is equal to the \underline{J} defined above. (At least the units match.)

Consider any component of Equation 34.19, for example $\mu = 2$, and any starting point \underline{X} . Thus, we wish to show $\underline{J}_{\text{alt}}^2(\underline{X}) = \underline{J}^2(\underline{X})$. Let \underline{X}_{\perp} denote just the 0, 1, and 3 components (all except the direction 2 that we chose to investigate). As in

¹⁸This was introduced in Section 0.3.8 (page 10).

Equation 34.7, let $\Delta^3 \underline{X}_\perp$ be a small region about \underline{X} obtained by varying everything except \underline{X}^2 . We will now integrate $\underline{J}_{\text{alt}}^2$ and \underline{J}^2 over this region and show that the answers are the same. Because the region was arbitrary, that result will suffice to show that $\underline{J}_{\text{alt}} = \underline{J}$.

Thus, we wish to simplify

$$\int_{\Delta^3 \underline{X}_\perp} d(ct) dx dz \underline{J}_{\text{alt}}^2 = \sum_{\ell} \int_{\Delta^3 \underline{X}_\perp} d(ct) dx dz \underbrace{cd\tau q_{\ell} U_{(\ell)}^2(\tau) \delta(\underline{X}^2 - \Gamma_{(\ell)}^2(\tau)) \delta^{(3)}(\underline{X}_\perp - \Gamma_{(\ell)\perp}(\tau))}_{(34.20)}.$$

The things in the brace don't depend on t , x , or z , so we may bring them to the front:

$$= \sum_{\ell} \int cd\tau q_{\ell} U_{(\ell)}^2(\tau) \delta(\underline{X}^2 - \Gamma_{(\ell)}^2(\tau)) \underbrace{\int_{\Delta^3 \underline{X}_\perp} d(ct) dx dz \delta^{(3)}(\underline{X}_\perp - \Gamma_{(\ell)\perp}(\tau))}_{(34.21)}.$$

The part of this expression in the second brace just gives 1 if particle $\#\ell$'s transverse coordinates fall anywhere inside $\Delta^3 \underline{X}_\perp$ at proper time τ ; otherwise it's zero. That is, as a function of τ it's either zero or a kind of step function.

Now turn to the rest of Equation 34.21. If trajectory $\#\ell$ is ever inside the range $\Delta^3 \underline{X}_\perp$ and crosses the fixed y that we are considering, then let τ_* be the proper time when that crossing occurs.¹⁹ Equation 34.18 gives the integrand in the first brace as

$$cq_{\ell} \frac{d\Gamma_{(\ell)}^2}{d\tau} \left| \frac{d\Gamma_{(\ell)}^2}{d\tau} \right|^{-1} \delta(\tau - \tau_*) = \pm cq_{\ell} \delta(\tau - \tau_*).$$

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 34.20 are those that actually pass through $\Delta^3 \underline{X}_\perp$ at the chosen y . We may thus restrict the sum to only those trajectories, which we denote by \sum'_{ℓ} , and so Equation 34.20 becomes

$$\int_{\Delta^3 \underline{X}_\perp} d(ct) dx dz \underline{J}_{\text{alt}}^2 = c \sum'_{\ell} (\pm q_{\ell}). \quad (34.22)$$

At last we can see that Equation 34.22 is the same property that we used to define the current \underline{J}^2 in Equation 34.8. Repeating the argument for the other three components yields that $\underline{J}_{\text{alt}} = \underline{J}$.

34.9.3 Another proof that \underline{J} is a 4-vector

Before proceeding, let's pause to show that $\delta^{(4)}(\underline{X})$ is a 4-scalar. Suppose that \underline{G}^{α} are a set of functions of \underline{X} that define a new set of coordinates, and that they all vanish at a point \underline{X}_* . We can generalize the derivation that led to Equation 34.18, finding

¹⁹For a small enough region $\Delta^3 \underline{X}_\perp$, there will be at most a single crossing. In Figure 34.1b (page 461), trajectory $\#4$ passes through $\Delta^3 \underline{X}_\perp$, but it's not there when it crosses the chosen y value, so it doesn't contribute to Equation 34.21. Trajectory $\#3$ never visits the chosen y at all.

that²⁰

$$\delta^{(4)}(\underline{G}^\alpha(\underline{X})) = \left| \det \frac{\partial \underline{G}^\alpha}{\partial \underline{X}^\nu} \right|^{-1} \delta^{(4)}(\underline{X}^\mu - \underline{X}_*^\mu). \quad (34.23)$$

For a Lorentz transformation, \underline{G} is a set of four *linear* functions, so the derivatives appearing in Equation 34.23 are a constant matrix, which we have called Λ^α_ν . The determinant of that matrix is ± 1 because $[\Lambda^t \underline{g} \Lambda] = [\underline{g}]$, so Equation 34.23 says $\delta^{(4)}(\underline{X})$ is a 4-scalar.

Now we can use our reformulation of the current (Equation 34.19) to show that \underline{J} is a 4-vector. Indeed, in that equation $d\tau$ is a 4-scalar, the q_ℓ are all 4-scalars, we just showed that the delta function is a 4-scalar, and \underline{U} is a 4-vector (it is the derivative of the 4-vector \underline{X} with respect to the invariant τ).

[T₂] Section 34.9' (page 470) formulates \underline{J} more invariantly.

34.10 A DIZZYING VISTA

Einstein famously said, “If you’re out to describe the truth, leave elegance to the tailor.” Should we care that Equations 34.12 are so beautiful?

One answer is that 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 could be 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 begun in the last few chapters led Einstein in seven more years 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. 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.

Should we try the same thing 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—looks nice. 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 \underline{g} itself be a dynamical object? (If so, then we’ll need to propose some new dynamical law for it!)

²⁰A linear function always has just one zero.

- Do Maxwell’s 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.” Moreover, Einstein found that *again* there is essentially only one acceptable equation of motion that a metric tensor could have.²¹ 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.²² 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), and once again, this vision was borne out.²³

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 Maxwell’s 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.

T2 Section 34.11’ (page 471) outlines the relativistic treatment of spin.

FURTHER READING

Intermediate:

T2 Spinors: Nonrelativistic: Landau & Lifshitz, 1977.

Relativistic: Wess & Bagger, 1992, Appendix A; Dreiner et al., 2010; Weinberg, 2005a.

p-form gauge fields: Weinberg, 2005b, §8.8.

Technical:

T2 Geometric status of the charge flux: Hehl & Obukhov, 2003.

²¹ **T2** Here “essentially” means there’s actually a two-parameter family of equations. One parameter is Newton’s constant, as expected. The other one is the “cosmological constant.” Despite some initial missteps, we now see that this parameter, too, corresponds to physical phenomena that are observed.

²² That is, we may always find normal coordinates (Equation 7.4, page 101).

²³ This thread was the insight needed for Hanging Question #G (page 21).

T₂

34.7'a 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,²⁴ but we found that the system of equations is singular: Two of the eight equations are tautologies, vacuously 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 $\partial_\mu \underline{J}^\mu = 0$ identically. So one combination of these four equations is vacuous.
- Apply $\underline{\varepsilon}^{\mu\nu\lambda\kappa} \partial_\kappa$ to the second set of equations and recall that partial derivatives commute. Here $\underline{\varepsilon}$ is the 4D analog of the Levi-Civita tensor, defined by a choice of orientation on spacetime and $\underline{\varepsilon}_{0123} = +1$. Again, you find that one combination of these four equations is vacuous (and the choice of orientation is immaterial).

A further reduction is possible if we use potentials (Section 34.8.1, page 463).

34.7'b Spatial inversion (parity) invariance

One of our goals is to eliminate the Levi-Civita tensor from all of classical physics (Hanging Question #E, page 13). Chapter 15 advocated rephrasing electrodynamics by replacing \vec{B} by the antisymmetric rank-3 tensor $\vec{\omega}$, and indeed we see that the spatial block of Equation 33.5 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 ($[\Lambda^t g \Lambda] = [g]$ and $\Lambda^0_0 > 0$).

34.7'c Time-reversal invariance

The main text pointed out that some physical laws contain dissipative processes, such as friction or electrical resistance, that violate time-reversal (or simply “T-”) invariance. For this reason, we supposed that a preferred time sense has been specified for spacetime, and we only explored orthochronous Lorentz transformations. However, we can still study specialized systems without dissipation, for example, the dynamics of a small number of point charges in vacuum, and the sense in which such systems have an additional discrete symmetry.

The tricky aspect of this discussion is that time reversal is not merely a Lorentz transformation with $[\Lambda] = \begin{bmatrix} -1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix}$. Indeed, when we apply that active transformation to a parameterized trajectory, the result is *no longer forward-directed in time*, as we have required in Sections 31.3.1 (page 411) and 32.6.5 (page 433). That is, $d\underline{\Gamma}^0/d\tau < 0$. We must instead let

$$\tilde{\underline{\Gamma}}(\tau) = \Lambda \underline{\Gamma}(-\tau),$$

which incurs an additional minus sign in the 4-velocity \underline{U} , and in quantities that involve it.²⁵

Quantities like position \underline{X} , which just suffer an ordinary Lorentz transformation by Λ , will be called “even” under time-reversal; quantities like \underline{U} will be called “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 ($m d^2 \underline{\Gamma} / d\tau^2$) equal to $q \underline{F}(\underline{U})$. Conventionally mass

²⁴Hanging Question #D.

²⁵In quantum physics, T invariance is also treated separately from other symmetries: Unlike other symmetries, it is implemented by an antilinear operator.

and electric charge are T-even scalars. In order for this law to be T-invariant, then, we must declare that the Faraday tensor is T-odd.

We now have no further freedom; we must check that the Maxwell equations are T-invariant with the convention we just pinned down. The homogeneous Maxwell equation (second of Equations 34.12, page 462) is straightforward.

To study the inhomogeneous equation, however, we need to investigate the charge 4-flux $\mu_0 \underline{J}$. Recall that:

- \underline{J}^1 = total charge per $dt dy dz$ crossing $\underline{X}^1 = \text{const}$ from $X^1 < 0$ to $X^1 > 0$, and so on. This definition depends on “from” and “to,” that is, on the sense of parameterization of the trajectories. So J is not a true 4-vector. Similarly,
- \underline{J}^0 = total charge per d^3r crossing $\underline{X}^0 = \text{const}$ from $\underline{X}^0 < 0$ to $\underline{X}^0 > 0$. Here there are two minus signs upon change of time orientation (from \leftrightarrow to as well as exchange of past and future); a true 4-vector’s zero component would have one minus sign.

All told, \underline{J} is T-odd. The inhomogeneous Maxwell equation thus sets a T-odd quantity (contraction of $\underline{\partial} \underline{F}$) equal to a T-odd quantity ($\mu_0 \underline{J}$), so it, too, is a T-invariant equation.

Finally, \underline{F} is built from derivatives of the four-vector potential \underline{A} , so \underline{A} is also T-odd.

T₂

34.8.1’a Counting equations, again

The main text arrived at four equations, Equation 34.15 (page 464). However, one degree of freedom in \underline{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 overconstrained by the four field equations. What rescues the equations is that one combination is vacuously satisfied, as we see by taking the 4-divergence of both sides and using the continuity equation for \underline{J} .

34.8.1’b p -form gauge fields

In the language of Section 15.9’c (page 227), the 4-vector potential is a 1-form field in four dimensions. Its field strengths are given by its exterior derivative. Therefore it is ambiguous; adding the exterior derivative of any 0-form leaves the field strengths unchanged.

Some exotic field theories derived from superstrings involve higher-rank antisymmetric tensor fields called “ p -form gauge fields.” They, too, are subject to the Poincaré lemma, and so can be written as the exterior derivative of a $(p - 1)$ -form potential. These potentials are again ambiguous (gauge invariant), because adding the exterior derivative of any $(p - 2)$ -form to the potential leave its exterior derivative unchanged.

T₂

34.9’ Geometric status of the charge flux

If you know a little differential geometry, then we can give a more general formulation of charge and charge flux, one that does not require any choice of coordinate system at all. The formulation in Section 34.6.1 assumed that an inertial coordinate system had been chosen, so that the 3-volume $\Delta^3 \underline{X}_\perp$ was defined. Now, we will not assume that any inertial coordinate system has been singled out, nor even that such systems exist at all; we only suppose that an orientation on spacetime has been chosen.

We now define a tensor field of rank $\binom{0}{3}$ called $\star \underline{J}$ 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 ϵ . Some charged particle trajectories cross this 3-volume. For each, multiply the charge carried by that particle by ± 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 and divide by ϵ^3 . This quantity is $\star \underline{J}(\underline{U}, \underline{V}, \underline{W})$. It is indeed trilinear and antisymmetric in its three arguments. Section 15.9'c (page 227) called such tensors **3-forms**. $\star \underline{J}$ describes the charges and their trajectories, and depends on the chosen orientation of spacetime, but not on the metric²⁶ nor on any coordinate choice. Indeed, it continues to make sense in curved spacetime.

We can now define the 4-vector field \underline{J} as the contraction of $\star \underline{J}$ with $\underline{\epsilon}$, extending the notion of Hodge dual mentioned in Section 15.9'a (page 226). Because the Levi-Civita 4-tensor depends on a choice of orientation on spacetime, \underline{J} does *not*. Because $\underline{\epsilon}$ involves the metric, so does \underline{J} . In fact, \underline{J} is the 4-vector field defined in the main text.

Maxwell's equations then say

$$d\underline{F} = 0, \quad d \star \underline{F} = \mu_0(\star \underline{J}),$$

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 \underline{F} and $\star \underline{F}$, an “electric–magnetic duality” transformation.

T₂

34.11' 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 to be defined 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 13.4, page 196), 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, $SO(3)$, 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** $SO(3)$. The simplest linear representation is the scalar; for example, charge transforms as $q \rightarrow q$. The next simplest is the vector: $[\vec{v}'] = [S][\vec{v}]$. In fact, any group that is defined as a set of matrices has a natural, or **fundamental**, linear representation, in which group elements act by ordinary matrix multiplication. The components of higher rank tensors form linear representations of the same group 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 $SO(3)$ can be decomposed into blocks obtained in this way.

²⁶Although 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.

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 $SO^+(3,1)$. Hence the components of a 4-tensor in such a system furnish a linear representations of that group, again starting with a fundamental representation we called “4-vector” (rank $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$): $\underline{X}' = \Lambda \underline{X}$. We also found *another* fundamental representation (called “4-covector,” rank $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$), but this was equivalent to a 4-vector via index raising. 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 $SO^+(3,1)$ can be decomposed into blocks obtained in this way.

Quantum mechanics permits a subtle extension to these ideas, in part due to the introduction of wavefunctions with complex values. In the following paragraphs, square brackets will denote complex 2×2 matrices and complex 2-component vectors, with the usual matrix multiplication rules when indices are suppressed. As usual, asterisk will represent complex conjugate. Dagger represents hermitian conjugate: $[M]^\dagger = [M^*]^\dagger$. If the hermitian conjugate equals the inverse, then M is called a **unitary matrix**.

3D

In quantum mechanics, the existence of a symmetry group G only implies that the Hilbert space of states is a representation of an extended form of G . The appropriate extension is called the “covering group.”²⁷ In nonrelativistic quantum mechanics, G is $SO(3)$, whose covering group is $SU(2)$.²⁸

Any group defined as a set of complex matrices has *four* natural, or **fundamental**, linear representations, in which group elements act by ordinary matrix multiplication:

$$[\eta'] = [U]\eta; \quad [\chi'] = [U^*]\chi \quad (34.24)$$

$$[\eta'] = [U^\dagger]^{-1}[\eta]; \quad [\chi'] = [U^{\dagger*}]^{-1}[\chi]. \quad (34.25)$$

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 **3-spinors** of spin rank 1/2. We will now show that the third is equivalent to the first, leaving only one 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

$$\eta'_\alpha = U_\alpha^\beta \eta_\beta, \quad (34.26)$$

whereas the third is distinguished from the first by index position:

$$\eta'^{\alpha} = ((U^\dagger)^{-1})^\alpha_\beta \eta^\beta.$$

We now find two “spin tensors from Heaven”: Let $[\epsilon] = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ and $[\tilde{\epsilon}] = -[\epsilon]$, and more explicitly²⁹

$$\epsilon_{\alpha\beta} = [\epsilon]_{\alpha\beta} \quad \epsilon^{\alpha\beta} = [\tilde{\epsilon}]^{\alpha\beta}. \quad (34.27)$$

²⁷Bargmann, 1954.

²⁸The special unitary 2×2 matrix group $SU(2)$ is sometimes instead called $Spin(3)$ in this context. Spatial inversions must be treated separately.

²⁹We may omit the tilde in the explicit-index notation, relying on index placement to specify which version is meant. But in contrast to $\underline{g}_{\mu\nu}$ and $\underline{g}^{\mu\nu}$ in Section 34.2.1, note that ϵ and $\tilde{\epsilon}$ are *not numerically equal*.

Next, notice that $[\mathbf{U}\epsilon\mathbf{U}^t]$ is again an antisymmetric 2×2 matrix, and hence has only one independent entry. For example,

$$[\mathbf{U}\epsilon\mathbf{U}^t]_{12} = \mathbf{U}_1^1 \epsilon_{12} \mathbf{U}_2^2 + \mathbf{U}_1^2 \epsilon_{21} \mathbf{U}_2^1 = -\det \mathbf{U} = -1.$$

Thus, $[\mathbf{U}\epsilon\mathbf{U}^t] = [\epsilon]$. Because the transpose of a unitary matrix is also unitary, we can equivalently say $[\mathbf{U}^t\epsilon\mathbf{U}] = [\epsilon]$. Changing the sign of both sides also yields $[\mathbf{U}^t\tilde{\epsilon}\mathbf{U}] = [\tilde{\epsilon}]$.

We may build up more complex linear representations by tensor products, possibly (anti)symmetrized. For example, the preceding paragraph established that although $\epsilon_{\alpha\beta}$ 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. Now notice that the first of Equations 34.24 implies that

$$\eta'^{\gamma} \equiv \epsilon^{\gamma\alpha} \eta'_{\alpha} = \epsilon^{\gamma\alpha} \mathbf{U}_{\alpha}^{\beta} \eta_{\beta} = [(\mathbf{U}^t)^{-1}(\mathbf{U}^t\tilde{\epsilon}\mathbf{U}\eta)]^{\gamma} = [(\mathbf{U}^t)^{-1}\tilde{\epsilon}\eta]^{\gamma} = ((\mathbf{U}^t)^{-1})^{\gamma}_{\delta} \eta^{\delta}.$$

In other words, raising a spin index transforms one of our remaining fundamental representations to the other one, much as in ordinary tensor analysis, establishing the claimed equivalence: In 3D, there is only one kind of rank-1/2 spinor. In Pauli’s nonrelativistic theory of the electron, the wavefunction is such a spinor.

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).

We can now set up the correspondence between $\text{SU}(2)$ and rotations. First, note that any real 3-vector \vec{v} corresponds to a traceless hermitian matrix, and vice versa, via $\vec{v} \leftrightarrow [\mathbf{M}] = [\vec{\sigma}] \cdot \vec{v}$, where $[\vec{\sigma}_i]$ are the three **Pauli matrices**:

$$[\vec{\sigma}_1] = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad [\vec{\sigma}_2] = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \quad [\vec{\sigma}_3] = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}. \quad (34.28)$$

Note that then

$$\det [\mathbf{M}] = -\|\vec{v}\|^2.$$

For any special unitary matrix, $[\mathbf{U}\mathbf{M}\mathbf{U}^{\dagger}]$ is traceless and hermitian with the same determinant as \mathbf{M} , and linear in \vec{v} , so it corresponds to a new vector that’s a rotation³⁰ of \vec{v} . This establishes a correspondence between $\text{SU}(2)$ and $\text{SO}(3)$ that preserves the product structures of the groups (and the inverse operation). However, the rotation specified by \mathbf{U} is the same as the one specified by $-\mathbf{U}$, so the correspondence is 2-to-1: $\text{SU}(2)$ double-covers $\text{SO}(3)$.

Pauli’s theory goes 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.

4D

This time, we need the covering group of the proper orthochronous³¹ Lorentz transformations, which we have called $\text{SO}^+(3,1)$. We will see in a moment that the covering group is 2×2 complex matrices, *not* necessarily unitary, but still with determinant one (called³² $\text{SL}(2, \mathbb{C})$).

³⁰Unitary matrices of the form $e^{i\alpha}\mathbf{1}$ would leave \mathbf{M} completely unchanged, but we have eliminated them by restricting to $\text{SU}(2)$.

³¹“Proper” means determinant +1; “orthochronous” means those transformations that do not reverse time; see Section 32.6.2. (Again, inversions must be treated separately.)

³²The special linear group of complex 2×2 matrices, $\text{SL}(2, \mathbb{C})$, is sometimes instead called $\text{Spin}(3,1)$ in this context.

Because we no longer impose the unitarity condition, this group is larger than $SU(2)$, as it must be to accommodate Lorentz boosts.

Again there are four fundamental representations (Equations 34.24–34.25). However, unlike in 3D our transformation matrices are not necessarily unitary, and so we *cannot* immediately conclude that two representations are redundant. We must therefore *keep all* of them, and distinguish them carefully. The traditional notation accomplishes this by dotting indices associated to two of the representations: For example, Equations 34.26 become

$$\eta'_\alpha = W_\alpha^\beta \eta_\beta; \quad \chi'_{\dot{\alpha}} = (W^*)_{\dot{\alpha}}^{\dot{\beta}} \chi_{\dot{\beta}}. \tag{34.29}$$

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 tensors from Heaven: Equation 34.27 is augmented by

$$\epsilon_{\dot{\alpha}\dot{\beta}} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}_{\dot{\alpha}\dot{\beta}} \quad \tilde{\epsilon}^{\dot{\alpha}\dot{\beta}} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}^{\dot{\alpha}\dot{\beta}}. \tag{34.30}$$

Just as we passed from Equation 34.24 to Equation 34.29, we also get two more representations from Equation 34.25, which we again distinguish by index placement. But the ϵ tensors let us raise and lower spin indices as before, without changing their dotted/undotted status. Thus, these two fundamental representations are equivalent to the ones in Equation 34.29.

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 m copies of the undotted representation, combined with the totally symmetric tensor product of n copies of the dotted one (“spin rank $(m/2, n/2)$ ”). Thus, we may construct spinor Rules paralleling those for 4-tensors.

We can now set up the correspondence between $SL(2, \mathbb{C})$ and Lorentz transformations. This time, note that any real 4-vector \underline{X} corresponds to a hermitian matrix (not necessarily traceless)³³ via $\underline{X} \leftrightarrow [M] = -\underline{X}^0 \mathbf{1} + \underline{X}^i [\sigma_i]$. Moreover,

$$\det M = -\|\underline{X}\|^2.$$

Let $[W]$ be any complex matrix with determinant equal to 1. Then $[W M W^\dagger]$ is again hermitian with the same determinant as M , so it corresponds to a new 4-vector that’s a Lorentz transformation of \underline{X} . The correspondence we have set up between $SL(2, \mathbb{C})$ and $SO^+(3, 1)$ preserves 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: $SL(2, \mathbb{C})$ double-covers $SO^+(3, 1)$.

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.³⁴ For example:

- A field with spin rank $(n/2, m/2) = (1/2, 0)$ (a **Weyl spinor**), when quantized, could represent chiral neutrinos.

³³Remarkably, we already made use of this correspondence when constructing the Stokes parameters (Section 24.2.2, page 324).

³⁴As mentioned in Chapter 40, a further twist is that the component fields representing spin 1/2, 3/2, . . . must take their values in an anticommuting number system.

- An electron field can be split into a $(1/2, 0)$ and a $(0, 1/2)$ (a **Dirac spinor**).
- Four-vectors appear as the case $(n/2, m/2) = (1/2, 1/2)$. The sum $n + m$ is an integer, so this is an ordinary representation of Lorentz.
- An antisymmetric rank-2 tensor (such as the Faraday tensor \underline{F}) 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.)

We can now generalize “Einstein thinking” to construct invariant differential equations as candidates for spinor field equations, much as Section 34.5 did for the Faraday tensor.³⁵ Here is one:

$$\underline{\sigma}_{\alpha\dot{\beta}}^{\mu} \cdot \partial_{\mu} \chi^{\dot{\beta}} = 0, \quad \text{Weyl equation} \quad (34.31)$$

where as before $[\underline{\sigma}^0]_{\alpha\dot{\beta}}$ is the unit matrix and $[\underline{\sigma}^i]_{\alpha\dot{\beta}}$ are Pauli matrices. Indeed, quantizing a spinor field that obeys Equation 34.31 yields states describing massless particles of spin $1/2$.

A little more tinkering is needed to accommodate massive particles, such as electrons, because the left side of Equation 34.31 does not transform in the same way as χ . We can do this by introducing a *second* spinor field, η :

$$\underline{\sigma}_{\alpha\dot{\gamma}}^{\mu} \cdot \partial_{\mu} \chi^{\dot{\gamma}} = im\epsilon_{\alpha\beta}\eta^{\beta}; \quad \partial_{\mu}\eta^{\alpha} \underline{\sigma}_{\alpha\dot{\gamma}}^{\mu} = -im\epsilon_{\delta\dot{\beta}}\chi^{\dot{\beta}}. \quad \text{Dirac equations} \quad (34.32)$$

Substituting one of these equations into the other shows that each field then satisfies the **Klein-Gordon equation** $\square\chi = m^2\chi$. Like the Schrödinger equation, this is second-order in space derivatives. Unlike Schrödinger, the K-G equation is relativistically invariant; it is also the appropriate generalization of the wave equation for a field associated to massive particles. The constant m has dimensions of inverse length; it is related to particle mass by the constant \hbar/c .

Opinions vary, but many physicists would say that the Weyl and Dirac equations have the surprising-yet-inevitable quality³⁶ that we prize, to an even greater degree than do the Maxwell equations.

Your Turn 34K

This section may seem to have wandered far from electromagnetism, so suggest a simple modification to Equations 34.32 that incorporate interaction with a 4-vector potential. Why can't you make a similar change to Equation 34.31?

³⁵To be clear, Einstein did not do this; relativistic spinors were developed by B. van der Waerden in 1929 following earlier work by E. Cartan.

³⁶Section 33.3.4 (page 445).

PROBLEMS

34.1 *Tensor types*

An antisymmetric 3-tensor of rank 2 (such as the magnetic dipole moment tensor, Section 17.3.1, page 244) 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 442) pointed out that an antisymmetric 4-tensor of rank $\binom{2}{0}$ (such as \underline{F}) has $(4 \times 3)/2 = 6$ 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 $\binom{2}{0}$ have (such as the energy–momentum flux tensor to be defined in a later chapter) got?
- c. How many independent entries has a totally antisymmetric 3-tensor of rank 3 got? (Equation 15.10, page 218 introduced one such object.)
- d. How many independent entries has a totally antisymmetric 4-tensor of rank $\binom{3}{0}$ got? (Equation 34.5, page 459 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 $\binom{4}{0}$ got?

34.2 *Uniformly moving charge revisited*

A charged point particle moves in a straight line with constant speed v . It creates electric and magnetic fields. Find a covariant expression for the Faraday tensor. That is, your formula should be an antisymmetric rank $\binom{2}{0}$ tensor constructed out of scalars and the four-vectors \underline{U} and \underline{X} using The Rules. Here \underline{X} is displacement from the particle to the observer. Check that your result is equivalent to the ones in Section 33.4.2.

[Hints:

- Sometimes it's easier to start by finding the 4-vector potential, as in Section 34.8.2.
- Again, your result must reduce to Coulomb's law if the particle is at rest in the chosen inertial coordinate system.
- The combinations

$$\underline{K}^{\mu\nu} = (\underline{U}^\mu \underline{X}^\nu - (\mu \Leftrightarrow \nu)) \quad \text{and} \quad \|\underline{K}\|^2 = \underline{K}^{\mu\nu} \underline{K}_{\mu\nu}$$

are useful intermediate building blocks for your answer, because the latter is equal to something useful when computed in the rest frame of the particle.]

34.3 *Bremsstrahlung I*

If you haven't done Problem 34.2 yet, do it first as a warmup.

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 maps out the electric field at time $t_0 > 0$ in the xy plane.

Close to the particle, the observer sees the usual $1/r^2$ field. Section 33.4.2 argued that 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, if it had continued

to move. Section 42.2 argued further that on the boundary between these regions, there is a pulse of radiation (bremsstrahlung). Verify these claims numerically, as follows.

- Express all lengths as dimensionless quantities times ct_0 . Find the region in the xy plane where the observer will see the 4-vector potential of a charge at rest.
- 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.³⁷
- Evaluate the 4-vector potential at each of the grid points satisfying the condition in (a).
- Use ideas from Problem 34.2 to evaluate the 4-vector potential at every grid point *not* satisfying that condition.
- Repeat (a–d) for later time $(1.001)t_0$. Subtract from your previous answer and divide by 0.001 to estimate the time derivative of \vec{A} throughout the xy plane.
- Do whatever else you need to do to find the electric field at time t_0 .
- Make a graphical depiction of the magnitude $\|\vec{E}(t_0, x, y, 0)\|$. If the range of values attained is too large to display properly, compress it by taking a logarithm before making the plot.
- Repeat for speed $0.1c$ and comment.

[Remarks:

- Luckily, \vec{E} points in the xy plane, so a two-dimensional plot is adequate.
- Unluckily, the 4-vector potential field is discontinuous, so you won't get an 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).
- Make 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 x and y axes.
- If you wish, you can compare your result to the more complicated formulas in Your Turn 33D, but that's not the approach you are to use in this problem.
- For the plots in (g,h), you may make a heatmap, a contour plot, or a surface plot. Use your judgement about what is clearest. Why don't you need to know the values of t_0 and q ?]

³⁷Python users will find useful information in the Kinder & Nelson, 2021, §6.4.1, or in the builtin help for `np.meshgrid`.

CHAPTER 35

Energy and Momentum of Fields

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* and include familiar bits corresponding to energy and momentum of point particles. But to get started, we need a good guess for what those quantities 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.

Electromagnetic phenomenon: Superconducting magnets can fail catastrophically.

Physical idea: Even a static magnetic field carries stored 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 the vacuum itself can store energy in static electric and magnetic fields. We need to make that more general and precise.

- Chapter 20 also studied energy and momentum *fluxes* in nonstatic situations. 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 provisional 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 31) 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' $\underline{p}_{(\ell)}^\mu$, 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 mutual long-range forces. Between collisions, each trajectory $\underline{\Gamma}_{(\ell)}(\tau)$ is therefore a straight line, which we parameterize by proper time. Let's suppose that each collision locally conserves energy and momentum, much as we assumed in Chapter 8 that collisions locally conserve electric charge. Analogously to the charge flux 4-vector \underline{J} , define the **energy–momentum flux 4-tensor**² by a recipe analogous to Equation 34.8 (page 461):

$$\begin{aligned} \underline{T}^{\mu\nu} = \text{net amount of } \underline{p}^\nu \text{ crossing the surface } \underline{X}^\mu = \text{constant,} \\ \text{from smaller to larger } \underline{X}^\mu, \text{ per } d^3X_\perp, \text{ times } c. \end{aligned} \quad (35.1)$$

¹Your Turns 27C (page 362) and 27D.

²Often abbreviated “energy–momentum tensor.” Some books call it the “stress-energy tensor” in light of Your Turn 35A.

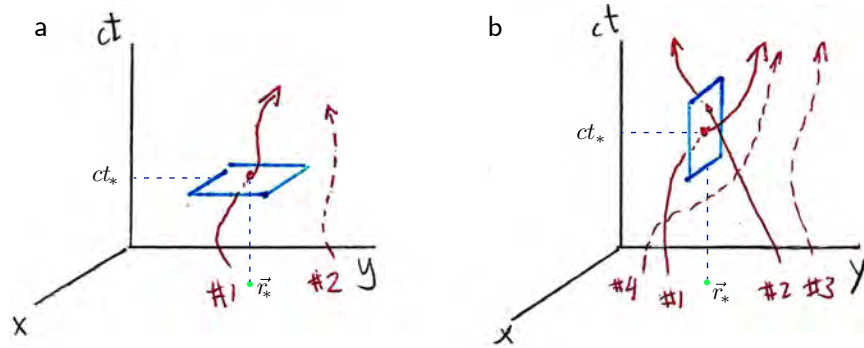


Figure 35.1: Graphical version of Equation 35.1. The figure is the same as Figure 34.1 (page 461). Again the z axis has been suppressed for visualization. (a) Particle #1 contributes to the density of charge or 4-momentum at the point shown (that is, to \underline{J}^0 or $\underline{T}^{0\nu}$ at (ct_*, \vec{r}_*)), whereas #2 does not. (b) Particles #1 and #2 contribute to the net flux of charge or 4-momentum (that is, to \underline{J}^2 or $\underline{T}^{2\nu}$), whereas #3 and #4 do not. #1 crosses from smaller to larger y and hence contributes $q_1/(\Delta x \Delta y \Delta(ct))$. #2 contributes minus its charge.

Your Turn 35A

Using Figure 35.1, convince yourself that

$$\underline{T}^{00} = c \text{ times the density of (energy}/c)$$

$$\underline{T}^{i0} = \text{net flux of (energy}/c)$$

$$\underline{T}^{0k} = c \text{ times density of the } k \text{ component of momentum}$$

$$\underline{T}^{ik} = \text{net flux along } i \text{ direction of the } k \text{ component of momentum.}$$

The last of those results says that the space-space components of \underline{T} constitute a 3D momentum flux tensor (page 194).

We will call the the energy–momentum flux tensor carried by particles $\underline{T}_{\text{part}}$, and write an equivalent formula like the one used for \underline{J} in Equation 34.19 (page 465): Just replace the charge on particle ℓ by the 4-momentum on particle ℓ at proper time τ :

$$\underline{T}^{\mu\nu}(\underline{X}) = \sum_{\ell} \int_{-\infty}^{\infty} c d\tau \underline{p}_{(\ell)}^{\nu}(\tau) \underline{U}_{(\ell)}^{\mu}(\tau) \delta^{(4)}(\underline{X} - \underline{\Gamma}_{(\ell)}(\tau)). \quad (35.2)$$

Your Turn 35B

Convince yourself that \underline{T} is a symmetric 4-tensor of rank $\binom{2}{0}$. Then show that it obeys

$$\frac{\partial}{\partial X^{\mu}} \underline{T}_{\text{part}}^{\mu\nu} = 0. \quad \text{if no long-range forces act} \quad (35.3)$$

That is, if no long-range forces act then $\underline{T}_{\text{part}}$ obeys four continuity equations, expressing the local conservation of each component of the 4-momentum.

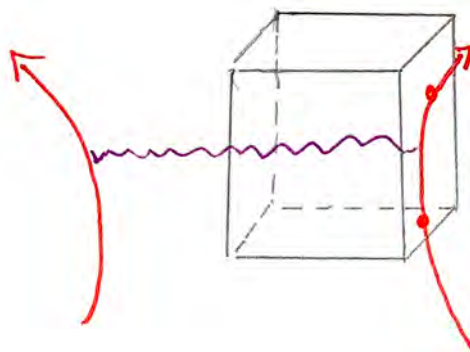


Figure 35.2: **Long-range repulsion** of two particles (*wavy arrow*) would spoil local conservation of particle energy and momentum. Net \vec{p}_2 appears in the small box due to force exerted on particle #2 while it is inside.

35.4 INTERACTIONS SEEM TO SPOIL LOCAL CONSERVATION

35.4.1 Long-range forces

Of course, if some external force acts 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 plus 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 2.6.1 and 18.11 argued *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 8.1). As with electric charge,⁴ that net change will equal $(-c^{-1}\partial_\mu T^{\mu\nu}_{\text{part}})(\Delta^4 X)$, regardless of any collisions among particles in the box (for example, the disintegration shown in the cartoon). Unlike that case, however, this time the net change isn't 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

$$\Delta_{\text{box}} p^\nu = \text{net } \underline{p}^\nu \text{ into 4-box} = - \sum'_\ell \int_{\tau_{\text{in},\ell}}^{\tau_{\text{out},\ell}} d\tau \left. \frac{dp^\nu}{d\tau} \right|_{\text{field}}. \quad (35.4)$$

In this formula, we only include those trajectories that actually enter the box; \sum'_ℓ

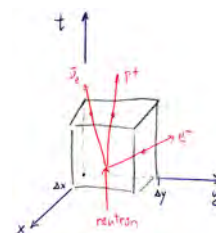


Fig. 8.1 (page 113)

³Also, each gets *not*-opposite amounts of kinetic energy, again seemingly from nowhere.

⁴See Equation 8.3 (page 114).

⁵To 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.

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 τ integral. Finally, we only need to include the contributions to $dp_\ell/d\tau$ 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 last factor in Equation 35.4 to the fields. The formula is cumbersome, however, 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 $1 = \int d^4X \delta^{(4)}(\underline{X} - \underline{X}_*)$ for any point \underline{X}_* in spacetime. For each term ℓ and each value of τ , make the choice $\underline{X}_* = \underline{\Gamma}_{(\ell)}(\tau)$. Then we move the integration over \underline{X} all the way to the left (do it last):

$$\Delta_{\text{box}} \underline{p}^\nu = - \int d^4X \sum'_\ell \int_{\tau_{\text{in},\ell}}^{\tau_{\text{out},\ell}} d\tau \frac{dp_{(\ell)}^\nu}{d\tau} \delta^{(4)}(\underline{X} - \underline{\Gamma}_{(\ell)}(\tau)).$$

This looks like it's making our formula more complicated, but now note what happens if we restrict the \underline{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 τ values for which a trajectory lies inside the box, so we don't need to restrict the τ integral either. Using that insight, and the Lorentz force law (Equation 33.3, page 442), gives

$$\Delta_{\text{box}} \underline{p}^\nu = - \int_{\text{box}} d^4X \sum_\ell \int_{-\infty}^{\infty} d\tau q_\ell \underbrace{F^{\nu\lambda}(\underline{\Gamma}_{(\ell)}(\tau)) U_{(\ell),\lambda}(\tau)}_{\delta^{(4)}(\underline{X} - \underline{\Gamma}_{(\ell)}(\tau))} \delta^{(4)}(\underline{X} - \underline{\Gamma}_{(\ell)}(\tau)).$$

Use the delta-function to re-express the factor in the brace as $F^{\nu\lambda}(\underline{X})$, and then push it to the left, outside of the τ integral. What remains is just c^{-1} times the electric charge flux four-vector (Equation 34.19, page 465):

$$\Delta_{\text{box}} \underline{p}^\nu = -c^{-1} \int_{\text{box}} d^4X \underline{F}^{\nu\lambda}(\underline{X}) \underline{J}_\lambda(\underline{X}). \quad (35.5)$$

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 $\Delta^4 X$ times the integrand. But Section 35.4.1 argued that this change is also $c^{-1} \Delta^4 X$ times minus the 4-divergence of \underline{T} , or

$$\partial_\mu \underline{T}_{\text{part}}^{\mu\nu} = \underline{F}^{\nu\lambda} \underline{J}_\lambda. \quad (35.6)$$

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 *fields*, with the properties that:

- $\underline{T}_{\text{field}}^{\mu\nu}$ is a symmetric 4-tensor given by a local expression in the fields; and
- $\partial_\mu(\underline{T}_{\text{part}}^{\mu\nu} + \underline{T}_{\text{field}}^{\mu\nu}) = 0$.

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* $\underline{T}^{\mu\nu}$ 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.⁶

Equation 35.6 shows what we need:

$$\partial_\mu \underline{T}_{\text{field}}^{\mu\nu} = -\underline{F}^{\nu\lambda} \underline{J}_\lambda. \quad (35.7)$$

But we can't prove this until we guess the correct formula for $\underline{T}_{\text{field}}^{\mu\nu}$!

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 ($\propto \vec{E}^2$). And stored magnetic energy is also a quadratic function of magnetic field, with no derivatives ($\propto \vec{B}^2$). Can we write any such expression that is a symmetric, rank- $\binom{2}{0}$ 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:

$$\underline{T}_{\text{field}}^{\mu\nu} = \alpha \underline{F}^{\mu\sigma} \underline{F}_\sigma{}^\nu + \beta g^{\mu\nu} \underline{F}_{\sigma\lambda} \underline{F}^{\sigma\lambda}. \quad \text{provisional formula} \quad (35.8)$$

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 α and β yet, but already we've made a huge simplification: Just those *two numbers* is all the freedom we have to construct a suitable tensor.

We now take the 4-divergence of our provisional formula:

$$\partial_\mu (\alpha \underline{F}^{\mu\sigma} \underline{F}_\sigma{}^\nu + \beta g^{\mu\nu} \underline{F}_{\sigma\lambda} \underline{F}^{\sigma\lambda}).$$

Use the fact that the fields obey Maxwell's equations, specifically the first of Equations 34.12 (page 462):

$$= \alpha \underbrace{(-\mu_0 \underline{J}^\sigma \underline{F}_\sigma{}^\nu)} + \underbrace{\underline{F}^{\mu\sigma} \partial_\mu \underline{F}_\sigma{}^\nu}_{\text{zero}} + \beta \underline{g}^{\mu\nu} 2(\partial_\mu \underline{F}_{\sigma\lambda}) \underline{F}^{\sigma\lambda}. \quad (35.9)$$

The first term (first brace) is just what we want! Just choose the value $\alpha = -\mu_0^{-1}$ and we get Equation 35.7.

⁶Sometimes called "Poynting's theorem," although independently codiscovered by Heaviside.

We are left with the unwanted other terms (second brace). Can we choose a value of β such that these terms cancel each other identically? That is, can we ensure that

$$0 \stackrel{?}{=} \frac{\alpha}{\beta} \underline{F}^{\mu\sigma} \underline{\partial}_\mu \underline{F}_{\sigma\nu} + 2 \underline{F}^{\sigma\lambda} \underbrace{\underline{\partial}_\nu \underline{F}_{\sigma\lambda}} \quad ? \quad (35.10)$$

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 \underline{F} equals zero.⁷ Specifically, the quantity enclosed by the brace in Equation 35.10 equals

$$\underbrace{-\underline{\partial}_\sigma \underline{F}_{\lambda\nu}} - \underline{\partial}_\lambda \underline{F}_{\nu\sigma}.$$

In Equation 35.10, this tensor is contracted on $\sigma\lambda$ with something antisymmetric, so we may replace its first term (in the brace) by $+\underline{\partial}_\lambda \underline{F}_{\sigma\nu}$. Then Equation 35.10 becomes

$$0 \stackrel{?}{=} \frac{\alpha}{\beta} \underline{F}^{\mu\sigma} \underline{\partial}_\mu \underline{F}_{\sigma\nu} + 2 \underline{F}^{\sigma\lambda} (\underline{\partial}_\lambda \underline{F}_{\sigma\nu} - \underline{\partial}_\lambda \underline{F}_{\nu\sigma}) \quad (35.11)$$

$$= \left(\frac{\alpha}{\beta} - 4\right) \underline{F}^{\mu\sigma} \underline{\partial}_\mu \underline{F}_{\sigma\nu}. \quad (35.12)$$

This will be identically true if we choose $\beta = \alpha/4$.

Substituting the values we found for α, β into Equation 35.8, we conclude that

$$\underline{T}_{\text{field}}^{\mu\nu} = -(\mu_0)^{-1} \left(\underline{F}^{\mu\sigma} \underline{F}_{\sigma}{}^{\nu} + \frac{1}{4} \underline{g}^{\mu\nu} \underline{F}_{\sigma\lambda} \underline{F}^{\sigma\lambda} \right).$$

energy–momentum flux tensor
of the electromagnetic field

(35.13)

This choice meets all the criteria listed at the start of this section.

Your Turn 35C

- a. Confirm that the 00 component (energy density), when written in terms of \vec{E} and \vec{B} , has the form that you expect from Sections 6.3 (page 75) and 18.3.4 (page 260).
- b. Then show that its $i0$ components (flux of energy, or density of momentum) also have a form anticipated in Section 20.4 (page 289).
- c. The ij components may be new to you; they are interesting, too, so work them out and interpret in terms of radiation pressure (Section 20.2.3, page 288).

Thus, the formulas that we already guessed for field energy density and Poynting vector need *no corrections* to account for relativity.⁸

35.6 WHAT HAS BEEN ACCOMPLISHED

⁷See the second of Equations 34.12 (page 462).

⁸Beware 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.”

35.6.1 Poynting's theorem fits with older ideas

We can now get global conservation laws in the usual way, by integrating $\underline{T}^{0\mu}$ over space (see Equation 8.6, page 114).

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:⁹

$$\partial_{\mu}(\underline{T}_{\text{part}}^{\mu\nu} + \underline{T}_{\text{field}}^{\mu\nu}) = 0. \quad \text{Poynting theorem}$$

Take a moment to appreciate how surprising this formula is, and how it vindicates Michael Faraday's intuitions:

- $\underline{T}_{\text{part}}^{\mu 0}$ 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.
- $\underline{T}_{\text{field}}^{\mu 0}$ 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 to overcome that force is preserved as potential energy, and can be recovered by later releasing the particles (letting them fly away from each other). Implicitly the framework is that #1 acts at a distance on #2, creating a potential energy well for it, a bit similarly to if there were a mechanical spring between them that is being 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 31).

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.

To see how the older viewpoint could be compatible with the new one in a concrete setting, let's return to two charged particles at rest. We know the electric potential everywhere from Coulomb's law and superposition, so we have

$$\frac{\epsilon_0}{2} \int d^3r \|\vec{\nabla}\psi\|^2 = -\frac{\epsilon_0}{2} \int d^3r \psi \nabla^2 \psi. \quad (35.14)$$

But $\nabla^2 \psi = \rho_q / \epsilon_0 = q(\delta^{(3)}(\vec{r} - a\hat{z}/2) + \delta^{(3)}(\vec{r} + a\hat{z}/2)) / \epsilon_0$. Multiplying this expression by ψ yields four terms on the left side of Equation 35.14. Two of those terms are

⁹This was Hanging Question #H.

infinite, but constant (self-energy is independent of a). The other two give

$$\frac{1}{2} \frac{q^2}{4\pi\epsilon_0} (a^{-1} + a^{-1}),$$

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 35C), 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 could be recovered by discharging it, *suggesting* that meantime that work 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, *suggesting* that work must be stored. And Section 20.4 argued that light imparts energy, *suggesting* that this energy “must” be transferred out of the light. Now we have *proved* local conservation as a property of Maxwell’s equations and the Lorentz force law.
- Previously we didn’t show that our expressions had the appropriate Lorentz transformation properties. Now it’s obvious because we followed The Rules.
- Previously we only got expressions for energy and momentum flux in plane waves, and we didn’t find the correct prefactor. Now we have complete and general formulas.
- Finally, the same derivation will also give us an analogous theorem when we later add media in Chapter 52.

T2 Section 35.6.1’ (page 487) 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 $\partial_\mu T_{\text{tot}}^{\mu\nu} = 0$, so what has been proved? But it was highly nontrivial that any such formula could be written at all. The only cookery allowed was the choice of *two constants*, α and β , but the theorem we proved was that *four functions* of space and time are everywhere zero.

It may also seem 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 Maxwell’s equations.

FURTHER READING

Intermediate:

T2 Angular momentum flux tensor: Zangwill, 2013, §22.7.4; Weinberg, 1972, p. 46.

T₂

35.5' Angular momentum flux tensor

Any first-year physics book claims to prove that for an isolated system of particles, overall angular momentum conservation 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- $\binom{3}{0}$ tensor

$$\underline{M}^{\mu\nu\lambda} = \underline{X}^\nu \underline{T}^{\lambda\mu} - \underline{X}^\lambda \underline{T}^{\nu\mu}.$$

It's antisymmetric on the last two indices, and you can readily show that $\partial_\mu \underline{M}^{\mu\nu\lambda} = 0$. Thus, we get six densities by taking $\mu = 0$, leading to six conserved quantities

$$\underline{L}^{\nu\lambda} = \int d^3r \underline{M}^{0\nu\lambda}(t, \vec{r}).$$

The spatial bits of this tensor, \underline{L}^{ij} , 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 \underline{T} . So it's appropriate to call \underline{M} the **angular momentum flux tensor**.

We may nevertheless worry that this orbital decay phenomenon dictates an “arrow of time,” despite the fact that electrodynamics is time-reversal invariant (its equations have no dissipative element). But on the contrary, two orbiting charges can also *gain* energy and angular momentum by interaction with an incoming, circularly-polarized, wave.

Even if the two particles are uncharged, for example, two neutron stars, they will still be gravitationally attracted, and their acceleration can generate gravitational radiation, leading to a decaying orbit and ultimately merger. Gravitational radiation with the time course characteristic of this process was detected on Earth by the LIGO-Virgo collaborations in 2016.

The orbit of a binary neutron star system decays via quadrupole gravitational radiation.

T₂

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.

Resistors

If $\int d^3r (\underline{T}_{\text{part}}^{00} + \underline{T}_{\text{field}}^{00})$ 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

An idealized battery really does require that we acknowledge an additional contribution to the energy density, from the chemical energy of its ingredients. That energy is ultimately quantum

mechanical in origin, and hence outside the domain of classical electromagnetism, but still we fit it into our framework in Section 10.3.4'f (page 150). 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).

PROBLEMS

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).

Strong electromagnets can explode.

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. 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 $(56 \text{ mm})^2$.

- 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 σ 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 .

- 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 In-



Figure 35.3: See Problem 35.1. [Image from press.web.cern.ch/press/PressReleases/Releases2008/PR14.08E.html .

sightful Comment.

35.3 *Fine point — energy and momentum of fields*

The main text stated that the expression

$$2F_{\lambda\sigma}\partial_{\sigma}F^{\lambda}_{\mu}$$

could be replaced by

$$-2F_{\lambda\sigma}\partial^{\lambda}F^{\sigma\mu},$$

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 289).

35.5 $\boxed{T_2}$ *Angular momentum flux*

[Not ready yet.]

CHAPTER 36

Vista: 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—but 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: TOOLKIT

Starting in 1821, 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 charges and magnet poles. The magnitude of the field increased as the lines were compressed laterally. The lines of force were under tension, like stretched rubber bands, yet repelled nearby lines with a transverse pressure-like force. This transverse pressure made the lines want to avoid each other, so they spread as they left a point charge; then the connection to the density of field lines gave rise to the $1/r^2$ law.

It sounds crazy! Even decades later, the Continental *philosophes* were particularly severe on Faraday and his successors. And yet Faraday, with practically no formal

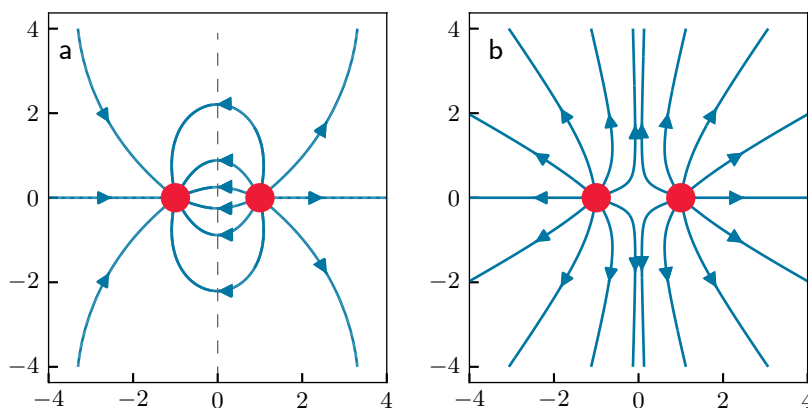


Figure 36.1: Field lines. (a) Electric field lines (streamlines of \vec{E}) set up by two opposite point charges. The magnetic field lines set up by two opposite pole tips look the same. The figure is antisymmetric upon reflection through the central plane (*dashed line*). (b) Two identical point charges or magnetic pole tips. This time the figure is symmetric upon reflection.

education and certainly no math, used his intuitive picture to make a discovery that had eluded everyone else: the law of induction. Maybe his viewpoint belongs in our *toolkit* alongside the others.¹

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 LINES ARE MATHEMATICALLY SIMILAR TO THE STREAMLINES OF AN INCOMPRESSIBLE FLUID

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* of a vector field define curves in space, and they do resemble the curves Faraday drew for the two situations in the figure. (Today they are often called **field lines**.) As in the figure, they spray out of a positive point charge, or the north pole tip of a magnet. They then spread apart, indeed as if by mutual repulsion.

To get more precise, let's warm up with a more intuitive system: an incompressible fluid flowing steadily through a pipe. There is a vector field (the local velocity near each point inside the pipe), whose streamlines are literally the paths taken by individual molecules (maybe averaged over thermal motion). Suppose that the flow encounters a constriction in the pipe. Then individual flowlines must converge. We know from watering our gardens that the fluid must also speed up as it passes through the constriction. Even though, if we sit at a fixed position, we see a time-independent (steady) fluid velocity there, still a speck of dust being swept along will be moving faster at the constriction.

Indeed, if $\vec{V}(\vec{r})$ is the velocity field and $\rho_m = \text{const}$ is the density of the incompressible fluid, then the flux of mass is $\rho_m \vec{V}$, and the continuity equation for mass says

$$\vec{\nabla} \cdot (\rho_m \vec{V}) = -\frac{\partial}{\partial t} \rho_m = 0.$$

That is, $\vec{\nabla} \cdot \vec{V} = 0$: Incompressible flow has divergence-free velocity. We know from Maxwell's equations that the magnetic field everywhere has this property, and the electric field has it in empty space.

Next, write \vec{V} in terms of its magnitude and direction: $\vec{V} = f(\vec{r})\hat{n}(\vec{r})$. The divergence-free property implies

$$\begin{aligned} \hat{n} \cdot \vec{\nabla} f &= -f \vec{\nabla} \cdot \hat{n} \\ \hat{n} \cdot \left(\frac{\vec{\nabla} f}{f} \right) &= -\vec{\nabla} \cdot \hat{n}. \end{aligned} \tag{36.1}$$

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

¹The graphical intuition will come back when we study radiation in Chapter 42.

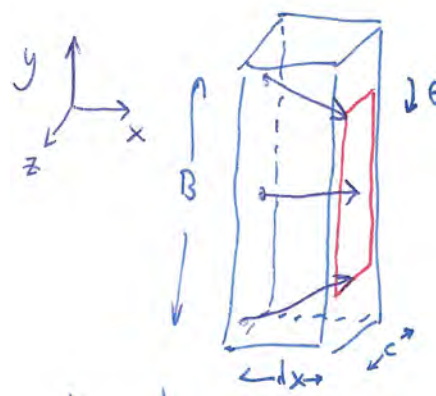


Figure 36.2: Fluid flow through a constricting channel.

should reflect changes in the *transverse density* of a set of neighboring streamlines, so let's see if the right hand side has any such interpretation.

Consider a simple situation, in which the constriction is just in one direction (y). Then \hat{n} lies always in the xy plane, as shown in Figure 36.2. In the middle of this small box $\partial\hat{n}/\partial x = 0$ but $\partial\hat{n}_y/\partial y < 0$. Thus, $\vec{\nabla} \cdot \hat{n} < 0$, as we expect for a converging flow. We ask what is happening to the transverse density of streamlines. If N lines enter at the left, spread over area BC , then they exit crammed into the smaller area $(B - 2dx \tan \theta)C$, where $\tan \theta \approx \theta \approx -\hat{n}_y$ evaluated at the top of the box. But $\hat{n}_y = 0$ at the center of the box, so by a Taylor expansion $\hat{n}_y(\text{top}) \approx \frac{1}{2}B \frac{\partial\hat{n}_y}{\partial y}$.

There's a similar shrinkage at the bottom, so the area of the rectangle containing the streamlines decreases from BC to $B(1 + dx \vec{\nabla} \cdot \hat{n})C$. Then the transverse density of streamlines increases from $N/(BC)$ to $N(1 + dx \vec{\nabla} \cdot \hat{n})^{-1}/(BC)$. Its relative change is then $-dx \vec{\nabla} \cdot \hat{n}$, which is 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 density of streamlines. (The constant is arbitrary because we could start with any number of streamlines.)

The same result holds for magnetic fields, and for electric fields in vacuum. (Electric charges act like sources or sinks.) Michael Faraday is smiling.

36.3 ELECTRIC AND MAGNETIC FORCES VIA DERIVATIVES OF FIELD ENERGY

The streamlines of \vec{E} thus 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} . Drawing in the lines for two opposite point charges, we see maximum density right at the charges, high density between them, and zero density out at infinity, as we should expect (Figure 36.1a). Moreover, bringing the two charges closer reduces the volume over which the lines are closely packed, and increases the volume in which

the lines are sparse. That reduces the integral of \vec{E}^2 , that is, the stored electrostatic field energy, so the opposite charges attract, as if the lines were real rubber bands under tension.

For two identical charges (Figure 36.1a), pushing them together increases the crowding at the central plane and increases energy, so the charges repel—as if the lines were real and created a transverse pressure.

36.4 FORCES VIA THE STRESS 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. So let’s now look directly at the electric and magnetic forces predicted by our formula for the momentum flux (or “stress”) 3-tensor, \vec{T}_{ij} .

Let

$$\vec{\vec{R}}_{ij} = \underline{F}^{0i} \underline{F}_0^j + \underline{T}^{ik} \underline{F}_k^j \quad \text{and} \quad S = \underline{F}^{0k} \underline{F}_k^0.$$

Recall that $\underline{F}^{0i} = \vec{E}_i/c$ and $\underline{F}^{ij} = \epsilon_{ijk} \vec{B}_k$. Thus,

$$\begin{aligned} \vec{\vec{R}}_{ij} &= \frac{1}{c^2} \vec{E}_i \vec{E}_j + (\delta_{im} \delta_{lj} - \delta_{ij} \delta_{lm}) \vec{B}_l \vec{B}_m \\ &= \frac{1}{c^2} \vec{E}_i \vec{E}_j - \delta_{ij} \vec{B}^2 + \vec{B}_i \vec{B}_j. \end{aligned}$$

Similar steps give $S = -(\vec{E}/c)^2$.

Your Turn 36A

Use these partial results to show that

$$\vec{\vec{T}}_{ij} = -\epsilon_0 \vec{E}_i \vec{E}_j + \frac{1}{2} \epsilon_0 \delta_{ij} \vec{E}^2 - \frac{1}{\mu_0} \vec{B}_i \vec{B}_j + \frac{1}{2\mu_0} \vec{B}^2 \delta_{ij}. \quad (36.2)$$

36.4.1 Electrostatic forces can be pictured as tension along, or pressure among, field lines

Electrostatic attraction and repulsion as momentum transfer.

Along the midplane in Figure 36.1a, the electric field points along \hat{y} , by symmetry. Everything to the left of the midplane transfers momentum to everything to the right with flux of \vec{p}_2 equal to

$$\vec{\vec{T}}_{22} = \epsilon_0 (-(\vec{E}_2)^2 + \frac{1}{2} \vec{E}^2). \quad (36.3)$$

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 it predicts a force on the right charge that is directed to the left, that is, attraction. Michael Faraday is smiling: This is his rubber-band tension at work.

Along the midplane in Figure 36.1b, the electric field is always perpendicular to \hat{y} . Thus,

$$\vec{\vec{T}}_{22} = \epsilon_0 (-(\vec{E}_2)^2 + \frac{1}{2} \vec{E}^2),$$

which is strictly positive. This time we predict repulsion. Michael Faraday is smiling: This is his transverse pressure at work.

36.4.2 Magnetostatic forces have similar pictorial expressions

The pictures look the same. And the magnetic terms of Equation 36.2 have the same forms as the electric terms. So we get the same results, and again Faraday is smiling.

Magnetostatic attraction and repulsion as momentum transfer.

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} (as in a transformer). Those statements eventually evolved into the magnetic part of the Lorentz force law and the field equation today called Faraday’s law, respectively.

PROBLEMS

36.1 *Push comes to shove*

a. Take the expression we found for the energy–momentum flux tensor:

$$\underline{T}_{\text{field}}^{\mu\nu} = -\mu_0^{-1} \left(\underline{E}^{\mu\lambda} \underline{E}_{\lambda}{}^{\nu} + \frac{1}{4} g^{\mu\nu} (\underline{E}^{\lambda\sigma} \underline{E}_{\lambda\sigma}) \right).$$

Consider a region where the magnetic field is zero. Write out the component \underline{T}_{33} in terms of the electric field.

b. Suppose that two identical point charges on the z axis are held close together. We know they will repel. Figure 36.1b shows some field lines near those two poles. Use your result in (a) to rederive the result about repulsion qualitatively. [*Hint*: Think about what crosses the xy plane.]

36.2 *Magnetic stress*

Consider the attraction between two bar magnets placed end-to-end with one’s N pole separated from the other’s S pole by a narrow gap. You can ignore fringe fields in this problem, and assume that \vec{B} is uniform in the gap and points in the \hat{x} direction.

- Show that, for a static, purely magnetic field, $\underline{T}^{\mu\nu}$ takes the form $u\underline{C}^{\mu\nu}$, where u is the energy density of the field and $\underline{C}^{\mu\nu}$ is a constant tensor that you are to find.
- Use the continuity equation for momentum to show that the force on each magnet (total rate of transport of momentum) equals $u\Sigma$, where Σ 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 35.13 for $\underline{T}_{\text{field}}^{\mu\nu}$ to evaluate this force.
- The total energy in the field is dominated by the contribution from the high-field space between the magnet poles, so it’s $ua\Sigma$, 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

Organize, systematize, consolidate, integrate. Let's see how some more of our earlier results reemerge in our new language. Chapter 20 studied energy and momentum *fluxes* in plane waves. Let's extend those results and show 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 density deposited by 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

$$\underline{F}^{\mu\nu} = \underline{\partial}^\mu \underline{A}^\nu - \underline{\partial}^\nu \underline{A}^\mu, \quad [34.13, \text{page 463}]$$

which cast Maxwell's equations as

$$-\underline{\partial}_\mu \underline{\partial}^\mu \underline{A}^\nu + \underline{\partial}_\mu \underline{\partial}^\nu \underline{A}^\mu = \mu_0 \underline{J}^\nu \quad [34.15, \text{page 464}]$$

with gauge invariance under

$$\underline{A}^\mu \rightarrow \tilde{\underline{A}}^\mu = \underline{A}^\mu + \underline{\partial}^\mu \Xi. \quad [34.14, \text{page 463}]$$

We could use this freedom to insist on Coulomb gauge as in Section 18.8.2 (page 268). But for some purposes, it's nicer to insist on a Lorentz-invariant condition,¹

$$\underline{\partial}_\mu \underline{A}^\mu = 0. \quad \text{Lorenz gauge} \quad (37.1)$$

Your Turn 37A

Show that in Lorenz gauge, Equation 34.15 becomes four *decoupled* copies of the d'Alembert equation: $\square \underline{A} = -\mu_0 \underline{J}$, or

$$c^{-2} \frac{\partial^2}{\partial t^2} \psi - \nabla^2 \psi = \rho_{\text{q}}/\epsilon_0 \quad \text{and} \quad c^{-2} \frac{\partial^2}{\partial t^2} \vec{A} - \nabla^2 \vec{A} = \mu_0 \vec{j}. \quad \text{Lorenz gauge} \quad (37.2)$$

Unlike our discussion in restricted Coulomb gauge,² Equations 37.2 are valid regardless

¹Named in honor of L. Lorenz. It's a Lorentz-invariant condition, but not named for H. Lorentz.

²See Section 25.1 (page 330).

of whether the charge density is zero or not. They are decoupled, but remember that the Lorenz gauge condition is a constraint linking the four variables ψ and \vec{A} .

37.2.2 It's permitted

Can we really insist on Lorenz gauge? Suppose that we had a vector potential not obeying Equation 37.1; that is, $\partial_\mu \underline{A}^\mu = f$ is some arbitrary function. Now apply a gauge transformation $\underline{A}^\mu \rightarrow \underline{A}^\mu + \partial^\mu \Xi$. Then $f \rightarrow f + \square \Xi$. But we have already found the solution to $\square \Xi = -f$ via its Green function in Chapter 25. So an appropriate Ξ 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

The scalar wave equation has plane-wave solutions of the form

$$\Phi(\underline{X}) = \frac{1}{2}(\exp(i\underline{k}_\mu \underline{X}^\mu) + \text{c.c.}),$$

characterized by a 4-vector $\underline{k}^\mu = \left[\frac{\omega/c}{\vec{k}} \right]^\mu$ (the 4-wavevector). These functions solve the scalar wave equation if $\|\underline{k}\|^2 = 0$ (“ \underline{k} is a **null 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 have plane wave solutions characterized by a null wavevector \underline{k} . Unlike the scalar field case, each wave also has a polarization 4-vector $\underline{\zeta}$:

$$\underline{A}^\mu(\underline{X}) = \frac{1}{2}\underline{\zeta}^\mu \exp(i\underline{k}_\nu \underline{X}^\nu) + \text{c.c.} \quad (37.3)$$

This 4-vector field will be in Lorenz gauge if $\underline{k}_\mu \underline{\zeta}^\mu = 0$.

We can apply a gauge transformation with Ξ that is itself of plane wave form.

Your Turn 37B

- Show that then $\tilde{\underline{A}}$ will still have the form Equation 37.3, but with $\underline{\zeta}$ replaced by its old value plus a multiple of \underline{k} .
- Show that $\tilde{\underline{A}}$ is still in Lorenz gauge, because $\underline{k}_\mu \underline{k}^\mu = 0$.

We can use the freedom you just found to impose the additional condition that $\underline{\zeta}^0 = 0$. With that choice,

$$[\underline{\zeta}] = \begin{bmatrix} 0 \\ \underline{P} \\ \underline{Q} \\ 0 \end{bmatrix}.$$

Your Turn 37C

- Work out the Faraday tensor and show that the electric field is parallel to $\vec{\zeta}$, and hence it is perpendicular to \vec{k} .
- Show that the magnetic field is perpendicular both to \vec{k} and to $\vec{\zeta}$.
- Also confirm that your formula for \underline{F} has the expected units.
- Suppose that we had not used our freedom to set $\zeta^0 = 0$. That is, suppose that $\underline{\zeta} = \begin{bmatrix} S \\ P \\ Q \\ S \end{bmatrix}$. What happens when you compute the Faraday tensor this time?

One way to express what you found in (d) is to note that the Faraday tensor involves the *projection* of $\vec{\zeta}$ onto the plane perpendicular to \vec{k} .

In short, we have recovered the key results that

- There are only two polarizations of light traveling in a given direction, and
- Both are transverse to the direction of propagation.

37.4 ENERGY AND MOMENTUM CARRIED BY A PLANE WAVE CONFIRM EARLIER EXPECTATIONS

Chapter 35 found the electromagnetic part of the energy–momentum tensor:

$$\underline{T}_{\text{field}}^{\mu\nu} = -\mu_0^{-1} \left(\underline{F}^{\mu\lambda} \underline{F}_{\lambda}{}^{\nu} + \frac{1}{4} g^{\mu\nu} (\underline{F}^{\lambda\sigma} \underline{F}_{\lambda\sigma}) \right). \quad [35.13, \text{page 484}]$$

Your Turn 37D

Show that for the Lorenz-gauge plane wave, the time-averaged energy–momentum flux tensor is

$$\langle \underline{T}_{\text{field}}^{\mu\nu} \rangle = \frac{1}{2\mu_0} k^\mu k^\nu \zeta_\mu^* \zeta^\nu = \frac{1}{2\mu_0} k^\mu k^\nu (|P|^2 + |Q|^2). \quad (37.4)$$

This compact formula 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.³

You should confirm that Equation 37.4 has units appropriate for energy density. Note that the two polarizations contribute *independently* to the energy and momentum (no cross-terms). This implies that they cannot interfere with each other; each polarization can only display interference phenomena with itself (Figure 37.1 and Media 12).

The two polarizations of light do not interfere with each other.

³Chapter 20.

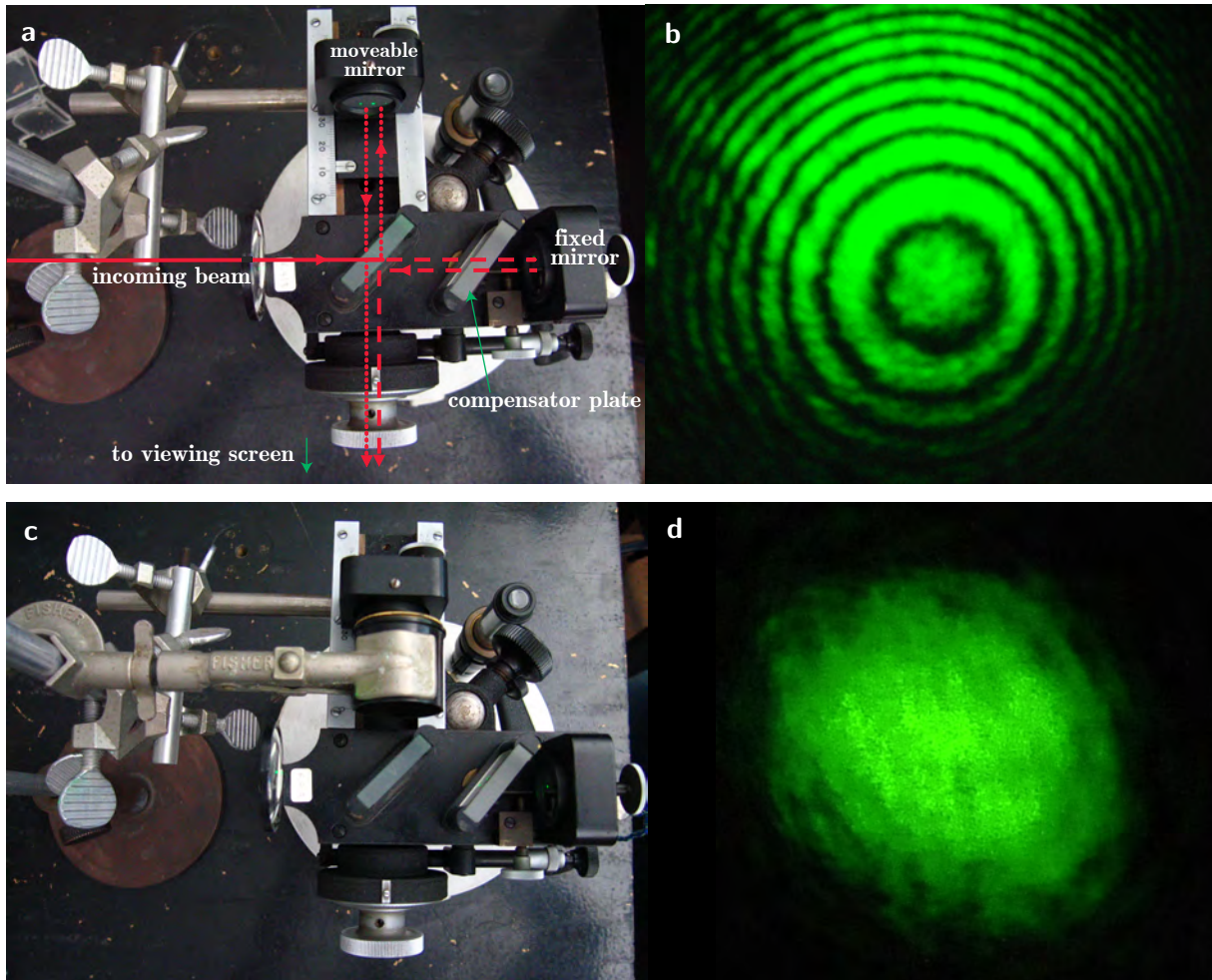


Figure 37.1: [Photographs.] **A Michelson interferometer setup.** (a) The incoming light is linearly polarized. (b) Interference fringes on the viewing screen produced by the setup in (a). (c) A quarter-wave plate has been added to the upper arm of the interferometer. Light passing through this arm gets its polarization rotated by up to 90 deg, depending on the orientation of the plate (see Problem 50.1, page 630). (d) Rotating the polarization in one beam by the full 90 deg, 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 37.1. (*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 12. [Realization and photos courtesy William Berner.]

Your Turn 37E

How would Equation 37.4 change if we had instead used a circular polarization basis?

The preceding expression is applicable for a pure plane wave. It implies that the energy density and the momentum flux 3-tensor are related in a simple way:

$$\langle \vec{T}_{\text{field}} \rangle = \langle \rho \varepsilon_{\text{field}} \rangle \hat{k} \otimes \hat{k},$$

regardless of the polarization of the wave. Now consider an isotropic mixture of many

different plane waves. More precisely,

- The directions \hat{k} are uniformly distributed.
- 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 $\langle\langle \dots \rangle\rangle$ denote averaging both over time and over the ensemble $\{\vec{k}_\ell, \vec{\zeta}_\ell\}$ of waves. Averaging the preceding equation gives

$$\langle\langle \vec{T}_{\text{field}} \rangle\rangle = \langle\langle \hat{k} \otimes \hat{k} \rho_{\mathcal{E}, \text{field}} \rangle\rangle.$$

The right side equals the angular average of $\hat{k} \otimes \hat{k}$ times $\langle\langle \rho_{\mathcal{E}, \text{field}} \rangle\rangle$.

The off-diagonal contributions to $\langle\langle \vec{T}_{\text{field}} \rangle\rangle$ average to zero. Energy density, however, is a nonnegative 3-scalar; its time average need not and will not be zero. Also, a symmetric rank-two 3-tensor such as $\hat{k} \otimes \hat{k}$ need not average to zero. For example, the identity tensor $\vec{\mathbb{1}}$ is unchanged by rotations (it’s a “tensor from Heaven”). Also, rotation does not affect the trace of a 3-tensor, so the rotational average of $\hat{k} \otimes \hat{k}$ must be⁴ $\frac{1}{3} \vec{\mathbb{1}}$.

The diagonal elements of the momentum flux 3-tensor give the *pressure*,⁵ so we get the simple conclusion that

$p = \frac{1}{3} \rho_{\mathcal{E}}.$ equation of state for isotropic EM radiation	(37.5)
--	--------

As mentioned in Section 20.2.3 (page 288), radiation pressure dominates over the gas pressure of ordinary matter in the early Universe, so Equation 37.5 is crucial for cosmology.

Radiation pressure dominates gas pressure in the early Universe.

⁴See Problem 14.2 (page 212).

⁵See Section 13.3.1 (page 190).



Gravitational radiation has just two polarizations.

37.2'a Gravitational waves

When we expand Einstein's gravitational 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 same old scalar wave equation, with the same old plane wave solutions! In particular, the polarization tensor is a symmetric, rank-2 tensor. Remarkably, a combination of a suitable Lorenz gauge and removal of residual gauge artifacts *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 spin-one states (for example, the p-orbitals of a hydrogen atom, or an s-wave, triplet bound state of two spin- $\frac{1}{2}$ particles) have *three* angular momentum states!"

This is interesting. You can always take a hydrogen atom, or a positronium "atom" in its triplet state, and view it in its rest frame. Then the usual analysis indeed guarantees three states. But a photon has *no rest frame*. There is thus no guarantee that the 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, 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.

PROBLEMS

37.1 Interference versus polarization

A monochromatic plane wave of light, with angular frequency ω , 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 then observe the flux of energy at various places on the screen (Figure 37.1, page 500 and Media 12).

Let Δ be the difference in path lengths traveled by the two beams.⁶ The original beam is linearly polarized in some direction $\vec{\zeta}$. Along the way, some optical element may rotate the polarization by an angle θ (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 θ and Δ . (You may neglect any overall constant.)
- c. **[T2]** 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?

37.2 Plane waves in Lorenz gauge

In Lorenz gauge, we studied 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. Extend your answer to Your Turn 37C by showing that the electric and magnetic fields are proportional to

$$\vec{\zeta} - \hat{z}(\hat{z} \cdot \vec{\zeta}) \quad \text{and} \quad \vec{\zeta} \times \hat{z},$$

respectively.

- b. What is the significance of these results for the paradox that the formula for \vec{A} appears to predict three independent polarizations of light?
- c. How might we have resolved that paradox without even bothering to compute \vec{E} and \vec{B} , and instead invoking gauge invariance?

37.3 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 $+\hat{z}$, in Lorenz gauge with angular frequency ω .
- b. Your answer involves a polarization 4-vector $\underline{\zeta}$. Write down an expression for the most general such $\underline{\zeta}^\mu$. 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.

⁶In 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 Δ .

- 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 ω , 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 $\underline{T}_{\text{field}}^{\mu\nu}$ in this situation. Make another Insightful Comment about the roles of the two polarizations in your answer.

[Hints: Use Equation 35.13 (page 484). 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} .]

37.4 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.⁷

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 $\pm \hat{z}$ 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 $-\hat{y}$. Now we wonder if there will be some apparent preference for the polarization along \hat{z} relative to \hat{x} , or vice versa.

- Write down the 4-vector potential corresponding to a plane wave of angular frequency ω , moving along $+\hat{y}$. Express the answer by using a wave 4-vector \underline{k} and a polarization 4-vector $\underline{\zeta}$. Use the usual complex exponential representation, and assume that $\underline{\zeta}$ is real (linear polarization). It will be convenient to work in Lorenz gauge, that is, to require $\partial_\mu \underline{A}^\mu = 0$. What conditions must \underline{k} and $\underline{\zeta}$ obey in order to have a solution to the vacuum Maxwell equations?
- Now apply a Lorentz boost to a coordinate system moving relative to the original one at speed (βc) 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?)
- Find the electric field of your wave solution in the original coordinate system. Show that it’s unchanged if you replace $\underline{\zeta}$ by $\underline{\zeta} + \xi \underline{k}$ for any constant ξ . Using this freedom, we can simplify the problem by also requiring that $\zeta^0 = 0$. Write the most general polarization 4-vector $\underline{\zeta}^\mu$ obeying all these requirements. Express it in terms of an amplitude b and the angle ψ that the electric field makes with the \hat{x} -axis.
- Take your boosted polarization vector from (b). Confirm that its electric field,

⁷Figure 30.5 (page 400).

- viewed in the new coordinate system, is still transverse. Use the trick in (c) to find an equivalent polarization vector $\tilde{\zeta}$ with the convenient property $\tilde{\zeta}'^0 = 0$. 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 (ω , b , and ψ), and β .
- 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 327) 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 ψ 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 nice, 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.

Electromagnetic phenomenon: Each far-field wavefront of a dipole spherical wave is isotropic, but the resulting energy flux is not.

Physical idea: The amplitude of the effectively plane wave in any direction depends on the projection of the underlying constant vector to the transverse plane—its polar graph looks like a *doughnut*.

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:

$$\vec{A}(t, \vec{r}) \stackrel{?}{=} \frac{1}{2} \vec{\xi} \frac{1}{kr} e^{-i\omega t \pm ikr} + \text{c.c.} \quad (38.1)$$

Here k is a scalar, r is distance from the origin, $\vec{\xi}$ is a constant vector, and as usual $\omega = ck$. The prefactor $1/k$ is a conventional choice designed to give $\vec{\xi}$ the same units as the polarization of a plane wave. The upper sign corresponds to outgoing spherical wavefronts; the lower sign to incoming.

To find the scalar potential, we make a similar trial solution for it:

$$\underline{A}^0(t, \vec{r}) \stackrel{?}{=} \frac{1}{2} \alpha(r) e^{-i\omega t \pm ikr} + \text{c.c.} \quad (38.2)$$

Here $\alpha(r)$ is an unknown function that we must find.

Your Turn 38A

- Now impose the Lorenz gauge condition to find what α must be. The insight is that α need not be a constant, nor even a constant divided by r , but you can nevertheless find it.
- Confirm that each of the three functions in Equation 38.1 indeed solves the wave equation.
- Also check that your answer to (a) has this property. Hence, conclude that Equation 38.1, along with your version of Equation 38.2, solves 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. Equation 38.1 has spherical wavefronts. Its amplitude $\|\vec{\xi}\|/(kr)$ 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.

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 wave¹ with wavevector $\vec{k}_{\text{pw}} = k\hat{n}$ and polarization vector $\vec{\zeta}_{\text{pw},i} = \vec{\xi}_i/(kL)$. We can therefore apply formulas from Chapter 37.

Your Turn 38B

- Work out the details, including ζ_{pw}^0 .
- Then find the electric and magnetic fields in terms of L , \hat{n} , and $\vec{\xi}$.
- 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*: Choose spherical polar coordinates with $\vec{\xi}$ pointing along the polar axis.]

Perhaps surprisingly, the fields (and therefore the energy flux) are *not at all isotropic*. It is true that the wavefronts (loci where $\vec{A} = 0$) are nice concentric spheres. But the amplitudes of the far \vec{E} and \vec{B} fields in various directions do depend on angle. They are all maximal in the directions perpendicular to $\vec{\xi}$, and *zero* when we view the wave from far away long the directions $\pm\hat{\xi}$.

There are many other spherical wave solutions, but the simplest one, which we just found, will turn out to be the most important contribution when we work out the radiation from an oscillating electric dipole in Chapter 43.² Accordingly, its pattern

Energy flux from a dipole radiator follows a “doughnut” pattern.

¹There are corrections that are higher order in powers of $1/L$.

²We'll encounter the other spherical waves when we study radiation systematically in Chapters 43–44.

of energy flux is sometimes called the **dipole doughnut** pattern.³

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 $1/L^2$. We can therefore say, a bit more carefully than in Section 25.5.3 (page 336), that the *total* energy output passing through a sphere of radius L approaches a *constant* as $L \rightarrow \infty$. 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 ω , sit at a *fixed* distance from the origin and consider the limit $\omega \rightarrow 0$, that is, keep only the leading behavior in powers of ω . You'll find that in this "near field" regime, \vec{E} dominates \vec{B} , and moreover \vec{E} has a very familiar form. The exact spherical wave solution considered in this section *interpolates* between this near-field form, which resembles the dipole field of electrostatics, 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 $\hat{x} + i\hat{y}$.

Your Turn 38C

Explore this case by using the same strategy as Section 38.2: What kind of plane wave does the solution look like when we stand far from the origin along some direction \hat{n} ? Is there any direction in which this wave is circularly polarized? Is there any direction in which it's *linearly* polarized? Can you explain your answers intuitively?

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.⁴

³A 3D polar plot depiction of $\sin^2 \theta$ resembles the surface of a toroidal pastry, at least when you are hungry. (Perhaps it looks more like a bialy, or a red blood cell.)

⁴Figure 37.1 (page 500) already showed this phenomenon. The striped interference pattern arose

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 differing by polarization.⁵

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 three-dimensional vector space of spherical waves (later we will find many more). Their wavefronts are spheres (hence the name), but they beam out energy in a “dipole doughnut” pattern that is maximal in the directions perpendicular to $\vec{\xi}$.

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 270) and 37.3 (page 498).

PROBLEMS

38.1 *Exact spherical wave solution*

Continuing Your Turn 38C, what direction is getting the largest energy flux, and why?

38.2 *Exact \vec{E} and \vec{B} fields*

- a. Work out the curl of Equation 38.1 (page 506) with the choice $\vec{\xi} = \xi \hat{z}$ and the outgoing (upper) sign in the exponential. It is simplest to find the derivatives using cartesian coordinates, then when you are done to express the answer in terms of polar unit vectors, finding

$$\vec{B} = \frac{\xi}{2k} \hat{\phi} \sin \theta (r^{-2} - ikr^{-1}) e^{-ikct+ikr} + \text{c.c.} \quad (38.3)$$

[Hint: First prove that $\hat{r} \times \hat{z} = -\hat{\phi} \sin \theta$.]

- b. Work out $-\vec{\nabla} \cdot \vec{A}^0 - \partial \vec{A} / \partial (ct)$ using your result from Your Turn 38A (page 507). Again the strategy suggested in (a) may be helpful; show that

$$\vec{E}/c = \frac{i\xi}{2k^2} \left[\hat{r} \cos \theta (2r^{-3} - 2ikr^{-2}) + \hat{\theta} \sin \theta (r^{-3} - ikr^{-2} - k^2 r^{-1}) \right] e^{-ikct+ikr} + \text{c.c.} \quad (38.4)$$

[Hint: First prove that $\hat{z} = \hat{r} \cos \theta - \hat{\theta} \sin \theta$.]

- c. Now take the low-frequency limit, or equivalently the short-distance limit, of your results in (a,b); that is, drop all but leading terms in powers of kr . Connect the result to an electrostatics problem we have discussed in a previous chapter.
- d. Next, take the opposite limit, that is, drop all but leading terms in powers of $(kr)^{-1}$. Connect your result to your answers to Your Turn 38B (page 507).

38.3 *Angular momentum of fields*

Background: EM waves can also carry *angular* momentum. You found the density of ordinary momentum in Your Turn 35C. So the density of *angular* momentum \vec{J}_3 , computed using the origin as reference point, is $(\mu_0 c^2)^{-1} [\vec{r} \times (\vec{E} \times \vec{B})]_3$. As usual, we will consider only the time average of \vec{J}_3 .

- 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 complex polarization⁶ $\vec{\xi} = C(\hat{x} + i\hat{y})$. 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 $1/r$.
- 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 $R \rightarrow \infty$. What saves us from that absurd conclusion?

⁶Chapter 43 will show that this solution could represent the radiation given off by a rotating electric dipole at the origin, in electric dipole approximation.

- d. Go back to (b) and keep also the next subleading 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 \vec{J}_3 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 \vec{J}_3 ? Why/why not?
- f. Now integrate the flux of \vec{J}_3 over the surface of the big sphere to get the rate at which angular momentum is transferred to the sphere.
- g. Also find the power absorbed by the sphere.
- h. Divide your answers to (f,g) and comment.

CHAPTER 39

Beams: Gaussian, Vortex, and Bessel

39.1 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:

$$\vec{A}(t, \vec{r}) = \frac{1}{2} \vec{\zeta} e^{-ikct + ikz} + \text{c.c.}, \quad [37.3, \text{page 498}]$$

where $\vec{\zeta}$ is a real vector. (In Lorenz gauge,¹ there is also a scalar potential ψ .) Another simple solution is a circularly polarized plane wave, obtained by replacing $\vec{\zeta}$ in the previous formula by one of the complex unit vectors that we called $\hat{\zeta}_{(\pm)}$ (Equation 18.32, page 272) times a suitable constant. Also there are 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 is finite in transverse extent, unlike a plane wave, and mainly traveling in one direction, unlike a spherical wave.

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, we might expect *diffractive spreading* to make the beam spray out on the other side in all directions, not just the one direction it was traveling in to begin with, like a plane water wave hitting a 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 pinhole, it continues for a considerable distance without spreading much. So in some sense the ideal of a real, physical beam seems realizable in practice. This chapter will begin by understanding and quantifying that statement. Then we will find some exotic beams that have recently proven very useful in applications.

Electromagnetic phenomenon: A structured beam of light can transfer angular momentum far greater than that of a circularly polarized plane wave.

Physical idea: Optical vortex beams exert gradient forces.

39.2 GAUSSIAN BEAMS

Let’s look for vacuum solutions to the Lorenz-gauge Maxwell equations (Equation 37.2, page 497) that have definite angular frequency $\omega = k_*c$. That is, we seek functions of the form $\underline{A}(\underline{X}) = \frac{1}{2} \underline{\bar{A}}(\vec{r}) e^{-i\omega t} + \text{c.c.}$, where

$$\boxed{(-k_*^2 - \nabla^2) \underline{\bar{A}} = 0 \quad \text{Helmholtz equation}} \quad (39.1)$$

¹Section 37.3 (page 498), Section 37.2.

and

$$-ik_*\vec{A}^0 + \vec{\nabla} \cdot \vec{A} = 0. \quad (39.2)$$

39.2.1 Paraxial approximation creates a mathematical analogy to the Schrödinger equation

The preceding equations were 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, $k_* = 2\pi/(600 \text{ nm})$, and the transverse extent is circular with radius $w = 1 \text{ mm}$, similar to the beam of a lab laser. 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.²

$$\vec{A}(t, \vec{r}) = \frac{1}{2}\vec{\zeta} e^{ik_*(-ct+z)}u(\vec{r}) + \text{c.c.} \quad \text{for constant } \vec{\zeta} \perp \hat{z}. \quad (39.3)$$

Substituting into the wave equation yields

$$\frac{1}{2}(\nabla^2 u + 2ik_*\vec{\nabla}_z u - \cancel{k_*^2 u} + \cancel{k_*^2 u})e^{-ik_*(ct-z)} + \text{c.c.} = 0.$$

We can now give a precise meaning to the phrase “traveling mostly along \hat{z} ”: If

$$(\vec{\nabla}_z)^2 u \ll \text{both } 2k_*\vec{\nabla}_z u \text{ and } (\vec{\nabla}_\perp)^2 u, \quad \text{paraxial conditions} \quad (39.4)$$

then we may drop a term in the wave equation, leaving

$$(\vec{\nabla}_\perp)^2 u + 2ik_*\vec{\nabla}_z u = 0. \quad \text{paraxial equation} \quad (39.5)$$

In words, the paraxial condition simply says that the unknown function $u(\vec{r})$ varies slowly along z compared with its transverse extent, and also compared to the wavelength. And the paraxial equation is an old friend—it's mathematically the same as the Schrödinger equation in two dimensions, with z playing the role of time.

It will shorten our formulas to nondimensionalize. Because the relevant transverse length scale is w , let $\bar{x} = x/w$ and $\bar{y} = y/w$. Because our problem is not isotropic, it will be convenient to rescale z differently,³ letting $\bar{z} = 2z/(k_*w^2)$. Then the paraxial equation and conditions take the simpler forms

$$(\vec{\nabla}_\perp)^2 u + 4i\frac{\partial u}{\partial \bar{z}} = 0 \quad \text{where} \quad (4wk_*)^{-2}\frac{\partial^2 u}{\partial \bar{z}^2} \ll \text{both } 4\partial u/\partial \bar{z} \text{ and } (\vec{\nabla}_\perp)^2 u. \quad (39.6)$$

²There will also be a scalar potential dictated by Lorenz gauge (Problem 39.3). Had we chosen to work in Coulomb gauge, then there would be no scalar potential (Section 18.8.3), but there would be a complicated extra condition on \vec{A} .

³The length scale $\frac{1}{2}k_*w^2$ is sometimes called the **Rayleigh range**.

You have some relevant experience with the Schrödinger equation from earlier courses. For example, you know that a confined wavepacket will spread over time, according to the Uncertainty Principle, but that a *gaussian* profile will spread minimally. That's because, although the packet falls off rapidly outside $\|\vec{r}_\perp\| = w$, nevertheless its Fourier transform is *also* gaussian, and hence also falls off rapidly. Thus, our trial solution will still be mostly propagating along \hat{z} . Let's make that expectation precise in our optical context.

Call the dimensionless transverse radial distance $\bar{\rho} = \|\vec{r}_\perp\|/w$. A Gaussian wavepacket at $\bar{z} = 0$ takes the form $\exp(-\bar{\rho}^2)$, but we must allow for it to spread and distort. So consider a trial solution of the form

$$u(\bar{z}, \vec{r}_\perp) = e^M, \text{ where } M = -\bar{\rho}^2 + \bar{z}S(\bar{\rho}) + \dots \quad (39.7)$$

In these formulas, S is a function we must find and the ellipsis represents higher-order terms in \bar{z} . We will be content with just the first-order correction, because in practice \bar{z} is numerically small:

Your Turn 39A

Estimate $k_* w^2$ using the typical numbers appearing earlier, and compare to the size of a typical lab setup. If we restrict to $\bar{z} \ll 1$, will that impose a severe limitation on the applicability of our calculation in the lab?

39.2.2 The gaussian beam spreads slowly, although its wavefronts curve

The preceding section argued that

We seek a solution to the Maxwell equations of the form Equations 39.3 with 39.7, satisfying the condition stated in Equation 39.6, in the region with $\bar{z} \ll 1$.

To find it, we now solve Equation 39.6.

Set up cylindrical coordinates $\bar{\rho}$, φ , \bar{z} . Recall⁴ that the two-dimensional Laplace operator is $(\vec{\nabla}_\perp)^2 = (\partial/\partial\bar{\rho} + \bar{\rho}^{-1})(\partial/\partial\bar{\rho}) + \bar{\rho}^{-2}\partial^2/\partial\varphi^2$. Hence in our case, the paraxial equation becomes

$$\begin{aligned} [(\partial/\partial\bar{\rho} + \bar{\rho}^{-1})(\partial/\partial\bar{\rho}) + \bar{\rho}^{-2}\partial^2/\partial\varphi^2 + 4i\partial/\partial\bar{z}]e^M &= 0 & (39.8) \\ 0 &= (\partial/\partial\bar{\rho} + \bar{\rho}^{-1})((-2\bar{\rho} + \bar{z}S')e^M) + 4iSe^M \\ 0 &= -2 + \bar{z}S'' + (-2\bar{\rho} + \bar{z}S')^2 - 2 + \bar{z}\bar{\rho}^{-1}S' + 4iS. \end{aligned}$$

We can now approximate by dropping all terms with the small quantity \bar{z} , to find⁵ $S \approx i(-1 + \bar{\rho}^2)$, or

$$u \approx \exp(-\bar{\rho}^2 + i\bar{z}(-1 + \bar{\rho}^2) + \dots). \quad (39.9)$$

Inspecting the modulus and phase of our solution, we see that

To first order in \bar{z} the beam width does not change, but its phase acquires a $\bar{\rho}$ dependence.

⁴Section 5.3.2 (page 66).

⁵Notice that $\bar{z}S'/\bar{\rho}$ is nonsingular, justifying this step.

Equations 39.3 and 39.9 give the small- \bar{z} limit of a solution to the paraxial equation called the **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 \bar{x} and \bar{y} at around 1, so the conditions in Equation 39.6 both amount to $(k_*w)^{-2} \ll 1$, which is certainly well satisfied in our example situation. Thus, Equation 39.9 indeed gives a beam that does not spread appreciably over typical lab dimensions. Had we kept higher powers in \bar{z} , we would have found eventual diffractive spreading away from the **beam waist** at $z = 0$ (and nonconstant $\vec{\zeta}$).

In lab conditions, gaussian beams propagate with very little spreading.

Your Turn 39B

Now put the pieces together (Equations 39.3 and 39.9). The phase of a wave is the imaginary part of the log of a complex function (the thing whose real part is a field component). A wavefront is a level set of the phase. Examine the phase of our solution at $t = 0$ and show that already at first order in \bar{z} , the wavefronts of \vec{A} begin to curve as \bar{z} increases from zero.

39.3 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 momentum is real (for example, it can be extracted by a charged particle), and that its quantity 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.

So it may be surprising to find that there are linearly polarized paraxial beams that can be engineered to carry arbitrarily high angular momentum per photon! Indeed, $30\hbar$ is readily achievable. Media 13 illustrates how such high transport can move large (micrometer) objects rapidly, overcoming the high viscous friction in the microworld.

You'll work out details of such **optical vortex** beams in Problems 39.2–39.3. Specifically, the solution you'll find is called a **Laguerre-gaussian** beam with “winding number” ± 1 ; much larger winding numbers can now be routinely created, for example via holographic construction.

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”).

Optical-vortex beams can exert forces transverse to their propagation direction.

Optical-vortex beams can transfer angular momentum far greater than that of a circularly polarized plane wave.

39.4 TRANSFER OF ANGULAR MOMENTUM TO A SPHERE

[Not ready yet.]

Bessel beams spread even less than gaussian beams.

39.5 BESSEL BEAMS

39.5.1 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,

$$\vec{A}(t, \vec{r}) = \frac{1}{2} e^{-ik_* ct + i\beta z} g(x, y) \vec{\zeta} + \text{c.c.} \quad (39.10)$$

For such a trial solution, the Helmholtz equation reduces to

$$k_*^2 g + \rho^{-1} \frac{\partial}{\partial \rho} \left(\rho \frac{\partial g}{\partial \rho} \right) + \frac{\partial^2}{\partial \varphi^2} g - \beta^2 g = 0. \quad (39.11)$$

As with our study of gaussian beams, we'll begin with the axially symmetric case, where $g = g(\rho)$ is independent of φ . Then we can solve

$$\rho^2 (k_*^2 - \beta^2) g + \rho \frac{dg}{d\rho} + \rho^2 \frac{d^2 g}{d\rho^2} = 0. \quad (39.12)$$

The solution of Equation 39.12 takes a standard form when we introduce the dimensionless variable $u = \rho \sqrt{k_*^2 - \beta^2}$; then $g = J_0(u)$, where the **Bessel function** J_0 is the solution to

$$u^2 \frac{d^2 J_0}{du^2} + u \frac{dJ_0}{du} + u^2 J_0 = 0.$$

39.5.2 A physically realizable approximation to the ideal

The **Bessel beam** solution that we just found has some remarkable features:

- It is an exact solution to the Maxwell equations (no paraxial approximation).
- It is exactly translation invariant, like a plane wave.
- But 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. Indeed, the fraction of energy flux within 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, with focal length f , placed at a distance f from the aperture. That's the complete setup, easily arranged in a lab.

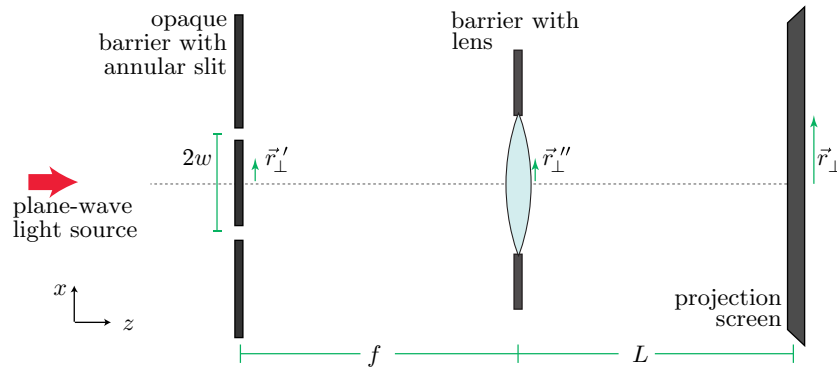


Figure 39.1: [Cartoon; not to scale.] **Durnin and coauthors' setup.** [Not ready yet.]

To find the downstream pattern of illumination, we first simplify the mathematical problem by going back to paraxial approximation (Section 39.2.1). 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:

$$u(\vec{r}_\perp, z) = -\frac{ik_*}{2\pi z} \int d^2r'_\perp u_0(\vec{r}'_\perp) e^{(ik_*/(2z))\|\vec{r}_\perp - \vec{r}'_\perp\|^2}. \quad (39.13)$$

Your Turn 39C

- Show that Equation 39.13 gives a solution to the paraxial equation.
- Show that as $z \rightarrow 0$, Equation 39.13 reproduces the starting profile u_0 .

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). Then we carry out the integral in Equation 39.13. 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 equivalently, an extra phase factor that depends on position \vec{r}_\perp'' in the plane $z = f$. The delay is maximal at the thickest part of the lens, that is, at $\vec{r}_\perp'' = \vec{0}$, and decreases as we move outward. In paraxial approximation we only need small values of $\|\vec{r}_\perp''\|$, so we can write a Taylor series:

$$\text{delay} = \text{const} - \frac{1}{2cf} \|\vec{r}_\perp''\|^2 + \dots \quad (39.14)$$

The constant in the second term must have dimensions $\mathbb{L}^{-2}\mathbb{T}$, so we may write it as c^{-1} divided by a length scale; calling that length scale $2f$ is convenient because then f is indeed the focal length of the lens.

We now use Equation 39.13 twice: 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:

$$u(\vec{r}'_{\perp}, f) = -\frac{ik_*}{2\pi f} e^{(ik_*/(2f))\|\vec{r}'_{\perp}\|^2} \int d^2r'_{\perp} u_0(\vec{r}'_{\perp}) e^{(ik_*/(2f))\|\vec{r}'_{\perp}\|^2} e^{-(ik_*/f)\vec{r}'_{\perp} \cdot \vec{r}'_{\perp}}.$$

The effect of the lens is therefore to cancel the first exponential. At $z = 2f$, we then have

$$u(\vec{r}_{\perp}, 2f) = \left(-\frac{ik_*}{2\pi f}\right)^2 e^{(ik_*/(2f))\|\vec{r}_{\perp}\|^2} \int d^2r''_{\perp} \left[e^{(-ik_*/(2f))\|\vec{r}'_{\perp}\|^2} e^{(ik_*/f)\vec{r}_{\perp} \cdot \vec{r}'_{\perp}} \right. \\ \left. \times \int d^2r'_{\perp} u_0(\vec{r}'_{\perp}) e^{(ik_*/(2f))\|\vec{r}'_{\perp}\|^2} e^{-(ik_*/f)\vec{r}'_{\perp} \cdot \vec{r}'_{\perp}} \right].$$

We now rearrange and do the \vec{r}'_{\perp} integral explicitly:

$$= (\text{const}) e^{(ik_*/(2f))\|\vec{r}_{\perp}\|^2} \int d^2r'_{\perp} \left[u_0(\vec{r}'_{\perp}) e^{(ik_*/(2f))\|\vec{r}'_{\perp}\|^2} \right. \\ \left. \times \int d^2r''_{\perp} e^{(-ik_*/f)(\vec{r}_{\perp} + \vec{r}'_{\perp}) \cdot \vec{r}'_{\perp}} e^{(ik_*/(2f))\|\vec{r}'_{\perp}\|^2} \right]. \quad (39.15)$$

If the integral over \vec{r}'_{\perp} were unrestricted, then combining the exponentials and completing the square shows that the inner integral would be a constant times $e^{(-ik_*/(2f))\|\vec{r}_{\perp} + \vec{r}'_{\perp}\|^2}$. A real lens is finite in extent, but we will neglect that limitation and substitute into Equation 39.15:

$$(\text{const}) \int d^2r'_{\perp} u_0(\vec{r}'_{\perp}) e^{(-ik_*/f)\vec{r}_{\perp} \cdot \vec{r}'_{\perp}}.$$

Note that the normalization constant may depend on z .

Within the approximations we made (paraxial; thin, wide lens), 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 $\|\vec{r}'_{\perp}\| = w$ and zero elsewhere, so

$$u(\vec{r}_{\perp}, 2f) = (\text{const})w \int d\varphi' e^{(-ik_*/f)w\rho \cos \varphi'} \quad \text{where } \rho = \|\vec{r}_{\perp}\|.$$

The remaining φ' integral is another representation of the Bessel function, so

$$u(\vec{r}_{\perp}, 2f) = (\text{const})J_0((k_*w/f)\rho). \quad (39.16)$$

The time-averaged intensity is then proportional to $|u|^2$.

Thus, the setup in Figure 39.1 should indeed create an approximately Bessel beam. Figure 39.2 shows a demonstration. Even at distances beyond $z = 2f$, we then expect a nearly invariant intensity profile. Figure 39.3 shows that in particular, 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 $10\ \mu\text{m}$ pinhole indeed yields a beam that spreads rapidly over just a few centimeters.

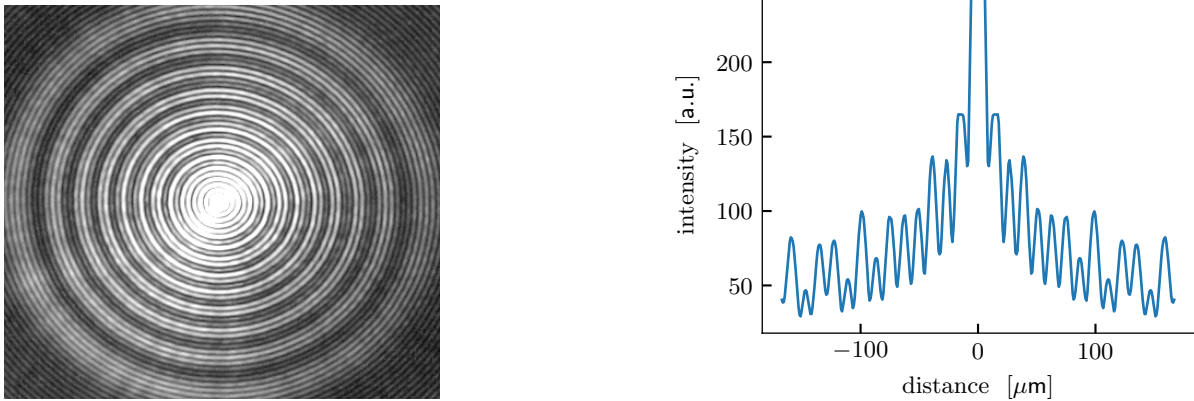


Figure 39.2: [Experimental data.] **Almost-Bessel beam**, produced via the setup described in the text. A laser beam with wavelength 532 nm impinged on an annular aperture: a ring with diameter $2w = 3.9$ mm and width $25 \mu\text{m}$. A lens with diameter 50 mm and focal length $f = 60$ mm was positioned 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 39.16). *Right:* intensity sampled along a radial line through the center. [Data courtesy Lucas Hanson.]

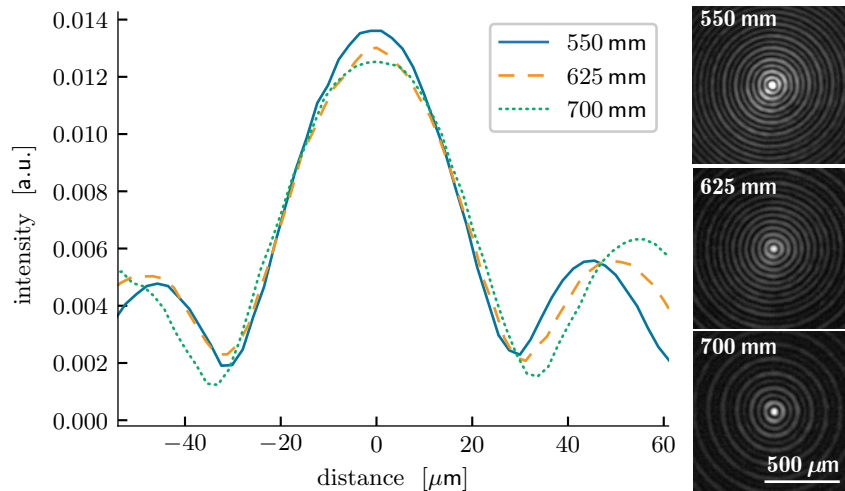


Figure 39.3: [Experimental data.] **Nearly diffractionless propagation of a beam.** The apparatus is the same as in Figure 39.2, but this time the light intensity was viewed at up to twelve focal lengths away from the lens. *Left:* Detail showing that the width of the central intensity maximum does not 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.]

39.5.3 Application to microscopy

[Not ready yet.] [“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 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 μm). 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 μm) as well as improved resolution over Bessel-beam and Gaussian-beam illumination have been demonstrated (70).” – Liu et al., 2020

63. Fahrbach FO, Simon P, Rohrbach A. 2010. Microscopy with self-reconstructing beams. *Nat. Photonics* 4(11):780–85
 64. Fahrbach FO, Gurchenkov V, Alessandri K, Nassoy P, Rohrbach A. 2013. Self-reconstructing sectioned Bessel beams offer submicron optical sectioning for large fields of view in light-sheet microscopy. *Opt. Express* 21(9):11425–40
 65. Chen Y, Liu JTC. 2015. Characterizing the beam steering and distortion of Gaussian and Bessel beams focused in tissues with microscopic heterogeneities. *Biomed. Opt. Express* 6(4):1318–30
 66. Gohn-Kreuz C, Rohrbach A. 2016. Light-sheet generation in inhomogeneous media using selfreconstructing beams and the STED-principle. *Opt. Express*. 24(6):5855–65
 67. Planchon TA, Gao L, Milkie DE, Davidson MW, Galbraith JA, et al. 2011. Rapid three-dimensional isotropic imaging of living cells using Bessel beam plane illumination. *Nat. Methods* 8:417–23
 68. Gao L, Shao L, Higgins CD, Poulton JS, Peifer M, et al. 2012. Noninvasive imaging beyond the diffraction limit of 3D dynamics in thickly fluorescent specimens. *Cell* 151(6):1370–85
 69. Yang Z, Prokopas M, Nylk J, Coll-Lladó C, Gunn-Moore FJ, et al. 2014. A compact Airy beam light sheet microscope with a tilted cylindrical lens. *Biomed. Opt. Express* 5(10):3434–42
 70. Vettenburg T, Dalgarno HIC, Nylk J, Coll-Lladó C, Ferrier DEK, et al. 2014. Light-sheet microscopy using an Airy beam. *Nat. Methods* 11:541–44

]]

FURTHER READING

Semipopular:

Simon, 2016.

Optical vortices as mechanical actuators: Media 13.

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:

Galvez, 2013; Götte & Barnett, 2013.

Micromanipulation with light: Grier, 2003. Holographic construction: physics.nyu.edu/grierlab/hot.html
 Ruffner & Grier, 2012; Curtis & Grier, 2003.

He et al. Direct observation of transfer of angular momentum to absorptive particles from a laser beam with a phase singularity. *Phys Rev Lett* (1995) vol. 75 (5) pp.

- 826-829;
- 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;
- Bessel beams: Experimental realization: Durnin et al., 1987. Other early work: Durnin et al., 1986; Durnin, 1987; Durnin et al., 1988. E. Betzig's application to light-sheet microscopy: Planchon et al., 2011; Martin et al., 2009; Gao et al., 2014.
- Chen, Y., Gao, J., Jiao, Z., Sun, K., Shen, W., Qiao, L., Tang, H., Lin, X. and Jin, X. (2018). Mapping Twisted Light into and out of a Photonic Chip. *Physical Review Letters*, 121(23). <https://doi.org/10.1103/PhysRevLett.121.233602>.
- Gibson, G., Courtial, J., Padgett, M., Vasnetsov, M., Pas'ko, V., Barnett, S. and Franke-Arnold, S. (2004). Free-space information transfer using light beams carrying orbital angular momentum. *Optics Express*, 12(22), p.5448. <https://doi.org/10.1364/OPEX.12.005448>.
- Ramachandran, S., Kristensen, P. (2013). Optical vortices in fiber. *Nanophotonics*, vol. 2, no. 5-6, pp. 455-474. <https://doi.org/10.1515/nanoph-2013-0047>.
- Wang, J., Yang, J., Fazal, I., Ahmed, N., Yan, Y., Huang, H., Ren, Y., Yue, Y., Dolinar, S., Tur, M. and Willner, A. (2012). Terabit free-space data transmission employing orbital angular momentum multiplexing. *Nature Photonics*, 6(7), pp.488-496. <https://doi.org/10.1038/NPHOTON.2012.138>.
- Ji et al., 2020.

PROBLEMS

39.1 *Display wavefronts*

Continuing Your Turn 39B (page 515), use a computer to display some wavefronts for $-0.2 \lesssim \bar{z} \lesssim 0.2$. *Hints:* By the axial symmetry of your solution to Your Turn 39B, it's enough to show just the intersections of the wavefronts with the $\bar{x}\bar{z}$ plane, that is, to show curves in that plane. Limit your graph to a relevant range of \bar{x} values, that is, a range where the amplitude is nonnegligible. To make the effect easier to see, set the constant $(k_*w)^2$ unrealistically small, say 30.

39.2 *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 39.2.

- a. Extend the discussion in Section 39.2 to find some more solutions of the form Equation 39.3. Again try $\vec{\zeta} = \zeta\hat{x}$, but now let your solution depend on azimuthal angle φ . The circular symmetry of the paraxial equation suggests that we look for solutions that are proportional to $e^{\pm i\varphi}$. But any such solution will have infinite derivatives on the symmetry axis ($\rho = 0$ in cylindrical coordinates), unless it also is proportional to ρ . So replace Equation 39.7 by the two trial solutions

$$u_m = e^{im\varphi} f_m(\bar{z}, \bar{\rho}), \quad \text{where } f_m = \bar{\rho} e^{-\bar{\rho}^2 + \bar{z}S_m(\bar{\rho}) + \dots}, \quad \text{and } m = \pm 1. \quad (39.17)$$

Here $\bar{\rho} = \rho/w$ is the dimensionless radial coordinate defined in the main text. Find the unknown functions $S_{\pm 1}(\bar{\rho})$ by using the same approach and approximations as in Section 39.2, and comment.

- b. Repeat Your Turn 39B and Problem 39.1 for your solution in (a), again with the value $(k_*w)^2 = 30$. 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. Are there also circularly-polarized beams of this type?
- d. *Optional:* Carry on with higher values of m .

39.3 *Twirl*

In this problem, you'll continue to study solutions to the Maxwell equations of the generic form Equation 39.17:

$$\vec{A}(t, \vec{r}) = \frac{1}{2}\vec{\zeta} e^{ik_*(-ct+z) + im\varphi} f_m(\bar{z}, \bar{\rho}) + \text{c.c.}, \quad (39.18)$$

where $\bar{\rho} = \sqrt{x^2 + y^2}/w$ and $\bar{z} = z/(2k_*w^2)$. Here $\vec{\zeta}$ is a constant vector. Section 39.2 found an approximate solution of this form with $m = 0$ and $f_0 = u = \exp(-\bar{\rho}^2 + i\bar{z}(-1 + \bar{\rho}^2))$. In the preceding problem, you studied the cases $m = \pm 1$. In this problem, you'll see a possibly surprising feature of the ± 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, continue to suppose that the solution is linearly polarized: $\vec{\zeta} = \zeta\hat{x}$. You can drop the overall constant ζ and reinstate it at the end; thus, \vec{A} will

temporarily be dimensionless. We would like to see whether the beams of light given by Equation 39.18 with $m \neq 0$ can make particles twirl around in the xy plane, despite not being circularly polarized. To do this, we need the flux of momentum (force per area) crossing the plane $\{z = 0\}$, as a function of x and y (or cylindrical coordinates ρ and φ). We are especially interested in the azimuthal component of this momentum flux, that is, in $\hat{\varphi} \cdot \vec{T}_{\text{field}} \cdot \hat{z}$.

- a. Starting from Equation 35.13 (page 484), derive a formula for this quantity in terms of components $\vec{E} \cdot \hat{z}$, $\vec{E} \cdot \hat{\varphi}$, $\vec{B} \cdot \hat{z}$, and $\vec{B} \cdot \hat{\varphi}$.

Before proceeding, note that all of our solutions have $\partial f / \partial z \propto 1 / (k_* w^2)$, but such quantities are much smaller than $\partial e^{ik_* z} / \partial z \propto k_*$. This means that when we calculate fields on the plane $z = 0$, the factors $\exp(\bar{z} S_m)$ will be negligible and may be dropped. For example, in this problem you may take $f_{\pm 1}$ to have the simple form $(\rho/w) \exp(-(\rho/w)^2)$, which is real and independent of z .

As so often in this book, it will be simplest to avoid the tricky formulas for divergence and curl in curvilinear coordinates, and instead to work in cartesian coordinates. It may be useful to recall that

$$\partial \rho / \partial x = x / \rho = \cos \varphi; \quad \partial \varphi / \partial x = -\rho^{-1} \sin \varphi; \quad \text{and so on.}$$

- b. Use the Lorenz gauge condition to find the scalar potential⁶ ψ corresponding to Equation 39.18 in terms of k_* , m , the function f_m , and its ρ derivatives.
- c. Find expressions for the four needed components of \vec{E} and \vec{B} fields and then set $z = 0$. Are the fields perpendicular to the z axis?
- d. Substitute your results in (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 $\rho \hat{\varphi} \cdot \vec{T}_{\text{field}} \cdot \hat{z}$. So evaluate this in terms of k_* , w , m , and the overall amplitude factor ζ from your result in (d). Comment on the distinction between the cases $m = -1, 0$, and 1 .

39.4 Twirl (circular polarization)

Work Problem 39.3, but with the following modifications: The beam is gaussian ($m = 0$), but circularly polarized.

⁶Recall our treatment of the spherical wave solution (Your Turn 38A).

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 translation of Lucretius (60 BCE)*

40.1 FRAMING: NOETHER THEOREM

Our derivation of $\underline{T}^{\mu\nu}$ 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 that the only candidate expression worked. But conservation laws should not be magical; they should be 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 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: [Not ready yet].

Physical idea: [Not ready yet].

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**³ \mathcal{L} is an ordinary function of two variables, with $x(t)$ substituted for the first argument and $\frac{dx}{dt}|_t$ for the second:

$$\mathcal{L}(x, \frac{dx}{dt}) = \text{KE} - \text{PE} = \frac{m}{2} \left(\frac{dx}{dt} \right)^2 - U(x).$$

Let's characterize those trajectories for which S is extremal over the space of all trajectories with fixed values at two time points: Substitute $\tilde{x} = x + \Delta x$, where $\Delta x(t_i) = \Delta x(t_f) = 0$. Expanding $S[x(t) + \Delta x(t)]$ to first order in $\Delta x(t)$ and using the Chain Rule gives

$$\Delta S = \int_{t_i}^{t_f} dt \Delta [\mathcal{L}(x, \frac{dx}{dt})] = \int_{t_i}^{t_f} dt \left(m \frac{dx}{dt} \frac{d\Delta x}{dt} - \Delta x \frac{dU}{dx} \right).$$

Now integrate the first term by parts. The boundary terms equal zero because we consider only variations that hold the endpoint values fixed:

$$\Delta S = - \int_{t_i}^{t_f} dt \left[-m \frac{d^2 x}{dt^2} - \frac{dU}{dx} \right] \Delta x.$$

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:

$$m \frac{d^2 x}{dt^2} = - \frac{dU}{dx}. \quad (40.2)$$

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

$$\mathcal{L}(\vec{r}_1, \vec{r}_2, \frac{d\vec{r}_1}{dt}, \frac{d\vec{r}_2}{dt}) = \frac{1}{2} \left(m \left\| \frac{d\vec{r}_1}{dt} \right\|^2 + m \left\| \frac{d\vec{r}_2}{dt} \right\|^2 + k \left\| \vec{r}_1 - \vec{r}_2 \right\|^2 \right). \quad (40.3)$$

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 α 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 variation with respect to the undifferentiated x_i , 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 is a scalar, and hence invariant under overall rotations; indeed, Your Turn 26A (page 347) involved a set of rotationally-invariant equations. We can also see at a glance that Equation 40.3 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

$$S[\text{traj}] = \int d^4X \mathcal{L}(\phi, \underline{\partial}\phi),$$

where for a scalar field ϕ , the lagrangian density \mathcal{L} is an ordinary function of five variables (the field and its four space and time derivatives at every 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.

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 Lorentz-invariant. 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

$$\underline{\partial}_\mu \left(\frac{\partial \mathcal{L}}{\partial (\underline{\partial}_\mu \phi)} \right) - \frac{\partial \mathcal{L}}{\partial \phi} = 0. \quad \text{Euler–Lagrange equation (fields)} \quad (40.5)$$

This time, the first term on the left denotes the result when we:

- Vary \mathcal{L} with respect to one of its four derivative variables, then
- Substitute values of ϕ and $\underline{\partial}_\mu \phi$, obtaining a function of \underline{X} , and
- Take a derivative with respect to \underline{X} .

⁴The Rules: Section 34.4. Or take the determinant of both sides of Equation 32.17 (page 430).

(The second term denotes the variation with respect to the undifferentiated ϕ , again followed by substituting values of ϕ and $\underline{\partial}_\mu\phi$.)

A specific choice for the lagrangian density of a scalar field could be a constant times

$$\mathcal{L}(\phi, \underline{\partial}\phi) = \frac{1}{2}(-\underline{\partial}^\mu\phi\underline{\partial}_\mu\phi - \lambda^{-2}\phi^2). \quad (40.6)$$

Evaluating Equation 40.5 then gives $-\underline{\partial}^\mu\underline{\partial}_\nu\phi - \lambda^{-2}\phi = 0$, in this context also called the **Yukawa equation** for its role in an early theory of nuclear forces.⁵

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 \underline{A} , so that half of the Maxwell equations become identities, not dynamical equations.⁶ The remaining equations are linear in \underline{A} , and second-order in its derivatives (Equation 34.15, page 464). So \mathcal{L} must be a *quadratic* function of \underline{A} , with at most two derivatives. It should be a Lorentz scalar, to ensure Lorentz-invariant field equations of motion, as well as being gauge- and translation invariant. There are very few such functions:

- $\underline{F}^{\mu\nu}\underline{F}^{\lambda\sigma}\underline{\varepsilon}_{\mu\nu\lambda\sigma}$: This term can be rewritten as $2\underline{\partial}^\mu(\underline{A}^\nu\underline{F}^{\lambda\sigma}\underline{\varepsilon}_{\mu\nu\lambda\sigma})$, that is, as 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 .
- The expressions $(\underline{\partial}_\mu\underline{A}^\mu)^2$ and $\underline{A}_\mu\underline{\square}\underline{A}^\mu$ are not gauge invariant. The expression $\underline{A}_\mu\underline{A}^\mu$ is not gauge invariant, and moreover contains no derivatives.
- $\underline{F}^{\mu\nu}\underline{F}_{\mu\nu}$ is the only remaining option, so we now explore it.

Following our recipe, we find the first order variation of our candidate action functional $S[\phi] = (\text{const}) \times \int d^4X \underline{F}^{\mu\nu}\underline{F}_{\mu\nu}$ and ask under what condition it will equal zero:

$$\Delta S = 0 = 2 \int d^4X \underline{F}_{\mu\nu}(\Delta\underline{F}^{\mu\nu}) = 2 \int d^4X \underline{F}_{\mu\nu}(\underline{\partial}^\mu\Delta\underline{A}^\nu - \underline{\partial}^\nu\Delta\underline{A}^\mu) \quad (40.7)$$

$$= 4 \int d^4X \underline{F}_{\mu\nu}(\underline{\partial}^\mu\Delta\underline{A}^\nu) \quad (40.8)$$

$$= -4 \int d^4X (\underline{\partial}^\mu\underline{F}_{\mu\nu})\Delta\underline{A}^\nu. \quad (40.9)$$

For this quantity to vanish regardless of $\Delta\underline{A}$, we must have that $\underline{\partial}^\mu\underline{F}_{\mu\nu} = 0$. That is indeed Maxwell's equations in vacuum.⁷

⁵When quantized, the field ϕ was once associated to particle states that could represent pions.

⁶Your Turn 34J (page 464).

⁷Set $\underline{J} = 0$ in the first of Equations 34.12 (page 462).

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 \underline{J} as in Section 34.9.2 (page 465). Then \underline{J} obeys the continuity equation for charge,

$$\partial_\mu \underline{J}^\mu = 0. \quad [34.10, \text{page } 462]$$

We may add $\underline{A}_\mu \underline{J}^\mu$ to our lagrangian density, because:

- This term is Lorentz invariant and its integral over d^4X is translation invariant.
- \underline{J} is gauge invariant, so under gauge transformation by Ξ we have (Equation 34.14, page 463)

$$\int d^4X \underline{J}_\mu \underline{A}^\mu \rightsquigarrow \int d^4X (\underline{J}_\mu \underline{A}^\mu + \underline{J}^\mu \partial_\mu \Xi).$$

Integrating by parts shows that the second term equals zero.

- This term is linear in \underline{A} , so it will contribute a term to the variational equations of order zero in \underline{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

$$\mathcal{L}(\underline{A}, \partial \underline{A}) = \mathcal{L}_f + \mathcal{L}_{\text{fp}} = \frac{1}{c} \left(-\frac{1}{4\mu_0} \underline{F}_{\mu\nu} \underline{F}^{\mu\nu} + \underline{A}_\mu \underline{J}^\mu \right). \quad (40.10)$$

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 (Js). The choice of sign 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 34.12, page 462 or Equation 34.15, page 464).

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 characterizing the particle's trajectory. It must be Lorentz invariant and local. The only obvious choice is the total elapsed proper time, but it has the wrong units. Fixing that defect with the available constants (particle mass and speed of light) yields a proposal for the particle part of the action:⁸

$$S_p[\text{trajectory}] = -mc \int d\xi \sqrt{-\|\underline{d}\underline{\Gamma}/d\xi\|^2}. \quad (40.11)$$

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

⁸The minus sign in the square root is needed because the length-squared of a timelike vector is negative.

on the trajectory to be explicit so that we know what we're doing when we vary it. Luckily, Equation 40.11 is invariant under change of parameter, so we may specify that ξ is any *fixed* parameter choice, for example, covering the fixed range from 0 to 1, and consider variations $\Delta \underline{\Gamma}$ 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 40.11 to the integral of Equation 40.10 and use Equation 34.19 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 442).

Solution: \mathcal{L}_f does not depend on the trajectory, so we wish to find the first-order variations of $\int d^4X \mathcal{L}_{fp}$ and of S_p . Start with the first of these:

$$\begin{aligned} S_{fp} + \Delta S_{fp} &= \frac{1}{c} \int d^4X \underline{A}_\mu(\underline{X}) q \int (cd\xi) \delta^{(4)}(\underline{X} - \underline{\Gamma}(\xi) - \Delta \underline{\Gamma}(\xi)) \left(\frac{d\underline{\Gamma}}{d\xi} + \frac{d\Delta \underline{\Gamma}}{d\xi} \right)^\mu \\ &= \int d\xi q \underline{A}_\mu(\underline{\Gamma}(\xi) + \Delta \underline{\Gamma}(\xi)) \left(\frac{d\underline{\Gamma}}{d\xi} + \frac{d\Delta \underline{\Gamma}}{d\xi} \right)^\mu. \end{aligned}$$

Taylor expand to find the displaced 4-vector potential, then integrate by parts:

$$\Delta S_{fp} = q \int d\xi \left(\frac{\partial \underline{A}_\mu}{\partial \underline{X}^\nu} \Big|_{\underline{\Gamma}(\xi)} \Delta \underline{\Gamma}^\nu \frac{d\underline{\Gamma}^\mu}{d\xi} - \Delta \underline{\Gamma}^\mu \frac{\partial \underline{A}_\mu}{\partial \underline{X}^\nu} \Big|_{\underline{\Gamma}(\xi)} \frac{d\underline{\Gamma}^\nu}{d\xi} \right).$$

We can pull out a common factor if we first rename the indices in the second term: $\mu \rightsquigarrow \nu$ and $\nu \rightsquigarrow \mu$:

$$\Delta S_{fp} = q \int d\xi \Delta \underline{\Gamma}^\nu \left(\frac{\partial \underline{A}_\mu}{\partial \underline{X}^\nu} \Delta \underline{\Gamma}^\nu \frac{d\underline{\Gamma}^\mu}{d\xi} - \frac{\partial \underline{A}_\nu}{\partial \underline{X}^\mu} \frac{d\underline{\Gamma}^\mu}{d\xi} \right) = q \int d\xi \Delta \underline{\Gamma}^\nu \underline{F}_{\nu\mu}(\underline{\Gamma}(\tau)) \frac{d\underline{\Gamma}^\mu}{d\xi}.$$

Next turn to the kinetic term and expand in $\Delta \underline{\Gamma}$:

$$\begin{aligned} S_p + \Delta S_p &= -mc \int d\xi \left(-\left\| \frac{d\underline{\Gamma}}{d\xi} + \frac{d\Delta \underline{\Gamma}}{d\xi} \right\|^2 \right)^{1/2} \\ &= -mc \int d\xi \frac{1}{2} \left(-\left\| \frac{d\underline{\Gamma}}{d\xi} \right\|^2 - 2 \frac{d\underline{\Gamma}^\mu}{d\xi} \frac{d\Delta \underline{\Gamma}_\mu}{d\xi} + \dots \right)^{1/2} \\ &= -mc \int d\xi \frac{1}{2} \left(-\left\| \frac{d\underline{\Gamma}}{d\xi} \right\|^2 \right)^{1/2} \left(1 - \frac{2}{-\|d\underline{\Gamma}/d\xi\|^2} \frac{d\underline{\Gamma}^\mu}{d\xi} \frac{d\Delta \underline{\Gamma}_\mu}{d\xi} + \dots \right)^{1/2} \\ \Delta S_p &= mc \int d\xi \left(-\left\| \frac{d\underline{\Gamma}}{d\xi} \right\|^2 \right)^{-1/2} \frac{d\underline{\Gamma}^\mu}{d\xi} \frac{d\Delta \underline{\Gamma}_\mu}{d\xi}. \end{aligned}$$

It is safe now to specialize to proper time parameterization, $\xi = \tau$, and use Equation 32.24 (page 433). Integrating by parts gives

$$= -mc \int d\xi \Delta \underline{\Gamma}_\nu \frac{d^2 \underline{\Gamma}^\nu}{d\tau^2} (c^2)^{-1/2}.$$

At last we can combine our two terms for the first-order variation and ask that they

equal zero for arbitrary $\Delta\underline{\Gamma}$. This happens only if the trajectory everywhere satisfies

$$0 = q\underline{F}_{\nu\mu} \frac{d\underline{\Gamma}^\mu}{d\tau} - m \frac{d^2\underline{\Gamma}_\nu}{d\tau^2}.$$

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 Maxwell's 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

To warm up, let's again begin with a simpler system, consisting of a single scalar field ϕ . 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

$$\phi(\underline{X}) \rightsquigarrow \tilde{\phi}(\underline{X}) = \phi(\underline{X}) + \epsilon D[\phi, \partial\phi](\underline{X}) + \dots \quad (40.12)$$

Here the ellipsis denotes terms of higher order in a bookkeeping parameter ϵ ; 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 \underline{X} . We suppose that the expression just given leaves S invariant for any trajectory $\phi(\underline{X})$, then ask for consequences in the situation where ϕ also obeys the variational equation associated to its action functional.

Here are two examples:

- A translation (shift of \underline{X} by a constant 4-vector $\epsilon\underline{b}$) corresponds to the local functional $D[\phi, \partial\phi] = \underline{b}^\mu \partial_\mu \phi$, as we see by Taylor expanding ϕ .
- Next, consider a set of *two* scalar fields, each with its own lagrangian density of the form Equation 40.6. Then

$$D \left[\begin{bmatrix} \phi_{(1)} \\ \phi_{(2)} \end{bmatrix} \right] = \begin{bmatrix} -\phi_{(2)} \\ \phi_{(1)} \end{bmatrix} = \mathbb{T} \begin{bmatrix} \phi_{(1)} \\ \phi_{(2)} \end{bmatrix}$$

implements an infinitesimal rotation in the *internal space* of ϕ 's components (not in physical space). Here \mathbb{T} is the generator of rotations in internal space (Section 3.7.2, page 44).

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 invariant under Equation 40.12). Thus, for each infinitesimal invariance of the system we must have

$$\mathcal{L} \rightsquigarrow \mathcal{L}(\tilde{\phi}, \partial\tilde{\phi}) = \mathcal{L}(\phi, \partial\phi) + \epsilon \partial_\mu \mathcal{M}^\mu(\phi, \partial\phi). \quad (40.13)$$

Here $\underline{\mathcal{M}}^\mu$ 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 40.6 gives

- For translation by \underline{b} ,

$$\underline{\mathcal{M}}^\mu = \frac{-1}{2}\underline{b}^\mu \underline{\partial}^\nu \phi \underline{\partial}_\nu \phi - \frac{1}{2}\lambda^{-2}\underline{b}^\mu \phi^2 = \frac{1}{2}\underline{b}^\mu (-\|\underline{\partial}\phi\|^2 - \lambda^{-2}\phi^2).$$

Your Turn 40C

Find $\underline{\mathcal{M}}$ for the case of internal rotations.

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 ϕ is a solution of the variational equation. To do this, first substitute Equation 40.12 into Equation 40.13:

$$\cancel{\mathcal{L}(\phi, \underline{\partial}\phi)} + \epsilon D[\dots] \left(\frac{\partial \mathcal{L}}{\partial \phi} \right) + \epsilon (\underline{\partial}_\mu D[\dots]) \frac{\partial \mathcal{L}}{\partial (\underline{\partial}_\mu \phi)} + \mathcal{O}(\epsilon^2) = \cancel{\mathcal{L}(\phi, \underline{\partial}\phi)} + \epsilon \underline{\partial}_\mu \underline{\mathcal{M}}^\mu + \mathcal{O}(\epsilon^2).$$

Next, rephrase the first term by using the Euler–Lagrange equation. Comparing the sides of this equation then shows that the 4-vector quantity⁹

$$\underline{\mathcal{J}}^\mu = \frac{\partial \mathcal{L}}{\partial (\underline{\partial}_\mu \phi)} D[\phi, \underline{\partial}\phi] - \underline{\mathcal{M}}^\mu \quad \text{obeys} \quad \underline{\partial}_\mu \underline{\mathcal{J}}^\mu = 0 \quad (40.14)$$

for any field trajectory that solves the equations of motion.

Equation 40.14 is the identity we were seeking, often called the **Noether theorem**. Returning to our two examples,

- For translation by \underline{b} ,

$$\underline{\mathcal{J}}^\lambda = -(\underline{\partial}^\lambda \phi \underline{b}^\mu \underline{\partial}_\mu \phi - \frac{1}{2}b^\lambda \underline{\partial}^\mu \phi \underline{\partial}_\mu \phi) + \frac{1}{2}\lambda^{-2}\underline{b}^\lambda \phi^2. \quad (40.15)$$

We can summarize all four of the associated continuity equations as

$$\underline{\partial}_\mu \underline{T}^{\mu\nu} = 0 \quad \text{where} \quad \underline{T}^{\mu\nu} = -\underline{\partial}^\mu \phi \underline{\partial}^\nu \phi - \frac{1}{2}g^{\mu\nu} (-\underline{\partial}_\sigma \phi \underline{\partial}^\sigma \phi - \lambda^{-2}\phi^2). \quad (40.16)$$

In fact, the symmetric tensor \underline{T} just defined is the energy–momentum flux 4-tensortensor!four@4D!energy–momentum flux of the scalar field theory under consideration. Tracing the derivation reveals that

Time-translation invariance implies energy conservation,
whereas the spatial translation invariances imply conserva-
tion of momentum.

- For the internal rotation invariance, $\underline{\mathcal{J}}^\mu = -(\underline{\partial}^\mu \phi)^t \mathbb{T} \phi$. Its corresponding conserved quantity is then $\int d^3r \underline{\mathcal{J}}^0$ (see Equation 8.6, page 114).

⁹We introduced the new generic symbol $\underline{\mathcal{J}}$ for the flux under construction, to distinguish it from $\underline{\mathcal{J}}$ which is always specifically electric charge flux.

40.4.3 Translational invariance of electrodynamics leads to the same \underline{T} as was found previously

Let's upgrade these ideas to electrodynamics, with the lagrangian density Equation 40.10. The derivation is a bit subtler than in the scalar field 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 $\epsilon \underline{b}$ is combined with a gauge transformation¹⁰ by $\Xi = -\epsilon \underline{b}_\mu \underline{A}^\mu$. (Other choices would yield an energy–momentum flux tensor that, although conserved, is not itself gauge invariant.)

The recipe given earlier starts by working out

$$\tilde{\underline{A}}^\mu = \underline{A}^\mu + \epsilon \underline{b}^\lambda (\partial_\lambda \underline{A}^\mu - \partial^\mu \underline{A}_\lambda), \quad \text{so} \quad D[\underline{A}, \partial \underline{A}]^\mu = \underline{b}_\lambda \underline{F}^{\lambda\mu}. \quad (40.17)$$

Equation 40.10 gives the change of field lagrangian density as

$$\mathcal{L}_f(\tilde{\underline{A}}, \partial \tilde{\underline{A}}) - \mathcal{L}_f(\underline{A}, \partial \underline{A}) = \frac{-1}{2\mu_0 c} \left[\partial_\mu (\epsilon \underline{b}_\lambda \underline{F}^{\lambda\mu}) - \partial_\nu (\epsilon \underline{b}_\lambda \underline{F}^{\lambda\nu}) \right] \underline{F}^{\mu\nu}.$$

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 $\partial_\mu \underline{F}_{\lambda\nu}$ by $-\partial_\lambda \underline{F}_{\nu\mu} - \partial_\nu \underline{F}_{\mu\lambda}$:

$$\begin{aligned} \Delta \mathcal{L}_f &= \frac{-\epsilon}{2\mu_0 c} \underline{b}^\lambda \left[-\partial_\lambda \underline{F}_{\nu\mu} - \cancel{\partial_\nu \underline{F}_{\mu\lambda}} + \cancel{\partial_\nu \underline{F}_{\lambda\mu}} \right] \underline{F}^{\mu\nu} \\ &= \frac{-\epsilon}{2\mu_0 c} \underline{b}^\lambda \left[-\frac{1}{2} \partial_\lambda (-\underline{F}_{\mu\nu} \underline{F}^{\mu\nu}) \right]. \end{aligned}$$

So indeed, the change is a total derivative (Equation 40.13), with

$$\underline{\mathcal{M}}_\lambda = \frac{-1}{4\mu_0 c} \underline{b}_\lambda \underline{F}_{\mu\nu} \underline{F}^{\mu\nu}, \quad (40.18)$$

Hence, for any \underline{b} we get a continuity equation for the quantities analogous to Equation 40.14:

$$\begin{aligned} \mathcal{J}^\lambda &= \frac{\partial \mathcal{L}}{\partial \partial_\lambda \underline{A}^\nu} D[\underline{A}, \partial \underline{A}]^\nu - \underline{\mathcal{M}}^\lambda \\ &= \frac{1}{\mu_0 c} \left[\underline{F}^{\lambda\nu} \underline{b}_\sigma \underline{F}^{\sigma\nu} + \frac{1}{4} \underline{b}^\lambda \underline{F}_{\mu\nu} \underline{F}^{\mu\nu} \right] = \underline{b}_\sigma \frac{1}{c\mu_0} \left[-\underline{F}^{\lambda\nu} \underline{F}^{\sigma\nu} + \frac{1}{4} \underline{g}^{\sigma\lambda} \underline{F}_{\mu\nu} \underline{F}^{\mu\nu} \right]. \quad (40.19) \end{aligned}$$

This expression is the contraction of $-\underline{b}/c$ with the electromagnetic energy–momentum flux tensor that we obtained in Chapter 35.¹² Its continuity equation, (40.14), is the result we already found in Section 35.5 (page 483), but now exposed as a consequence of translation symmetry.

¹⁰Equation 34.14 (page 463). The virtue of this approach will become apparent when we obtain gauge-invariant expressions in Equations 40.17–40.19.

¹¹Again see Your Turn 34J (page 464).

¹²Equation 35.13 (page 484).

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 40.16 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:

Coopersmith, 2017; Neuenschwander, 2017.

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; 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, and never 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.

T2 Supersymmetry: Weinberg, 2005a.

Technical:

Historical: Noether, 1918; English translation at arxiv.org/abs/physics/0503066.



40.4'a 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', page 487). 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 \underline{b} in the main text) are also anticommuting variables. We have seen that the parameters of a symmetry transformation may themselves transform, for example under rotations. The anticommuting parameters of supersymmetric transformations transform as spinors under rotations and other Lorentz transformations.

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. 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.

PROBLEMS

40.1 *Consequences of galilean invariance*

Illustrate the reasoning in Section 40.4 with the mechanical example of two masses joined by a spring (Equation 40.3, page 525). There are ten symmetries corresponding to the infinitesimal Galilean group transformations.

CHAPTER 41

Radiation Green Function Revisited

The past is not dead. It is not even past.

— *William Faulkner*

41.1 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 straightforwardly. 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 as what we found in Chapter 38! We need a more general understanding of spherical waves (see Chapter 44).

Electromagnetic phenomenon: 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 influence falls as r^{-3} .

41.2 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

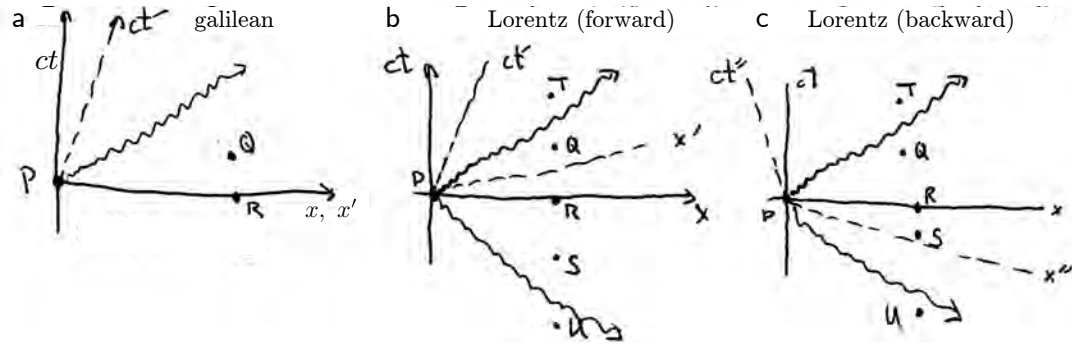


Figure 41.1: Relativity of simultaneity. In each panel, the wavy lines depict light trajectories. (a) See text. (b) The unprimed coordinate system says that events **P** and **R** are simultaneous, **Q** and **T** occur later than **P**, whereas **S** and **U** precede **P**. The primed system disagrees and says that **R** precedes **P**. (c) The doubly primed system disagrees about the time ordering of **R** and **S** relative to **P**.

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 41.1a): Any two *G*-inertial coordinate systems¹ 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 leftward-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 wavy lines at $\pm 45^\circ$ 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 $|t_Q - t_P| > \|\vec{r}_Q - \vec{r}_P\|/c$. We can restate this by using the invariant interval:³ The temporal ordering of two events **P** and **Q** is unambiguous if $(c\Delta\tau)^2$ is nonnegative, that is, if $\|\Delta\mathbf{X}_{PQ}\|^2 \leq 0$. Section 32.6.4 introduced the terms timelike separation if $\|\Delta\mathbf{X}_{PQ}\|^2 < 0$, 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 spacelike.

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

¹Recall that a *G*-inertial coordinate system is one in which the equations of motion take their usual (newtonian) form (Section 26.6.1, page 347).

²In some even faster-moving coordinate systems, **Q** precedes **P**!

³Equation 32.20 (page 432).

happened first? But it's not a complete disaster: When two events have timelike or lightlike separation, 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 would have to send a signal to the other moving faster than the speed of light in vacuum.⁴ 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.

41.3 GREEN FUNCTION FOR THE D'ALEMBERT EQUATION

41.3.1 Lorentz invariance tightly constrains the Green function

Section 37.2 obtained a version of Maxwell's equations valid in Lorenz gauge (Equation 37.2, page 497):

$$\square \underline{A}^\mu = -\mu_0 \underline{J}^\mu. \quad (41.1)$$

This is four decoupled copies of a single equation, so to simplify the notation let's first solve the scalar d'Alembert equation:

$$\square \phi = -\mathcal{J} \quad (41.2)$$

and later add the 4-vector index and factor of μ_0 . Chapter 25 found a solution to Equation 41.2, but we had to make an unobvious guess, and Equation 25.4 (page 331) didn't look exactly like a Green function solution. Let's use "Einstein thinking" to do better.

Equation 41.2 is linear and translation-invariant, so we expect that the solution can be written in terms of a Green function:

$$\phi(\underline{X}) = \int d^4 X_* D_r(\underline{X} - \underline{X}_*) \mathcal{J}(\underline{X}_*). \quad (41.3)$$

We now use invariance to constrain the possible form of the unknown function D_r , show that there is only one reasonable choice, then confirm that with that choice, the formula Equation 41.3 solves Equation 41.2 for any source function \mathcal{J} .

The constraints are that:

- D_r must be a Lorentz-invariant, scalar function of the 4-vector $\Delta \underline{X} = \underline{X} - \underline{X}_*$.
- It must have dimensions (length)⁻², by Equation 41.2. But it cannot involve any constants of Nature, because the equation doesn't contain any.

⁴What about quantum entanglement? Luckily that's not part of this course, 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 .

- It should vanish when $\Delta\underline{X}^0 < 0$, because the behavior of charges in the future cannot affect the values of fields in the past.⁵

The first constraint suggests that D_r must be a function of the invariant interval.

Your Turn 41A

One scalar quantity with the desired units (second constraint) is $\|\Delta\underline{X}\|^{-2}$. What's wrong with that choice?

Luckily, there is another option: We can satisfy all the constraints with a function of this form:

$$D_r(\Delta\underline{X}) = \frac{1}{2\pi} \delta(\|\Delta\underline{X}\|^2) \Theta(\Delta\underline{X}^0). \quad \text{radiation Green function} \quad (41.4)$$

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 that⁶ electromagnetic influences always travel at speed c . Two points can be joined by a path traversed at speed c only if they are lightlike-separated.
- The last factor is a “Heaviside step function,” and it enforces causality. Together with the delta function, it says that fields at \underline{X} can only be influenced by sources lying in the past light cone⁷ of \underline{X} .

The delta function has dimensions⁸ inverse to $(\text{length})^2$. The step function is dimensionless. So our proposal has the desired units.

The argument of the delta function is the invariant interval, so this whole factor is Lorentz invariant. The step function looks noninvariant at first, because Lorentz transformations can affect the temporal ordering of two events: $\Delta\underline{X}'^0$ may not have the same sign as $\Delta\underline{X}^0$. However, this problem can only arise for spacelike-separated events, that is, a pair of events with invariant interval 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 41.2 argued that the temporal ordering of any such pair of events is unambiguous.

In short, D_r is a 4-scalar function. The other ingredient in Equation 41.3 is d^4X , which we saw in Section 34.9.3 is also Lorentz-invariant. Thus, Equation 41.3 is overall a Lorentz-invariant recipe to obtain ϕ from \mathcal{J} , as desired.

Our trial solution has all the qualitative properties we expect it should have. Now we need to confirm that it really solves the d'Alembert equation. But once that's done, *everything about radiation will follow from Equation 41.4.*

⁵Strictly speaking, the *fields* must be causal; the *potentials* could be nonzero outside the light cone, as indeed they are in a noncovariant gauge choice like Coulomb gauge. We are only setting out heuristic expectations that will help us to formulate a promising guess.

⁶See Section 25.3 (page 331).

⁷Section 32.6.4 (page 432).

⁸See Section 0.3.8 (page 10).

41.3.2 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.

We want to substitute our guess Equation 41.4 into Equation 41.3 and ultimately confirm that Equation 41.2 is valid. After the substitution, we've got four integrals and one delta function. We will now use the delta function to eliminate one of the integrals, specifically the one over $\underline{X}_*^0 = ct_*$.

Recall from Section 34.9.1 how delta functions transform:

$$\delta(f(t_*)) = \sum_{\ell} |f'(t_{*,\ell})|^{-1} \delta(t_* - t_{*,\ell}), \quad [34.18, \text{page 465}]$$

where $t_{*,\ell}$ are all the values of t_* at which $f(t_{*,\ell}) = 0$. For our application,

$$f(t_*) = -c^2(t - t_*)^2 + R^2 \quad \text{where } R = \|\vec{r} - \vec{r}_*\|.$$

The quantities t and R are constants for purposes of evaluating the integral over t_* .

There are two solutions to $f = 0$: $t_{*-} = (ct - R)/c$ and $t_{*+} = (ct + R)/c$. Of these, however, the second is acausal and so cannot contribute (the step function eliminates it). Turning to the first,

$$\left. \frac{df}{dt_*} \right|_{t_{*-}} = 2c^2(t - t_{*-}) = 2c^2(t - t + R/c) = 2cR.$$

Thus,

$$\delta(\|\Delta \underline{X}\|^2) \Theta(\Delta \underline{X}^0) = \frac{1}{2cR} \delta(t_* - t + R/c).$$

That result lets us easily do the t_* integral in Equation 41.3. The three remaining integrals become

$$\phi(t, \vec{r}) = \frac{1}{4\pi} \int d^3r_* \frac{1}{R} \mathcal{J}(t - R/c, \vec{r}_*). \quad [25.4, \text{page 331}]$$

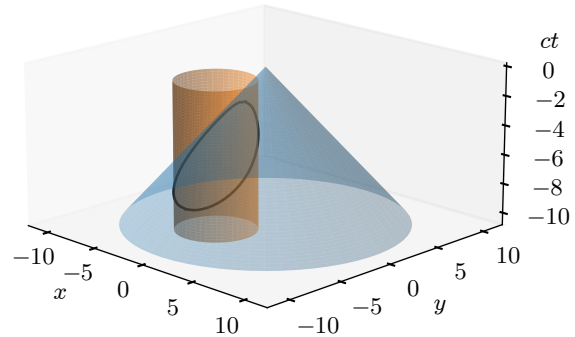
Chapter 25 already confirmed that ϕ defined by this formula solves the scalar d'Alembert equation. This time, however, we found it without having to make such a lucky guess, by using "Einstein thinking" (imposing manifest Lorentz invariance and causality).

41.4 REMARKS

41.4.1 Upgrade to 4-vector fields

Equation 25.4 is pretty simple: For each location \vec{r}_* inside the source, it tells us to look back in time to the moment when charges and currents at that location could have influenced the fields at (t, \vec{r}) , then introduce a factor of $1/(4\pi R)$. To upgrade

Figure 41.2: [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 $\vec{r} = \vec{0}$ 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} .



this result to electrodynamics,⁹ just use the scalar solution four times with $\mathcal{J} = \mu_0 \underline{J}^\mu$:

$$\underline{A}^\mu(\underline{X}) = \mu_0 \int d^3 r_* \frac{1}{4\pi \|\vec{r} - \vec{r}_*\|} \underline{J}^\mu(\underline{X}^0 - \|\vec{r} - \vec{r}_*\|, \vec{r}_*). \quad \text{Lorenz gauge} \quad (41.5)$$

This result looks a bit like the one we found in Coulomb gauge (Chapter 25). Unlike that result, however, this one assigns a nonzero value to the scalar potential. It is also valid even when the charge density is not everywhere zero.

Our recipe gets especially simple for a point charge sitting at rest at the origin, because $\vec{J} = 0$ and ρ_q is time-independent. So our solution reproduces the static Coulomb potential of a point charge.

More generally, we have shown that Equation 41.5 with Equation 41.4 gives the fields created (caused by) a general distribution of charges and currents. Other names for this **causal Green function** are **retarded Green function** or **retarded propagator**. The names refer to the fact that the formula “looks back in time.” For example, Figure 41.2 shows a current loop.

41.4.2 Check self-consistency

We’re not quite done. Equation 41.1 is *not equivalent to Maxwell* unless \underline{A} is in Lorenz gauge. Does our solution really have that property?

To find out, we must compute

$$\partial_\mu \underline{A}^\mu = \int d^4 X_* \underline{J}^\mu(\underline{X}_*) \frac{\partial}{\partial \underline{X}^\mu} D_r(\underline{X} - \underline{X}_*).$$

First, note that by the Chain Rule

$$\frac{\partial}{\partial \underline{X}^\mu} D_r(\underline{X} - \underline{X}_*) = -\frac{\partial}{\partial \underline{X}_*^\mu} D_r(\underline{X} - \underline{X}_*).$$

⁹L. Lorenz was the first to write the retarded potentials in Lorenz gauge, though not in the manifestly covariant form shown here.

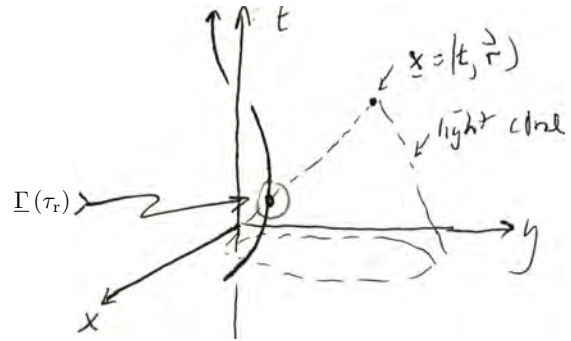


Figure 41.3: Retarded time. An observer at \underline{X} will measure electromagnetic fields set by a moving point charge at an earlier time t_r , determined by the intersection of the observer's past light cone with the particle's trajectory $\underline{\Gamma}$.

After that substitution, we can integrate by parts to find¹⁰

$$\partial_\mu \underline{A}^\mu = \int d^4 X_* D_r(\underline{X} - \underline{X}_*) \frac{\partial}{\partial X_*^\mu} J^\mu(\underline{X}_*).$$

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 41.1, for an isolated blip source.
- If we assemble a lot of blips together into a \underline{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.

41.5 POINT PARTICLE EXECUTING SPECIFIED MOTION

41.5.1 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 41.1 of a single electron undergoing specified motion. If we are given the particle trajectory parameterized by proper time, $\underline{\Gamma}(\tau)$, then Equation 34.19 gave the 4-current as

$$\underline{J}(\underline{X}) = \int d(c\tau) q \underline{U}(\tau) \delta^{(4)}(\underline{X} - \underline{\Gamma}(\tau)). \quad [34.19, \text{page 465}]$$

Substitute into Equation 41.3 with Equation 41.4:

$$\frac{2\pi}{\mu_0 q c} \underline{A}(\underline{X}) = \int d^4 X_* \Theta(\underline{X}^0 - \underline{X}_*^0) \delta(\|\underline{X} - \underline{X}_*\|^2) \int d\tau \underline{U}(\tau) \delta^{(4)}(\underline{X}_* - \underline{\Gamma}(\tau)).$$

¹⁰The reasoning here is similar to something we used in magnetostatics, Section 15.5.4 (page 221), and again later in our first look at radiation, Your Turn 25A (page 331).

Use the 4D delta function to eliminate four integrals:

$$= \int d\tau \underline{U}(\tau) \Theta(\underline{X}^0 - \underline{\Gamma}^0(\tau)) \delta(\|\underline{X} - \underline{\Gamma}(\tau)\|^2). \quad (41.6)$$

The step and delta functions tell us that the only contribution to the integral comes from the intersection of the particle's trajectory with the observer's past light cone (Figure 41.3). Let τ_r denote the corresponding proper time along the trajectory; similarly, the subscript “r” on other quantities will denote this event, for example, its time t_r in the lab coordinate system. As in Section 25.3 (page 331), we will call t_r the “retarded time.” In words:

The retarded time is the time when the past light cone of the observation event intersects the particle trajectory.

The definition implies that

$$ct - \|\vec{r} - \vec{\Gamma}(t_r)\| = ct_r \quad \text{and} \quad t_r < t. \quad (41.7)$$

Thus, $t_r(\underline{X})$ depends on the observation position and time. This remark will be important in a moment,¹¹ when we begin varying \underline{X} . Similarly, define $\vec{\beta} = d\vec{\Gamma}/d(ct)$ and $\vec{\beta}_r = \vec{\beta}(t_r)$.

Next, use the remaining delta function to remove the remaining integral, via Equation 34.18 (page 465)

$$(41.6) = \int d\tau \underline{U}(\tau) \left| \frac{d}{d\tau} \Big|_r (\|\underline{X} - \underline{\Gamma}(\tau)\|^2) \right|^{-1} \delta(\tau - \tau_r). \quad (41.8)$$

The derivative inside the absolute value is

$$-2(\underline{X} - \underline{\Gamma}(\tau_r))_\mu \underline{U}_r^\mu. \quad (41.9)$$

Let $\vec{R}_r = \vec{r} - \vec{\Gamma}(t_r)$. Then the derivative is

$$\begin{aligned} &= 2(c(t - t_r) \underline{U}_r^0 - \vec{R}_r \cdot c\vec{\beta}_r \frac{dt}{d\tau} \Big|_r) \\ &= 2c \frac{dt}{d\tau} \Big|_r (c(t - t_r) - \vec{R}_r \cdot \vec{\beta}_r) \end{aligned}$$

The quantity $c(t - t_r)$ equals R_r . Thus, we have

$$= c \frac{dt}{d\tau} \Big|_r R_r (1 - \hat{R}_r \cdot \vec{\beta}_r). \quad (41.10)$$

This quantity is always positive, because $\|\vec{\beta}\| < 1$, so we may drop the absolute value. Combining Equations 41.8–41.9 yields

$$\underline{A} = \frac{\mu_0 q c}{4\pi} \underline{U}_r (-(\underline{X} - \underline{\Gamma}(\tau_r))_\mu \underline{U}_r^\mu)^{-1}. \quad \text{Liénard–Weichert potentials} \quad (41.11)$$

¹¹The retarded time also depends implicitly on the particle's trajectory, but we will hold this fixed.

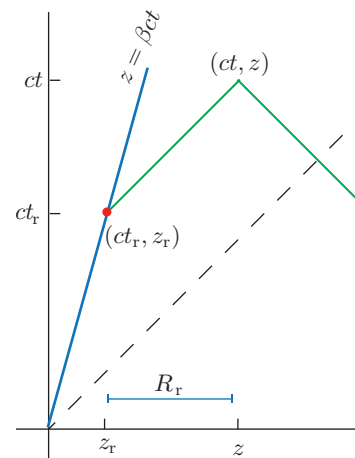


Figure 41.4: [Spacetime diagram.] **Graphical solution of Equation 41.13**, for the special case in which the field point \vec{r} lies on the z axis. The x and y directions have been suppressed for clarity. One of the two past-directed, light-speed lines from (ct, z) hits the particle trajectory exactly once, at (ct_r, z_r) ; the other such line misses the trajectory altogether.

Equation 41.11 is manifestly Lorentz-invariant. However, sometimes it is useful to render it in another form by using Equation 41.10:¹²

$$\underline{A}(\underline{X}) = \frac{\mu_0 q c}{4\pi} \begin{bmatrix} 1 \\ \vec{\beta}_r \end{bmatrix} |R_r - \vec{\beta}_r \cdot \vec{R}_r|^{-1}. \quad (41.12)$$

Again, to use this compact formula first solve Equation 41.7 for $t_r(\underline{X})$ and substitute into Equation 41.12. If we like, we can also replace $\mu_0 c$ by $1/(\epsilon_0 c)$.

41.5.2 Uniform motion once again

To gain confidence in the Liénard–Weichert formula, let’s revisit the problem of a constant-velocity trajectory, whose fields we have already found by other means.¹³ Suppose that a point charge q moves along the z axis at speed βc . Thus, its trajectory can be written as $\vec{\Gamma}(t) = \beta c t \hat{z}$ and Equation 41.7 says that t_r is the value of t_* for which

$$c(t - t_*) = R_{\text{traj}}. \quad (41.13)$$

For this simple situation, we can see explicitly that Equation 41.13 always has exactly one solution.

First proof (1D)

Figure 41.4 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 41.5a then illustrates that, for

¹²Equation 41.5 has reinstated the absolute value because Chapter 51 will revisit the situation in media, where the speed of a charged particle may exceed the local speed of light.

¹³See Section 33.4.2 (page 447) and 34.8.2 (page 464).

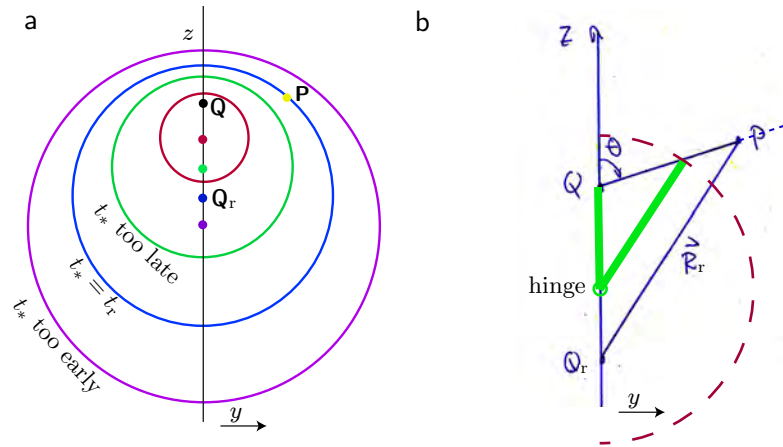


Figure 41.5: General solutions to Equation 41.13. In these figures, the time direction is suppressed but the y direction is shown. (a) Circles centered on $\beta c \hat{z} t_*$, of radii $c(t - t_*)$, for four choices of time t_* prior to the observation time t . The locations of the charge at those times are shown as *dots* on the z axis, and circles centered on those points are shown. The circles cover the entire plane, so one of them will certainly hit the observation point \mathbf{P} . Moreover, the circles never intersect, so *only one* of them hits \mathbf{P} . (If the x direction had been shown, the circles would instead be nested spheres.) (b) Two sticks (*green*) are joined by a hinge. See text for the argument that again establishes a unique solution.

any observation point \mathbf{P} in the yz plane, exactly one of the circles drawn intersects \mathbf{P} . Hence, there is exactly one contribution to Equation 41.12.

Third proof

Later it will be useful to have yet another graphical proof of the point just made. Again suppose that we have been given a choice of field point \mathbf{P} and observation time t . Figure 41.5b shows an example, along with the charged particle's position \mathbf{Q} at observation time t . This information determines the angle θ between the line $\overline{\mathbf{QP}}$ and the z axis.

What we need to find is another point, called \mathbf{Q}_r in Figure 41.5b, which is the charge's position at some earlier time t_r . Thus, the distance $\overline{\mathbf{QQ}_r}$ equals $\beta c(t - t_r)$. We want to know whether we may choose t_r such that also the distance $\overline{\mathbf{Q}_r\mathbf{P}}$ equals $c(t - t_r)$ (Equation 41.13), and if so, how many such choices exist.

Imagine two sticks joined by a hinge. The ratio of the sticks' lengths is β . Place the free end of the shorter stick at \mathbf{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 \mathbf{Q} through \mathbf{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 \mathbf{Q} . There will always be exactly one rescaling that makes the free endpoint pass through \mathbf{P} .

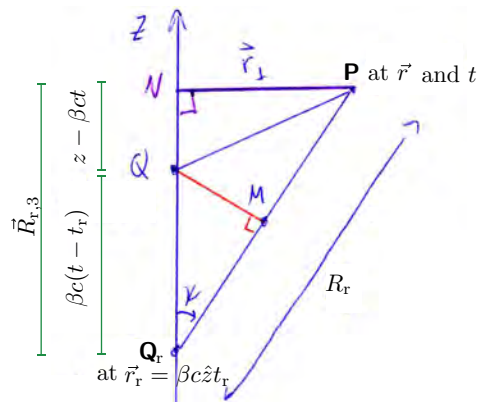


Figure 41.6: Geometry needed to evaluate Equation 41.12. We are given observation time t and position \vec{r} (the point \mathbf{P}). We also know where the charged particle is located at t (point \mathbf{Q}). We wish to find a prior point \mathbf{Q}_r on the trajectory that satisfies $R_r = c(t - t_r)$, which will also allow us to evaluate that quantity and the rest of Equation 41.12.

Your Turn 41B

Figures 41.4–41.5 were drawn assuming that the observer is ahead of the charged particle at the time of observation, that is, $z > \beta ct$. Redraw them to make sure they still work in the contrary case.

We again conclude that there is exactly one contribution to Equation 41.12 for any field point $\underline{X} = [\frac{ct}{\vec{r}}]$.

Evaluation of the potentials and fields

Now we must evaluate the expression $R_r - \beta \vec{R}_{r,3}$ appearing in Equation 41.12. Figure 41.6 shows a perpendicular dropped from \mathbf{Q} to the segment $\mathbf{Q}_r\mathbf{P}$ in red. Notice that there are two right triangles with a common angle ψ , so they are similar: $\triangle \mathbf{Q}_r\mathbf{NP} \sim \triangle \mathbf{Q}_r\mathbf{MQ}$, or

$$\frac{R_r}{\vec{R}_{r,3}} = \frac{\beta c(t - t_r)}{\overline{\mathbf{Q}_r\mathbf{M}}}.$$

Rearranging gives

$$\frac{R_r}{\beta c(t - t_r)} = \frac{\vec{R}_{r,3}}{\overline{\mathbf{Q}_r\mathbf{M}}}.$$

Also, Equation 41.13 gives $R_r = c(t - t_r)$, so we have

$$\beta \vec{R}_{r,3} = \overline{\mathbf{Q}_r\mathbf{M}}.$$

Hence, the quantity we need is

$$\begin{aligned} R_r - \beta \vec{R}_{r,3} &= R_r - \overline{\mathbf{Q}_r\mathbf{M}} = \overline{\mathbf{MP}} \\ &= \sqrt{\overline{\mathbf{QP}}^2 - \overline{\mathbf{MQ}}^2} = \sqrt{\vec{r}_\perp^2 + (z - \beta ct)^2 - (\beta R_r \sin \psi)^2} \\ &= \sqrt{(1 - \beta^2)\vec{r}_\perp^2 + (z - \beta ct)^2}. \end{aligned}$$

The square root is always real, because $\beta < 1$. Finally, substitute this result into Equation 41.12 and the similar formula for vector potential:

$$\begin{aligned}\psi(t, \vec{r}) &= \frac{q}{4\pi\epsilon_0} \left((1 - \beta^2) \vec{r}_\perp^2 + (z - \beta ct)^2 \right)^{-1/2} \\ \vec{A}(t, \vec{r}) &= \frac{q\mu_0}{4\pi} \beta c \hat{z} \left((1 - \beta^2) \vec{r}_\perp^2 + (z - \beta ct)^2 \right)^{-1/2}.\end{aligned}$$

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 using the following shortcut. Let $g(\vec{u}) = (\gamma^{-2} \vec{u}_\perp^2 + \vec{u}_3^2)^{-1/2}$. Thus,

$$\begin{aligned}\psi(t, \vec{r}) &= \frac{q}{4\pi\epsilon_0} g(\vec{r} - \beta ct \hat{z}) \\ \vec{A}(t, \vec{r}) &= \frac{q\mu_0}{4\pi} \beta c \hat{z} g(\vec{r} - \beta ct \hat{z}).\end{aligned}$$

So using cylindrical coordinates $u_\perp, \varphi, \vec{u}_3$,

$$\begin{aligned}\vec{B} = \vec{\nabla} \times \vec{A} &= \frac{q\mu_0\beta c}{4\pi} \left(\hat{u}_\perp \frac{1}{u_\perp} \frac{\partial g}{\partial \varphi} - \hat{\varphi} \frac{\partial g}{\partial u_\perp} \right) \\ &= \frac{q\mu_0\beta c}{4\pi} (-1) \left(-\frac{1}{2}\right) g^3 \gamma^{-2} 2u_\perp \hat{\varphi} = \frac{q\mu_0\beta c}{4\pi} \frac{\gamma u_\perp}{(u_\perp^2 + \gamma^2 \vec{u}_3^2)^{3/2}} \hat{\varphi}.\end{aligned}\quad (41.14)$$

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:

$$\begin{aligned}\vec{E} &= \frac{q}{4\pi\epsilon_0} \left(-\vec{\nabla}g - (\beta/c) \hat{z} (-\beta c) \frac{\partial g}{\partial \vec{u}_3} \right) \\ &= \frac{q}{4\pi\epsilon_0} \left(-\hat{u}_\perp \left(-\frac{1}{2}\right) g^3 \gamma^{-2} 2u_\perp - \hat{z} \left(-\frac{1}{2}\right) g^3 2\vec{u}_3 + \hat{z} \beta^2 \left(-\frac{1}{2}\right) g^3 2\vec{u}_3 \right) \\ &= \frac{q}{4\pi\epsilon_0} g^3 (\hat{u}_\perp \gamma^{-2} u_\perp + \hat{z} \gamma^{-2} \vec{u}_3).\end{aligned}$$

Electric field of a uniformly moving charge still falls as r^{-3} .

Next, note that $\hat{u}_\perp u_\perp + \hat{z} \vec{u}_3$ is just \vec{u} , which is $\vec{r} - \beta ct \hat{z}$. Thus,

$$\vec{E} = \frac{q\gamma}{4\pi\epsilon_0} \frac{\vec{r} - \beta ct \hat{z}}{(r_\perp^2 + \gamma^2(z - \beta ct)^2)^{3/2}}.\quad (41.15)$$

Equations 41.14–41.15 are the same results we obtained by applying a Lorentz transformation to the electrostatic field surrounding a static point charge.¹⁴

A charged particle in uniform motion carries fields along with it but does not radiate.

The solution that we have found corresponds to a spatial region with nonzero electric and magnetic field strengths, which moves at speed βc . The energy flux $\vec{E} \times \vec{B}$ 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 by a region with fields as it moves. 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.

¹⁴See Your Turns 33D and 33E, which however had the particle moving along the x axis and evaluated only in the plane $z = 0$.

41.5.3 Coda

The preceding section was a lot of work just to rediscover results we obtained earlier (Equations 41.14 and 41.15)! 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.¹⁵ Also, even the uniform velocity derivation will be useful in another context (Čerenkov radiation, Chapter 51), where the solution by Lorentz transformation will not be available.

¹⁵See Problem 42.3.

CHAPTER 42

Vista: J. J. Thomson's Pictorial Explanation of Radiation

Faraday read that electricities certainly existed, whereas there was much contention as to the forces exerted by them; but he saw that the effects of these forces were clearly displayed, whereas he could perceive nothing of the electricities themselves. And so he formed a quite different, opposite conception of the matter. To him, the electric and magnetic forces became the actually present, tangible realities; to him electricity and magnetism were the things whose existence might be disputed.

— *Heinrich Hertz, 1889*

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 such a field configuration falls as r^{-4} , too fast to transport any energy to infinity. 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 charges¹ by abstracting just one more qualitative fact from Chapter 41: 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 let us understand the magnetic field as well. Other 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 equatorial plane, and fall with distance as $1/r$.

Physical idea: Field lines must be continuous in empty space; to connect regions with simple, known behavior before and after the impulse, they must *kink*.

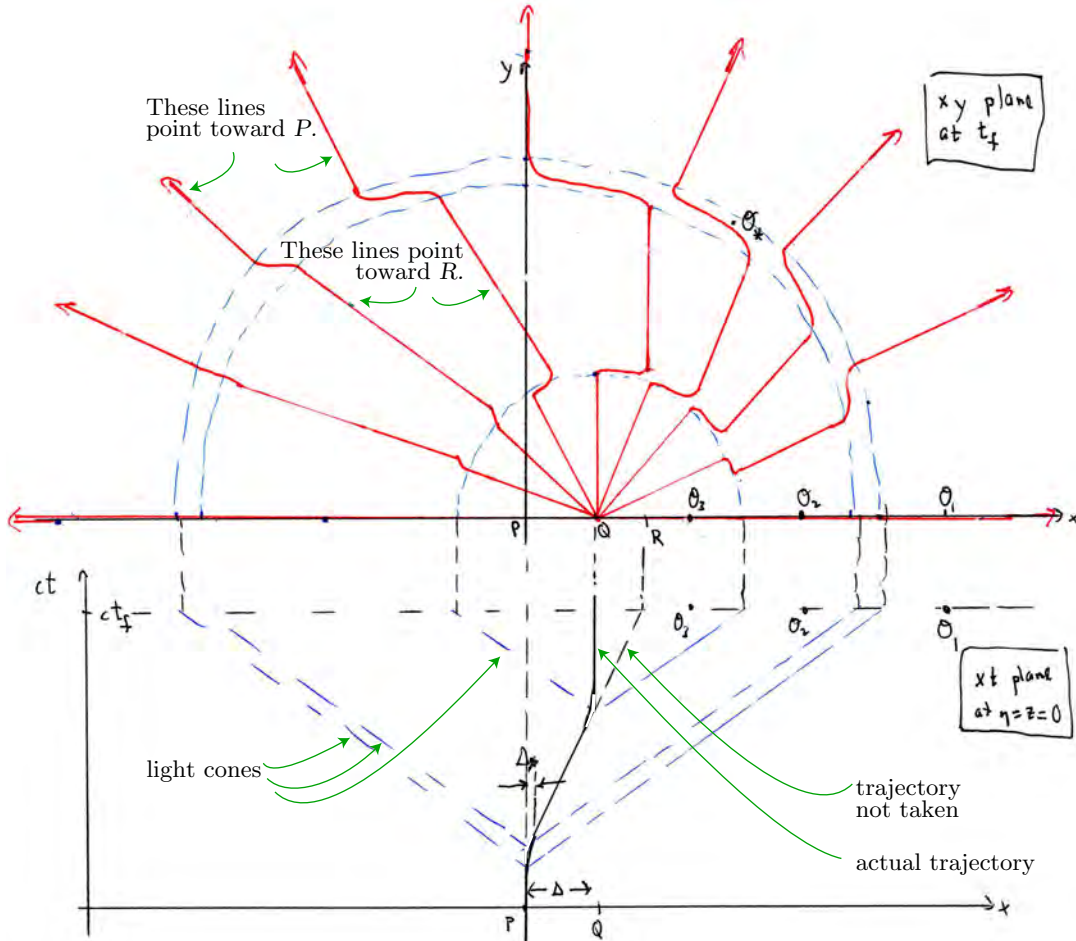


Figure 42.1: [Sketches.] **Electric field lines from a start/stop trajectory.** *Bottom:* A charged particle trajectory in the xt plane. *Top:* Snapshot of the corresponding field lines at one time, in the upper xy half-plane. The full 3D picture would be a figure of revolution about the \hat{x} axis. The bunching of field lines in the intermediate region, expected when the intermediate velocity is close to c , is not shown.

42.2 ELECTRIC FIELDS FROM A SUDDENLY ACCELERATED CHARGE

The lower panel of Figure 42.1 shows the trajectory in spacetime of a particle that is motionless from time $-\infty$ till time zero, then accelerates along \hat{x} , then decelerates to rest. At some time t_f after that last step, we ask what the fields look like throughout space.

- A very distant observer, at \mathcal{O}_1 , has not yet learned that the particle is no longer stationary at P , so it sees radial \vec{E} pointing outward from P toward \mathcal{O}_1 . A ring of such observers, all at the same distance, see uniformly spaced field lines with transverse density $1/r_{OP}^2$ (outermost arrows in the upper part of the figure).
- At the other extreme, a very nearby observer, at \mathcal{O}_3 , sees the up-to-date infor-

¹You began this program in Problem 34.3 (page 476).

mation, that is, radial \vec{E} pointing from q .

- In between, an observer, at \mathcal{O}_2 , sees radial \vec{E} pointing from R because that's where the charge *would have been* at time t_f , had it not decelerated, and this observer has not yet had a chance to learn that the charge has decelerated.²

We now connect up the three regions whose fields we just described. We know that $\vec{\nabla} \cdot \vec{E} = 0$, so *field lines cannot terminate* anywhere except on the charge itself. Thus, in the two joining regions the field lines must look as they are drawn in the figure:

- An observer at \mathcal{O}_* , for example, sees a pulse of \vec{E} directed *transversely* to her line of sight to P (and \vec{B} = out of page). These kinks lie on a spherical shell whose radius expands outward in time at speed c .
- There is an opposite kink associated to the deceleration, on another spherical shell that is also expanding outward.
- The kinks are most pronounced at 90 deg to the direction of acceleration (on the $\pm y$ axis); there is no kink along the direction of acceleration (on the $\pm x$ axis). Specifically, the kink is directed along $\hat{r} \times (\hat{r} \times \vec{a})$ where \vec{a} is the acceleration.

A suddenly accelerated charge emits bremsstrahlung.

Next let's ask about the strength of the transverse fields, for example in the first kink region (corresponding to the initial acceleration). We learned in Chapter 36 that $\|\vec{E}\|$ is proportional to the transverse density of the field lines, which in turn is the total length of all the lines in a volume, divided by that volume. And the *stretching* needed to accommodate the kink without breaking any line crowds more length into the thin shell than there would otherwise be (without any acceleration)!

Consider the situation at \mathcal{O}_* , a particular angle θ from the \hat{x} axis. The charge accelerates from velocity 0 to v , so the kink joins a line pointing toward P to one pointing toward R , a distance $\Delta = vt_f$ to the right in the \hat{x} direction (Figure 42.1), or $\Delta \sin \theta$ in the direction transverse to the field line. The acceleration occurs over a time interval v/a , so the thickness of the shell between the dashed lines is $\Delta_* = cv/a$.

Imagine drawing a total of N lines emerging from the charge. We wish to find the total length of all the field lines passing through a shell of thickness $c\Delta_*$ and cross-sectional area $d\Sigma$. A total of $Nd\Sigma/(4\pi r^2)$ lines enter, bend sideways, travel a distance $vt_f \sin \theta$, bend again, and emerge. Thus,

$$\frac{\text{total length of lines}}{\text{volume}} = \frac{Nd\Sigma vt_f \sin \theta / (4\pi r^2)}{d\Sigma \Delta_*}.$$

The radius of the sphere is ct_f , so we find that $\|\vec{E}\|$ is proportional to the acceleration, to $\sin \theta$, and to $1/r$. These are the key features of radiation from an accelerated charge:

- The electric field is transverse to the line of sight from observer to source.
- The electric field is mainly 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.

²This has nothing to do with the mental state of the observer. The causal structure of the theory is such that no instrument can, at the point in space and moment in time, distinguish the trajectory from one that is in eternal, uniform straight-line motion, and Section 33.4.2 showed that the field created in that situation is as described here.

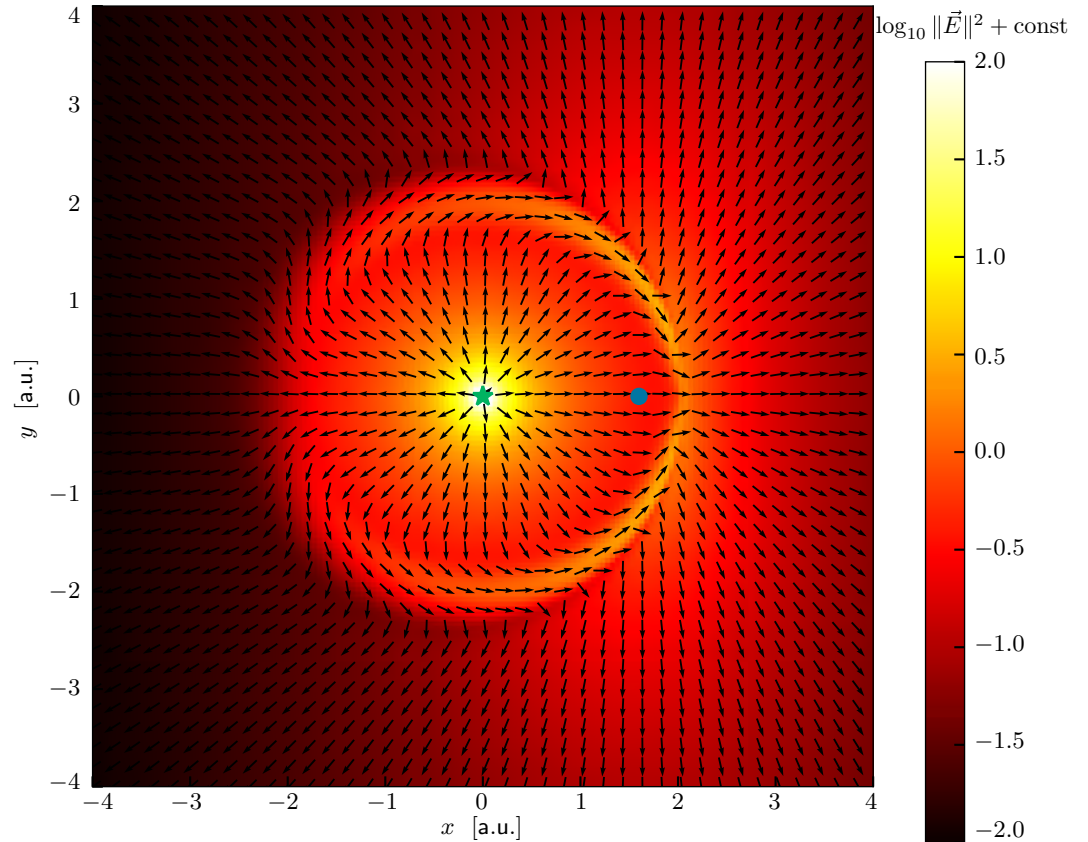


Figure 42.2: [Mathematical functions.] **Electric field** at time $ct_f = 2.0$ a.u., near a charged particle that was initially moving along $+\hat{x}$ at $0.8c$, but brought to rest at the origin (*star*) over a time interval $c\Delta t = 0.2$. Color indicates magnitude; arrows indicate direction. The *dot* indicates where the charge would have arrived by this time, had it continued in uniform motion. The figure was made by evaluating Equation 41.4 (page 540) on the trajectory, then numerically differentiating to find the electric field. [See also Media 14.]

42.3 MAGNETIC FIELDS ALSO HAVE A RADIATION CONTRIBUTION

The magnetic Gauss law does not give \vec{B} any sources or sinks. Hence, \vec{B} field lines must all be closed loops. Why, then, should they exist at all? The answer comes from the Faraday law.

We consider the same trajectory as before, but focus on only the final deceleration. The top panel of Figure 42.3 shows the electric field lines at a time t_1 . We argued a fixed point S in the xy plane, an observer will initially see a small, radial electric field ($\propto r^{-2}$), then around t_1 a pulse of \vec{E}_x ($\propto r^{-1}$), and then back to small field at later time t_2 .

An observer outside the expanding shell has not yet learned about the deceleration, so she sees the magnetic field of a particle in uniform, straight-line motion, which falls as r^{-2} . An observer inside the shell sees the magnetic field of a charge at rest, which is zero. But at the leading (outer) boundary of the shell (point T), the observer sees

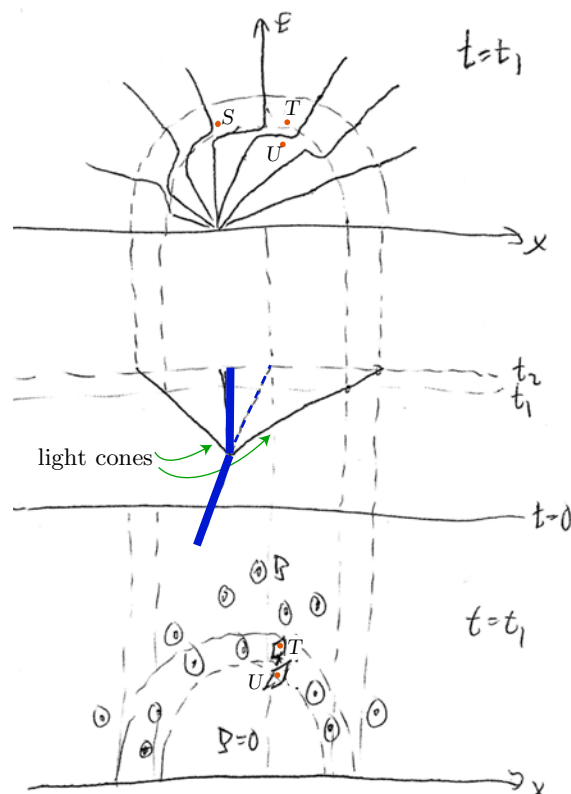


Figure 42.3: [Sketches.] **Electric and magnetic field lines.** *Middle:* The decelerating part of the trajectory in Figure 42.1 (*heavy blue line*). *Bottom:* Determination of magnetic fields. The small rectangular paths surrounding points T and U are parallel to the yz plane, that is, they extend out of the page.

a rising \vec{E}_x . The figure shows a small rectangular surface area surrounding T , coming out of the page in the yz plane. Two of the four edges of this rectangle straddle the boundary of the shell. Integrating Faraday's law over this surface element shows that either \vec{B}_y or \vec{B}_z must be nonzero there.

The field lines must form a figure of revolution about the x axis, by axial symmetry, and they must also be closed curves by the Gauss law. A radial (y) component of \vec{B} would require field lines to extend to infinity, and hence not close. Also, $\vec{B} = 0$ on the outermost edge of the rectangle. But an azimuthal (z) component is allowed on the inner edge of the rectangle. So at T we have \vec{B} pointing into the page, with field lines forming rings in planes parallel to the xz plane.

A similar argument applies at point U on the trailing boundary of the shell. Here \vec{E}_x is *falling* over time, but only on the *inner* edge of the rectangle. So again we find \vec{B} pointing into the page.

Throughout the shell we have $\|\vec{B}\| \propto r^{-1}$ because \vec{E} has that behavior. Thus, we find that

$$\vec{E} \times \vec{B} \text{ is directed radially outward and falls as } r^{-2}.$$

FURTHER READING

Intermediate:

Freeman et al., 2019, §3.3.

PROBLEMS

42.1 *Relativistic bremsstrahlung*

Consider a charge that is motionless for a long time, then gets rapidly accelerated to uniform straight-line motion at speed $V \ll c$, then gets rapidly decelerated back to rest. The lower panel of Figure 42.1 shows the trajectory of the charge in the xt plane. The upper panel depicts a snapshot of the electric field lines at a time after the particle has returned to rest, throughout the upper xy half-plane.

Now you sketch two similar pictures for the case where V is *not* much smaller than c . Discuss the differences from Figure 42.1 physically.

42.2 *Bremsstrahlung II*

[Not ready yet.]

42.3 *Bumper car*

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 from the unrealistic assumption of instantaneous deceleration. In this problem, you'll do better, in a different but related situation.

A point charge q sits motionless at the origin $\vec{r} = \vec{0}$ 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

$$\vec{\Gamma}(t) = \begin{bmatrix} x_0 e^{-t^2/(2\tau^2)} \\ 0 \\ 0 \end{bmatrix}.$$

Here x_0 and τ are constants. Measure all lengths and ct values in some 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 $0.4c$ (relativistic motion).

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.

- What is x_0 ?
- 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 solution to an ordinary (not differential) equation.
- Before actually calculating anything, show that 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. The grid should be fine enough to get reasonably accurate estimates of derivatives by numerical differentiation. Then use a computer to evaluate your result from (b) numerically at each grid point, at each of the time values that you chose.
- e. Do whatever you need to do to convert your result from (d) into an evaluation of \vec{E} on the xy plane at each time value. (Why is it good enough just to show the xy cross-section?)
- f. 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 heatmap; that is, show this 3-scalar quantity as color on a plane. Or you may prefer a contour plot or surface plot. Use your judgement about what is clearest.
- g. Point out all the visual features of the result that you can explain, and explain why they arose. This crucial step also serves as reasonableness checking. 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, and do your graphics show that behavior?
- h. A picture may be worth a thousand words, but a *movie* is worth many pictures, so get your computer to make an animated graphic from individual video frames.
- i. *Optional:* Use your superpowers to create some other meaningful representation, using your own judgement, 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 a small array.)
- 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.)

CHAPTER 43

Electric Dipole Radiation

43.1 FRAMING: *DOUBLE EXPANSION*

This chapter will show in a special situation that, as foreshadowed in Chapter 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 realized 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: Homonuclear molecules have little infrared activity, but more complex ones can be strong greenhouse gases.

Physical idea: O_2 and 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 a centered on the origin of coordinates. So they satisfy $\|\vec{r}_*\| < a$. We observe fields at \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, 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 given frequency ω . Then the quantity c/ω has dimensions of length, and indeed, it will be the wavelength of emitted radiation (divided by 2π). One common situation is then

$$r \gg c/\omega \gg a. \quad \text{far field (dynamic)} \quad (43.1)$$

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.

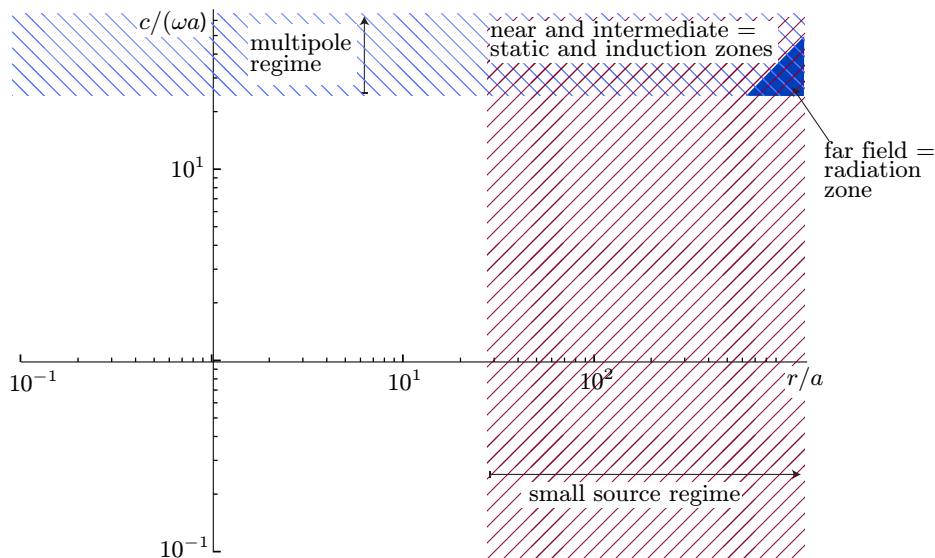


Figure 43.1: Three limiting cases. The small source regime is $r \gg a$ (hatched region extending to right). The multipole limit is $c/\omega \gg a$ (hatched region extending upward). The far field regime is the intersection of these, with the additional condition $r \gg c/\omega$ (solid triangle extending up and to the right).

Accordingly, the following sections will introduce

$$r \gg a \quad \text{small source regime} \quad (43.2)$$

and

$$c/\omega \gg a. \quad \text{multipole regime} \quad (43.3)$$

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 $\vec{R} = \vec{r} - \vec{r}_*$. Please review why¹

$$R = r - \hat{r} \cdot \vec{r}_* + \dots \quad (43.4)$$

$$R^{-1} = r^{-1} \left(1 + \frac{\hat{r} \cdot \vec{r}_*}{r} + \dots \right). \quad (43.5)$$

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 smaller in the small source regime (Equation 43.2).

¹Section 3.4 (page 39).

Our general, Green-function solution gives the vector potential in Lorenz gauge as²

$$\underline{A}^\mu(t, \vec{r}) = \frac{\mu_0}{4\pi} \int d^3r_* \frac{1}{R} \underline{J}^\mu(t - R/c, \vec{r}_*). \quad [41.5, \text{page } 542]$$

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 \underline{A} that are suppressed by powers of a/r relative to the first term. But in the argument of \underline{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 $r \rightarrow \infty$, and
- When we take \underline{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 will 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,

$$\underline{A}^\mu(t, \vec{r}) = \frac{\mu_0}{4\pi r} \int d^3r_* \underline{J}^\mu(t - r/c + \hat{r} \cdot \vec{r}_*/c, \vec{r}_*). \quad \text{small source} \quad (43.6)$$

Equation 43.6 is the generalization of Equation 25.9 (page 334) 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 \underline{J} vary harmonically in time with some angular frequency ω . That is, assume³

$$\underline{J}^\mu(t, \vec{r}_*) = \frac{1}{2} e^{-i\omega t} \bar{\underline{J}}^\mu(\vec{r}_*) + \text{c.c.},$$

where $\bar{\underline{J}}^\mu$ are four complex functions of position \vec{r}_* only. Then

$$\underline{A}^\mu(t, \vec{r}) = \frac{1}{2} \frac{\mu_0}{4\pi r} e^{-i\omega(t-r/c)} \int d^3r_* e^{-i\omega \hat{r} \cdot \vec{r}_*/c} \bar{\underline{J}}^\mu(\vec{r}_*) + \text{c.c.} \quad (43.7)$$

Everything inside the integral is independent of the observer's distance r . However, the observer's *direction* \hat{r} is still present inside the integral.

²We used relativity to obtain this formula. However, this chapter will consider a nonrelativistic problem (charges moving much slower than light), so there will be little benefit to writing only manifestly Lorentz-invariant formulas.

³If \underline{J} is not harmonic, we may nevertheless be able to decompose it into Fourier components, use the analysis below on each one, and ultimately add all their contributions. But see the caveat in Section 41.1 (page 537).

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/ω is much bigger than the source size⁴ a (the multipole regime, Equation 43.3). That is, the dimensionless quantity

$$\epsilon_{\text{multi}} = \omega a/c \quad \text{multipole parameter} \quad (43.8)$$

is much smaller than 1.

In that case, we may replace the exponential inside the integral by its Taylor series: $1 - i\epsilon_{\text{multi}}(\hat{r} \cdot \vec{r}_*/a) + \dots$. Making this approximation, and truncating after a finite number of terms, is called **multipole approximation**. 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 a/r and ϵ_{multi} , and from now on in this chapter will keep only the leading nonzero term, which for a generic source is electric dipole.

43.3 ELECTRIC DIPOLE RADIATION

43.3.1 A time-varying ED moment leads to $1/r$ potentials

Equation 43.7 has become

$$\underline{A}^\mu(t, \vec{r}) = \frac{1}{2} \frac{\mu_0}{4\pi r} e^{-i\omega(t-r/c)} \underbrace{\int d^3r_* \bar{J}^\mu(\vec{r}_*)}_{\text{multipole expansion}} + \text{c.c.} = \frac{\mu_0}{4\pi r} \int d^3r_* \underline{J}^\mu(t_c, \vec{r}_*). \quad (43.9)$$

In this expression, t_c is shorthand for $t - r/c$. This quantity is simpler than the retarded time from Chapter 41,⁵ because we dial back the time based on the *center* of the distribution, not the location of any particular charge. In particular, t_c does not depend on \vec{r}_* .

We can now get an even simpler formula⁶ for the spatial components of \underline{A} . First, the divergence theorem implies

$$\int d^3r_* \vec{\nabla}_i(\vec{r}_m \vec{j}_i)|_{\vec{r}_*} = 0$$

for each of $m = 1, 2, 3$. (Remember that $\vec{j} \rightarrow 0$ outside the finite region where the source is located.) So

$$\int d^3r_* \delta_{im} \vec{j}_i(\vec{r}_*) = - \int d^3r_* \vec{r}_{*m} \vec{\nabla} \cdot \vec{j}|_{\vec{r}_*} = + \int d^3r_* \vec{r}_{*m} \frac{\partial}{\partial t} \rho_q(\vec{r}_*) = \frac{d}{dt} \vec{D}_{E,m}. \quad (43.10)$$

⁴If the charges are oscillating or doing circular motion, this condition says that their speed $\approx \omega a$ is much smaller than c . This is certainly true of electrons in a radiating atom or molecule, or in a radio antenna.

⁵Section 41.5.1 (page 543).

⁶The 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.

The final step made use of the definition of electric dipole moment $\vec{\mathcal{D}}_E$.

Overall, Equation 43.10 says that $\int d^3r_* \vec{j}_m = \frac{d}{dt} \vec{\mathcal{D}}_{E,m}$. So the three spatial components of Equation 43.9 reduce to

$$\vec{A}^{[ED]}(t, \vec{r}) = \frac{\mu_0}{4\pi r} \frac{d\vec{\mathcal{D}}_E}{dt} \Big|_{t-r/c}. \quad \text{ED approximation, small source} \quad (43.11)$$

Again, note that the derivative is to be evaluated at time $t_c = t - r/c$, not the retarded time from Chapter 41.

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 $\vec{\mathcal{D}}_E$. In this limit, the ED approximation really does become exact, and it recovers 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 506). Relate the constant ξ used then to \mathcal{D}_E used in this chapter.

Your Turn 43B

Evaluate \underline{A}^0 using Equation 43.9 and check that it agrees with Your Turn 38Aa (page 507).

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 \vec{E} and \vec{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).⁷ 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.

$$\vec{B}_k = \varepsilon_{kmi} \vec{\nabla}_m \vec{A}_i = \varepsilon_{kmi} \frac{\mu_0}{4\pi} \frac{\partial}{\partial \vec{r}_m} \left(\frac{1}{r} \frac{d\vec{\mathcal{D}}_{E,i}}{dt} \Big|_{t-r/c} \right).$$

The Chain Rule gives

$$= \varepsilon_{kmi} \frac{\mu_0}{4\pi r} \frac{d^2 \vec{\mathcal{D}}_{E,i}}{dt^2} \Big|_{t-r/c} (-\hat{r}_m/c) + \text{subleading}.$$

⁷The following derivation is essentially a solution to Your Turn 38B.

Even more compactly,

$$\vec{B}^{[ED]} = -\frac{\mu_0}{4\pi r c} \hat{r} \times \frac{d^2 \vec{\mathcal{D}}_E}{dt^2} \Big|_{t-r/c}. \quad \text{ED approx., far-field} \quad (43.12)$$

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.”
- 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 \hat{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 \underline{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 $d\vec{E}/dt = c^2 \vec{\nabla} \times \vec{B}$, 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 \hat{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: $\vec{\nabla}(t - r/c) = -\hat{r}/c$. So

$$\frac{d\vec{E}}{dt} = c^2 \left(-\frac{\mu_0}{4\pi r c} \right) \left(-\frac{\hat{r}}{c} \right) \times \left(\hat{r} \times \frac{d^3 \vec{\mathcal{D}}_E}{dt^3} \Big|_{t-r/c} \right).$$

Because everything is harmonic in time, we can just drop one time derivative from both sides of this equation:

$$\vec{E}^{[ED]} = \frac{\mu_0}{4\pi r} \hat{r} \times \left(\hat{r} \times \frac{d^2 \vec{\mathcal{D}}_E}{dt^2} \Big|_{t-r/c} \right). \quad \text{ED approx., far-field}$$

Like \vec{B} , the electric field is transverse to the line of sight \hat{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

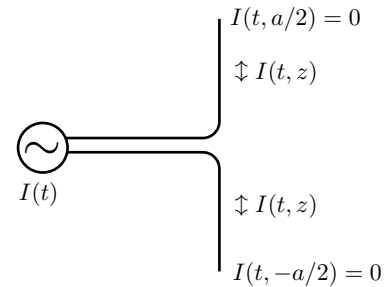


Figure 43.2: Center-fed linear microwave antenna.

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

$$I(t, z) = \bar{I} \cos(\omega t) (1 - |z|/(a/2)) \quad \text{for } |z| < a/2.$$

Current is 1D charge flux, so the 1D continuity equation says

$$\frac{d\rho_q^{[1D]}}{dt} = -\frac{dI}{dz} = -(\bar{I} \cos \omega t)(\pm 2/a)$$

for the upper and lower wires respectively. Thus, $\rho_q^{[1D]} = \pm \frac{2\bar{I}}{a\omega} \sin \omega t$.

We can now find the dipole moment:

$$\vec{D}_E = \hat{z} \sin \omega t \left[\int_{-a/2}^0 z dz \frac{-2\bar{I}}{a\omega} + \int_0^{a/2} z dz \frac{2\bar{I}}{a\omega} \right] = \hat{z} \frac{\bar{I}a}{4\omega} \sin \omega t.$$

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 Greenhouse gases absorb and radiate via molecular dipole moments

Absorption and emission by single molecules should properly be treated quantum mechanically (Chapter 56); however, some qualitative features can be understood in our classical picture.

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.

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 O_2 and N_2 , 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 light in the infrared region. 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 (H_2O), also has a permanent dipole moment; water vapor is also a potent infrared-active gas. The carbon dioxide molecule 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 an oscillating dipole moment in some of its vibrational states; transition dipoles therefore exist for these and also for mixed rotation-vibration states, making CO_2 another infrared-active gas (Figure 43.3).

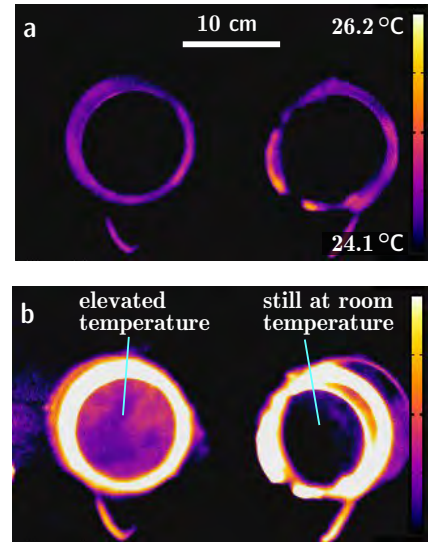
Homonuclear molecules have little infrared activity, but more complex ones can be strong greenhouse gases.

43.6 ENERGY FLUX AND TOTAL POWER SCALE AS ω^4

In this section, we'll sometimes drop the suffix “ $|_{t-r/c}$ ” for brevity. We continue to work in the far field, in electric dipole approximation.

⁸See Chapter 56.

Figure 43.3: [Infrared photographs.] **Energy absorption by an IR-active gas.** Two identical, cylindrical chambers with transparent ends, viewed in the wavelength band $7.5\text{--}14\ \mu\text{m}$. False color indicates radiance in this band (reds are higher than blues); the scale bar is labeled with approximate inferred temperature values in degrees Celsius. The chamber on the right contains dry air. The one on the left contains carbon dioxide. Both have axial length $23\ \text{cm}$ 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 15 = youtu.be/0eI9zxZoiPA and Sieg et al., 2019.]



Now at last we can see how energy is transported: Its flux is

$$\vec{S}^{[ED]} = \frac{1}{\mu_0} \vec{E} \times \vec{B} = -\mu_0^{-1} \left(\frac{\mu_0}{4\pi r} \right)^2 \frac{1}{c} \left[\hat{r} \times \left(\hat{r} \times \frac{d^2 \vec{\mathcal{D}}_E}{dt^2} \right) \right] \times \left[\hat{r} \times \frac{d^2 \vec{\mathcal{D}}_E}{dt^2} \right].$$

The factor in the brace is $\hat{r}(\hat{r} \cdot \frac{d^2 \vec{\mathcal{D}}_E}{dt^2}) - \frac{d^2 \vec{\mathcal{D}}_E}{dt^2}$. Now use the triple cross product formula again:

$$\vec{S} = -\mu_0^{-1} \left(\frac{\mu_0}{4\pi r} \right)^2 \frac{1}{c} \left(\hat{r} \left[(\hat{r} \cdot \frac{d^2 \vec{\mathcal{D}}_E}{dt^2}) - \frac{d^2 \vec{\mathcal{D}}_E}{dt^2} \right] \cdot \frac{d^2 \vec{\mathcal{D}}_E}{dt^2} - \frac{d^2 \vec{\mathcal{D}}_E}{dt^2} \left[(\hat{r} \cdot \frac{d^2 \vec{\mathcal{D}}_E}{dt^2}) - \frac{d^2 \vec{\mathcal{D}}_E}{dt^2} \right] \cdot \hat{r} \right)$$

$$\vec{S}^{[ED]} = \hat{r} \frac{\mu_0}{(4\pi r)^2} \frac{1}{c} \left(\left\| \frac{d^2 \vec{\mathcal{D}}_E}{dt^2} \Big|_{t-r/c} \right\|^2 - \left(\hat{r} \cdot \frac{d^2 \vec{\mathcal{D}}_E}{dt^2} \Big|_{t-r/c} \right)^2 \right). \quad \text{far-field} \quad (43.13)$$

Thus, the energy flux vector always points radially outward. It's not spherically symmetric, however, because its magnitude depends on the direction \hat{r} to the observer.

The *total* power output is the rate at which energy passes through a large spherical shell:⁹

$$\mathcal{P}^{[ED]} = \lim_{B \rightarrow \infty} \int_{r=B} d^2 \vec{\Sigma} \cdot \vec{S}^{[ED]} = \frac{\mu_0}{(4\pi)^2 c} \frac{d^2 \vec{\mathcal{D}}_E}{dt^2} \Big|_{t-r/c} \cdot \left[\int d^2 \hat{r} (\vec{\mathbf{1}} - \hat{r} \hat{r}) \right] \cdot \frac{d^2 \vec{\mathcal{D}}_E}{dt^2} \Big|_{t-r/c}.$$

The first term inside the square brackets is the integral over all directions of a constant tensor, that is, $4\pi \vec{\mathbf{1}}$. The second term is -4π times the average over all directions of

⁹Because we only want energy that makes it all the way out to infinity, the far-field approximation is automatically satisfied.

$\hat{r}\hat{r}$. 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 $-\int d^2\hat{r} = -4\pi$, which fixes the constant to be $-1/3$. All together, then, the factor in square brackets is $4\pi(1 - \frac{1}{3})\vec{\mathbb{I}}$, and we have

$$\mathcal{P}^{[ED]} = \frac{\mu_0}{4\pi c} \frac{2}{3} \left\| \frac{d^2\vec{\mathcal{D}}_E}{dt^2} \Big|_{t-r/c} \right\|^2. \quad \text{total power output, ED approximation} \quad (43.14)$$

43.7 LINEAR POLARIZATION RECOVERS THE DIPOLE DOUGHNUT PATTERN

Consider the case in which $\vec{\mathcal{D}}_E$ is always directed along a single direction as in Figure 43.2. We can choose coordinates to make that direction be the z -axis: $\vec{\mathcal{D}}_E = \mathcal{D}_E(t)\hat{z}$. First note a relation between the spherical directions:

$$\hat{z} = \hat{r} \cos \theta - \hat{\theta} \sin \theta.$$

Your Turn 43C

Show that

$$\begin{aligned} \vec{B}^{[ED]} &= \hat{\phi} \frac{\mu_0}{4\pi r c} \frac{d^2\mathcal{D}_E}{dt^2} \Big|_{t-r/c} \sin \theta \\ \vec{E}^{[ED]} &= \hat{\theta} \frac{\mu_0}{4\pi r} \frac{d^2\mathcal{D}_E}{dt^2} \Big|_{t-r/c} \sin \theta. \end{aligned}$$

Hence, in any direction, a distant observer sees a linearly polarized plane wave.

Turning now to the energy flux,

$$\begin{aligned} \left\| \frac{d^2\vec{\mathcal{D}}_E}{dt^2} \right\|^2 &= \left[\frac{d^2\mathcal{D}_E}{dt^2} \right]^2 \\ \left(\hat{r} \cdot \frac{d^2\vec{\mathcal{D}}_E}{dt^2} \right)^2 &= \left(\frac{d^2\mathcal{D}_E}{dt^2} \mathcal{D}_E \hat{r} \cdot \hat{z} \right)^2 = \left[\frac{d^2\mathcal{D}_E}{dt^2} \right]^2 \cos^2 \theta \\ \vec{S}^{[ED]} = \mu_0^{-1} \vec{E} \times \vec{B} &= \hat{r} \frac{\mu_0}{(4\pi r)^2} \frac{1}{c} \left[\frac{d^2\mathcal{D}_E}{dt^2} \Big|_{t-r/c} \right]^2 \sin^2 \theta. \end{aligned} \quad (43.15)$$

Equation 43.15 shows the angular dependence explicitly: Energy mostly comes out near the equatorial plane (here the xy plane).

¹⁰You showed this in Problem 14.2.

We can now get the total power output from Equation 43.14. If the dipole varies harmonically in time, then we can write $\mathcal{D}_E(t)$ in terms of the amplitude (maximum value) $\bar{\mathcal{D}}_E$ as $\mathcal{D}_E(t) = \frac{1}{2}e^{-i\omega t}\bar{\mathcal{D}}_E + \text{c.c.}$ Then the time-averaged power output is

$$\langle \mathcal{P}^{[ED]} \rangle = \frac{\mu_0}{12\pi c} \omega^4 |\bar{\mathcal{D}}_E|^2, \quad \begin{array}{l} \text{total power output,} \\ \text{harmonic source} \end{array} \quad (43.16)$$

a famous result.

Your Turn 43D

Repeat the exercise, but with $\vec{\mathcal{D}}_E(t) = \bar{\mathcal{D}}_E \begin{pmatrix} \cos \omega t \\ \sin \omega t \\ 0 \end{pmatrix}$ and interpret the result.

FURTHER READING

Greenhouse gases:

Bohren & Clothiaux, 2006, chap. 2.

PROBLEMS

43.1 Beyond far-field approximation

Background: The main text derived an expression for the exact vector potential outside an arbitrary localized charge/current distribution. Then we simplified the result by assuming harmonic time dependence of the sources and far field regime (blue triangular region in Figure 43.1, page 561), so that we could discard all $\mathcal{O}(r^{-2})$ terms in the fields as well as all but the leading term in an expansion in powers of a/λ , 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/ω (region marked “near and intermediate” in Figure 43.1); for example, we may consider the limit of an oscillating pure (point) dipole ($a \rightarrow 0$), with

$$\vec{\mathcal{D}}_{\text{E}}(t) = \hat{z}\bar{\mathcal{D}}_{\text{E}} \cos(\omega t).$$

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 510), 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 $\hat{E} = \vec{E} / \|\vec{E}\|$ and superimpose a heat map of $\|\vec{E}\|$ as in Figure 42.2 (page 556). The normalized vector field has the same streamlines as \vec{E} .

- b. Next, try nonzero frequency: Assume $\omega = 2\pi c/(1.5 \text{ cm})$. Again plot \vec{E} at time $t = 0$.

- c. Continuing (b), make another plot with a lot of representative streamlines of \vec{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. Suppose 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 38.3 (page 510) described how EM waves can carry angular momentum: The density of angular momentum \vec{J}_z , computed using the origin as reference point, is $\hat{z} \cdot \frac{1}{\mu_0} [\vec{r} \times (\vec{E} \times \vec{B})]$. As usual, we will suppose that the fields are harmonically varying in time and consider only the time average of our answers.

Do:

- Suppose that we have two oscillating dipoles of strength \vec{D}_E at the origin, pointing at right angles to each other and both in the xy plane. The dipoles oscillate at the same angular frequency ω 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 \vec{J}_z is then your answer divided by c .
- 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 \vec{J}_z ? Why/why not? Now do the calculation using (a), to get the rate at which \vec{J}_z is transferred to the sphere.
- Also find the power absorbed by the sphere.
- Divide your answers to (b,c) and comment.

43.3 [Not ready yet.]

CHAPTER 44

Higher-Multipole Radiation

44.1 FRAMING: SUPPRESSION

Chapter 43 found a solution to Maxwell's 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 in the 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 ω :

$$\underline{A}^\mu(t, \vec{r}) = \frac{\mu_0}{4\pi r} e^{-i\omega(t-r/c)} \int d^3r_* e^{-i\epsilon_{\text{multi}} \hat{r} \cdot \vec{r}_* / a} \frac{1}{2} \underline{J}^\mu(\vec{r}_*) + \text{c.c.} \quad [43.7, \text{page 562}]$$

Recall that in this formula, t and \vec{r} (and hence also \hat{r}) refer to the observation, whereas \vec{r}_* is a source point. Again, we restrict to the multipole regime: ϵ_{multi} is the small quantity controlling the multipole expansion (Equation 43.8, page 563), and a is the overall source size (upper bound on $\|\vec{r}_*\|$). Now, however, we will retain higher terms in the expansion in ϵ_{multi} that were dropped in Chapter 43.

44.2.1 Order-one terms in ϵ_{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 $\vec{A}^{[0]}$; instead, now we focus on first order in ϵ_{multi} . We'll call the three spatial components

of that term $\vec{A}^{[1]}$:

$$\vec{A}^{[1]}(t, \vec{r}) = \frac{\mu_0}{4\pi r} e^{-i\omega(t-r/c)} \int d^3r_* (-i\epsilon_{\text{multi}} \hat{r} \cdot \vec{r}_*/a) \frac{1}{2} \vec{j}(\vec{r}_*) + \text{c.c.}$$

We can write $-i\epsilon_{\text{multi}}/a$ as $c^{-1} \frac{d}{dt}$:

$$= \frac{\mu_0}{4\pi r c} \hat{r} \cdot \frac{d}{dt} \left[\underbrace{\int d^3r_* \vec{r}_* \otimes \vec{j}(t-r/c, \vec{r}_*)}_{\vec{\Gamma}(t_c)} \right].$$

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 antisymmetric pieces, which we'll call

$$\vec{\Gamma} = \vec{\Gamma}^{[EQ]} + \vec{\Gamma}^{[MD]} \quad (44.1)$$

respectively.

44.2.2 Antisymmetric part of the moment: magnetic dipole radiation

Like any antisymmetric second-rank 3-tensor, we may re-express the three independent entries of $\vec{\Gamma}^{[MD]}$ in terms of a single pseudovector:

$$\vec{\Gamma}_{np}^{[MD]} = \epsilon_{npi} \vec{\mathcal{D}}_{M,i} \quad \text{where} \quad \vec{\mathcal{D}}_{M,i} = \frac{1}{2} \epsilon_{iks} \int d^3r_* \vec{r}_{*k} \vec{j}_s. \quad [17.6, \text{page 245}]$$

Your Turn 44A

Show that $\vec{\Gamma}^{[MD]}$ contributes

$$\vec{A}^{[MD]} = -\frac{\mu_0}{4\pi r c} \hat{r} \times \frac{d}{dt} \vec{\mathcal{D}}_M \Big|_{t-r/c} \quad \text{MD approximation, small source}$$

to $\vec{A}^{[1]}$.

Your result implies that

- Once again, this part of the field is a spherical wave (because the wave crests of $\vec{A}^{[MD]}$ lie on the spherical shells $ct - r = \pi nc/\omega$ for integer n).
- $\vec{A}^{[MD]}$ 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 560)):

Your Turn 44B

a. Do a calculation similar to the one in Section 43.4 to show that

$$\vec{B}^{[MD]} = \frac{\mu_0}{4\pi r c^2} \hat{r} \times \left(\hat{r} \times \frac{d^2}{dt^2} \vec{\mathcal{D}}_M \Big|_{t-r/c} \right). \quad \text{far field} \quad (44.2)$$

b. Then use Ampère's law to find $\vec{E}^{[MD]}$.

Remarkably,

The MD contribution to the magnetic far field looks just like the ED contribution to the electric field. The MD contribution to the electric far field looks just like the ED contribution to the magnetic field.

Consider a circular loop of wire in the xy plane, with area Σ and carrying current with amplitude \bar{I} and angular frequency ω . It has no net charge anywhere, and hence vanishing electric dipole and quadrupole moments. But you found in Your Turn 17A (page 245) that the magnetic dipole moment is nonzero: $\vec{\mathcal{D}}_M = (\hat{z}\Sigma)(\bar{I} \cos \omega t)$.

Your Turn 44C

- Find the far electric and magnetic fields and compare to your earlier result obtained in Coulomb gauge (Your Turn 25E, page 335).
- Find the Poynting vector and compare with the result in Coulomb gauge (Your Turn 25E).
- Integrate the Poynting vector over all directions \hat{r} to obtain the power \mathcal{P} .
- Time-average your result from (c) to show that

$$\langle \mathcal{P}^{[MD]} \rangle = \left(\frac{\omega}{c}\right)^4 \frac{(\bar{I}\Sigma)^2}{12\pi\epsilon_0 c}. \quad (44.3)$$

44.2.3 Symmetric part of the moment: electric quadrupole radiation

Next we turn to the first term of Equation 44.1. To simplify $\vec{\Gamma}^{[EQ]}$, we now use a trick remembered from magnetostatics (Section 17.2): The divergence theorem gives that

$$0 = \int d^3r_* \vec{\nabla}_{*i} (\vec{r}_{*k} \vec{r}_{*m} \vec{j}_i(\vec{r}_*)),$$

where $\vec{\nabla}_*$ denotes partial derivatives with respect to \vec{r}_* . Thus,

$$\vec{\Gamma}_{mk}^{[EQ]} = \frac{1}{2} \int d^3r_* (\vec{r}_{*m} \vec{j}_k + \vec{r}_{*k} \vec{j}_m) = -\frac{1}{2} \int d^3r_* \vec{r}_{*m} \vec{r}_{*k} \vec{\nabla} \cdot \vec{j}.$$

For a static current distribution, this quantity would be zero by the continuity equation. More generally, however, we get

$$= \frac{1}{2} \frac{d}{dt} \int d^3r_* \vec{r}_{*m} \vec{r}_{*k} \rho_q \Big|_{t-r/c}.$$

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 37):

$$\frac{1}{3} \vec{\mathcal{Q}}_{E,ml} \Big|_{t-r/c} + \frac{1}{3} \mathbb{I}_{ml} \int d^3r_* r_*^2 \rho_q \Big|_{t-r/c}.$$

So the contribution of $\vec{\Gamma}^{[EQ]}$ to the first-order term of the vector potential, $\vec{A}^{[1]}$, can be written as

$$\vec{A}^{[EQ]} = \frac{1}{6} \frac{\mu_0}{4\pi r c} \frac{d}{dt} \hat{r} \cdot \left(\frac{d}{dt} \vec{\mathcal{Q}}_E \Big|_{t-r/c} + \mathbb{I} \int d^3r_* r_*^2 \frac{d}{dt} \rho_q(t-r/c, \vec{r}_*) \right)$$

$$= \frac{\mu_0}{24\pi c} \left[r^{-1} \hat{r} \cdot \frac{d^2 \vec{Q}_E}{dt^2} \Big|_{t-r/c} + \hat{r} r^{-1} \int d^3 r_* r_*^2 \frac{d}{dt} \rho_q(t-r/c, \vec{r}_*) \right].$$

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

$$\vec{A}^{[EQ]} = \frac{\mu_0}{24\pi c r} \hat{r} \cdot \frac{d^2 \vec{Q}_E}{dt^2} \Big|_{t-r/c} \quad \text{far field} \quad (44.4)$$

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 $\epsilon_{\text{multi}} = \omega a/c$, but it can be the leading term for a source with dipole moments everywhere equal to zero.

44.3 HIGHER ORDER TERMS CAN CONTRIBUTE EVEN IF LOWER ONES ARE ZERO

44.3.1 Magnetic dipole and electric quadrupole contributions can also transport energy to infinity

Clearly we could carry out the expansion to next order in ϵ_{multi} 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

$$\frac{d}{dt} \vec{E} = c^2 \vec{\nabla} \times \vec{B} \approx -c \hat{r} \times \vec{B}$$

so the energy flux is

$$\vec{S} = \frac{1}{\mu_0} \vec{E} \times \vec{B} = \frac{c}{\mu_0} \vec{B} \times (\hat{r} \times \vec{B}) = \frac{c}{\mu_0} \hat{r} \|\vec{B}\|^2,$$

and each nonzero term of

$$\|\vec{B}^{[0]} + \vec{B}^{[1]} + \vec{B}^{[2]} + \dots\|^2 \quad (44.5)$$

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 568) gave the $\|\vec{B}^{[0]}\|^2$ term (electric

dipole), and Equation 43.16 (page 570) 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.¹ So the next most important terms involve $\|\vec{B}^{[1]}\|^2$ (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, N^{13} has a state that emit a photon with half-life $T_{\text{N}} \approx 10^{-15}$ s, whereas Hg^{197} has a state with half-life $T_{\text{Hg}} \approx 7 \cdot 10^{-9}$ s.

Nuclear radiative lifetimes vary widely.

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 N^{13} may transition via its electric dipole moment, whereas Hg^{197} can only emit via electric quadrupole radiation.

Gamma ray emission is quantum mechanical in character (usually just one photon is emitted).² 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 dipole radiation, Equation 43.14 (page 569) gives

$$\text{transition rate} = \frac{c^{-1}\mu_0 e^2 a^2 \omega^4}{\hbar\omega}.$$

The present chapter has shown that quadrupole radiation is suppressed by an additional factor of $(a\omega/c)^2$. We thus have

$$\frac{\text{transition rate } \text{N}^{13}}{\text{transition rate } \text{Hg}^{197}} \approx \frac{a_{\text{N}}^2 \omega_{\text{N}}^4}{\hbar\omega_{\text{N}}} \bigg/ \frac{a_{\text{Hg}}^4 \omega_{\text{Hg}}^6 c^{-2}}{\hbar\omega_{\text{Hg}}}.$$

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 just given, using observed energies for these transitions: $\hbar\omega_{\text{N}} = 2.38 \text{ MeV}$ and $\hbar\omega_{\text{Hg}} = 0.13 \text{ MeV}$. (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.

¹See Problem 44.5.

²See Chapter 56.

44.4 PLUS ULTRA

A *spherically symmetric* charge distribution will not radiate, no matter how it depends on time. For example, its monopole moment is fixed by charge conservation, and hence has vanishing time dependence. We also saw above how the first orders of the expansion involve $\vec{\mathcal{D}}_E$, $\vec{\mathcal{D}}_M$, $\vec{\mathcal{Q}}_E$, and so on, all of which are zero for a spherically symmetric distribution.

FURTHER READING

Intermediate:

General: Zangwill, 2013, §20.8.

Anapole radiation: Rovenchak & Krynytskyi, 2018.

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 $\vec{r} = (L, 0, 0)$, and we only asked for the leading order term in powers of $1/L$. 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 b . However, each loop also has a larger current than the previous one, in such a way that the magnetic dipole moment $\mathcal{D}_M(t) = \bar{\mathcal{D}}_M \cos(\omega t)$ is the same for all.³ In this small source limit (and also the far-field limit), 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 $\pm \frac{1}{2}(\bar{I}e^{i\omega t} + \text{c.c.})$ 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.]

³This “pure dipole” limit is similar to the one in Section 43.3.2 (page 564).

44.3 Pulsar

A pulsar is a compact star (specifically a neutron star), with a large magnetic dipole moment $\vec{\mathcal{D}}_M$. As with an ordinary permanent magnet, the moment is frozen into the star. That is, $\vec{\mathcal{D}}_M$ rotates with the star's angular velocity ω . 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 α relative to the rotation axis. The moment \mathcal{D}_M is not directly observable, but it is related to the strength of B_{pole} of the magnetic field at the star's surface, which is indirectly observable via the Zeeman splitting of spectral lines (Problem 18.8 (page 281)). So let $\|\vec{\mathcal{D}}_M\| = \kappa B_{\text{pole}}$, where κ is a constant.

- Find the rate at which the pulsar radiates electromagnetic energy, in terms of κ , B_{pole} , ω , and α .
- If the source of the energy is the pulsar's rotational kinetic energy, $\mathcal{E} = \frac{1}{2}J\omega^2$ with $J =$ pulsar's moment of inertia, find the characteristic slowdown time scale $T \equiv -\frac{\omega}{d\omega/dt}$ at time zero, as a function of κ , B_{pole} , ω , and α , and J .
- Suppose that the pulsar has radius R , and derive a formula for κ in terms of R .
- Thus, we get a prediction of the slowdown in terms of B_{pole} , R , ω , α , 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 τ for $\omega(0) = 10^4$ s $^{-1}$, a frequency thought to be typical of newly formed pulsars.

44.4 $\boxed{T_2}$ Exact MD wave

First do Problem 44.1. In this problem, however, find the \vec{E} and \vec{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 $a \rightarrow 0$ 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, analogous to the one in Chapter 38 (Problem 38.1 (page 510)).]

44.5 $\boxed{T_2}$

Using Equations 43.12 (page 565), 44.2, and 44.4, show that the cross term $2\vec{B}^{[0]} \cdot \vec{B}^{[1]}$ integrated over angles gives zero. Thus, there is no term in the total radiated power that is first order in the small parameter ϵ_{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. There's no such thing, so this average must equal zero.]

44.6 $\boxed{T_2}$ Electric quadrupole radiation

When we expand Equation 44.5, the term $\|\vec{B}^{[1]}\|^2$ includes the cross term $2\vec{B}^{[EQ]} \cdot \vec{B}^{[MD]}$. 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 44.3 (page 575), plus one other subterm that you are to find.

CHAPTER 45

Synchrotron Radiation

45.1 FRAMING: *BEAMING*

This chapter will explore one particular example of radiation by an accelerating charge: the case of uniform circular motion. 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.
- 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.”

The analysis in Chapter 43 may seem to lead to some qualitative conclusions: A charge executing an circular orbit of radius b has a rotating electric dipole moment, so we might expect the emitted radiation to be predominantly circularly polarized and directed perpendicular to the charge's orbital plane, but with a broad angular distribution characteristic of electric dipole radiation. Moreover, the centripetal acceleration $\omega^2 b = (\omega b)^2/b$ may never exceed c^2/b , so we might expect a limiting amount of radiation.

Remarkably, both of the predictions just made fail as we push the charge up to relativistic speeds. Perhaps this should not be surprising—the multipole approximation 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 important for technological applications like crystallography.

45.2 THE LIÉNARD–WEICHERT FIELDS

We wish to find the electromagnetic far fields associated to a point charge in uniform circular motion. For this, we must compute some derivatives of the Liénard–Weichert potentials. Section 41.5.2 (page 545) 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.

¹Section 43.2.3.

At first, the potentials look pretty simple:

$$\underline{A}(\underline{X}) = \frac{\mu_0 q c}{4\pi} \underline{U}_r (-\underline{X} - \underline{\Gamma}(\tau_r))_\mu \underline{U}_r^\mu)^{-1}. \quad [41.11, \text{page 544}]$$

But recall that in addition to their explicit dependence on the observation point \underline{X} , the potentials involve the retarded proper time τ_r , which also implicitly depends on \underline{X} . Fortunately, we only need the variation of τ_r in order to take the first derivatives needed for the Faraday tensor. So let's vary the defining relation (Section 41.5.1, page 543), which is that $\underline{\Gamma}(\tau_r)$ must lie on the past light cone of \underline{X} :

$$0 = \|\underline{X} - \underline{\Gamma}(\tau_r)\|^2 = \|\underline{X} + d\underline{X} - \underline{\Gamma}(\tau_r + d\tau_r)\|^2.$$

Let $\underline{\Gamma}_r = \underline{\Gamma}(\tau_r)$ and collect the first-order terms:

$$\begin{aligned} 0 &= (\underline{X} - \underline{\Gamma}_r)^\mu (d\underline{X} - d\tau_r \underline{U}_r)_\mu \\ d\tau_r &= \frac{(\underline{X} - \underline{\Gamma}_r)_\mu d\underline{X}^\mu}{(\underline{X} - \underline{\Gamma}_r)_\nu \underline{U}_r^\nu} \quad \text{or} \quad \frac{\partial \tau_r}{\partial \underline{X}^\mu} = \frac{(\underline{X} - \underline{\Gamma}_r)_\mu}{(\underline{X} - \underline{\Gamma}_r)_\nu \underline{U}_r^\nu}. \end{aligned} \quad (45.1)$$

Introduce the abbreviation

$$\alpha(\underline{X}, \tau) = -(\underline{X} - \underline{\Gamma}(\tau))_\nu \underline{U}^\nu(\tau). \quad (45.2)$$

Then the derivatives of Equation 41.11 become

$$\frac{4\pi}{\mu_0 q c} \partial^\mu \underline{A}^\nu = \partial^\mu \left(\frac{\underline{U}^\nu(\tau_r(\underline{X}))}{\alpha(\underline{X}, \tau_r(\underline{X}))} \right).$$

In the preceding formula, the derivative acts on every \underline{X} dependence. We now reformulate by using the chain rule:

$$\begin{aligned} &= -\frac{\underline{U}_r^\nu}{\alpha_r^2} \frac{\partial \alpha}{\partial \underline{X}_\mu} \Big|_r + \frac{\partial \tau_r}{\partial \underline{X}_\mu} \frac{\partial}{\partial \tau} \Big|_r \left(\frac{\underline{U}^\nu}{\alpha} \right) \\ &= \left[\alpha_r^{-2} \underline{U}_r^\nu \underline{U}_r^\mu - \alpha_r^{-1} (\underline{X} - \underline{\Gamma}_r)^\mu \frac{\partial}{\partial \tau} \Big|_r \left(\frac{\underline{U}^\nu}{\alpha} \right) \right]. \end{aligned} \quad (45.3)$$

The Faraday tensor (here called the **Liénard–Weichert fields**) is this expression minus ($\mu \rightleftharpoons \nu$), so we need not keep the symmetric first term.

Next, recall that Section 41.5.1 (page 543) made α more explicit. As usual, let $\vec{\beta} = d\vec{r}/d(ct)$ and $\gamma = (1 - \vec{\beta}^2)^{-1/2}$. Also let $t_r = \underline{\Gamma}^0(\tau_r)$, $\vec{R}_r = \vec{r} - \vec{\Gamma}_r$, and so on. Thus, t_r is the lab time at which a disturbance destined to be observed at \underline{X} is created by the charge. Then Equation 41.10 (page 544) gave that

$$\alpha_r = c\gamma_r R_r (1 - \hat{R}_r \cdot \vec{\beta}_r). \quad (45.4)$$

Two qualitative features of this formula will be important:

- $1/\alpha$ falls with distance as $1/R_r$.
- For a fast particle ($\beta_r \approx 1$), $1/\alpha$ is sharply peaked in the direction of the particle's retarded velocity.

The first point suggests that we may find fields with the slow falloff characteristic of radiation. The second 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 “rotating searchlight” pattern. The following analysis will confirm these physical expectations.

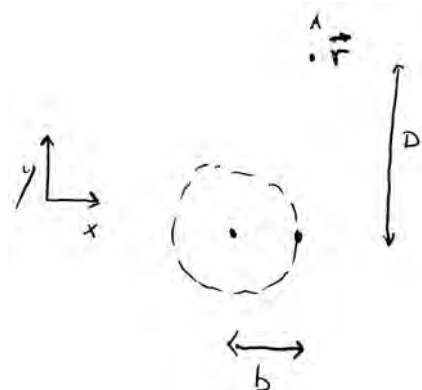


Figure 45.1: Variables defined in the text. The observation point \vec{r} (top) is offset horizontally by the radius b of the circular orbit. The dot on the orbit denotes the position of a charged particle when the lab time, and its proper time, are both zero. The observation is made later, at lab times close to $t_1 = D/c$.

45.3 UNIFORM CIRCULAR MOTION

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 charge in uniform circular motion in the xy plane (Figure 45.1).

Accordingly, take the trajectory to be

$$\vec{\Gamma}(t_*) = b(\hat{x} \cos(\omega t_*) + \hat{y} \sin(\omega t_*)).$$

This motion has the nice feature that $\beta = b\omega/c$, $\gamma = (1 - \beta^2)^{-1/2}$, and $\underline{U}^0 = \gamma c$ are all constants; hence $t_* = \gamma\tau$ is a linear relation. Also,

$$\vec{\beta}(\tau) = \beta(-\hat{x} \sin(\omega\gamma\tau) + \hat{y} \cos(\omega\gamma\tau)) \quad \text{and} \quad \vec{U} = \gamma c \vec{\beta}. \quad (45.5)$$

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 $\vec{r} = b\hat{x} + D\hat{y}$, where D is some large, constant distance (Figure 45.1). 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 taken 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 ω^{-1} . We have some freedom in how we nondimensionalize time; it will be convenient to represent times as dimensionless quantities times $(\omega\gamma)^{-1}$:

$$\bar{\underline{X}} = \underline{X}/b, \quad \bar{\underline{\Gamma}} = \underline{\Gamma}/b, \quad \bar{D} = D/b, \quad \bar{t} = \omega\gamma t, \quad \bar{\tau} = \omega\gamma\tau.$$

To summarize, we then have

$$\bar{\underline{\Gamma}}(\bar{\tau}) = \begin{bmatrix} \bar{\tau}/\beta \\ \cos \bar{\tau} \\ \sin \bar{\tau} \end{bmatrix} \quad \text{and} \quad \bar{\underline{X}} = \begin{bmatrix} \bar{t}/(\beta\gamma) \\ 1 \\ \bar{D} \end{bmatrix}, \quad (45.6)$$

where $\bar{\tau}$ is dimensionless proper time for the particle and \bar{t} is dimensionless observation time. Equation 45.6 has been abbreviated by dropping the fourth entry in each 4-vector, because in each case it is the constant zero. Finally, it will be convenient to nondimensionalize α_r :

$$\bar{\alpha} = \alpha_r \omega / (\gamma c^2).$$

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. To confirm this idea, and predict it quantitatively, we will examine a small range of observer times t close to some initial time t_i . According to our guess, t_i should be chosen such that the retarded proper time is $\tau_{r,i} = 0$; that way, the retarded particle velocity at t_i will be directed along \hat{y} , that is, parallel to the line of sight of the observer in Figure 45.1. Then the light-cone condition says $\|\underline{X}_i - \underline{\Gamma}_{r,i}\| = 0$ or

$$c(t_i - \gamma \tau_{r,i}) = \|b\hat{x} + D\hat{y} - b\hat{x}\| = D, \quad \text{or} \quad t_i = D/c.$$

To evaluate formulas like Equation 45.2, we must find the retarded proper times τ_r corresponding to lab times $t \approx t_i$. Although the relation is complicated, we will later confirm that we only need it for times very close to t_i . Hence we may expand to first order about that moment and use Equation 45.1:

$$\begin{aligned} \tau_r &= \tau_{r,i} + c(t - t_i) \left. \frac{\partial \tau_r}{\partial X^0} \right|_i + \dots \approx -c(t - t_i) \frac{ct_i}{-ct_i U_{r,i}^0 + (b\hat{x} + D\hat{y} - b\hat{x}) \cdot \vec{U}_{r,i}} \\ &= (t - t_i) \gamma^{-1} (1 - \beta)^{-1} \end{aligned} \quad (45.7)$$

In dimensionless variables,

$$\bar{\tau}_r \approx (\bar{t}/\gamma - \bar{D}\beta)(1 - \beta)^{-1}. \quad (45.8)$$

Reassuringly, in the limit $\omega \rightarrow 0$ this formula says that retarded proper time advances as a constant plus observation time.

For circular motion, Equations 45.2, 45.5, and 45.6 give

$$\alpha(\underline{X}, \tau) = -b \left[\bar{t}/(\beta\gamma) - \bar{\tau}/\beta, 1 - \cos \bar{\tau}, \bar{D} - \sin \bar{\tau} \right] \begin{bmatrix} -1 & & \\ & 1 & \\ & & 1 \end{bmatrix} \gamma c \begin{bmatrix} -\beta \sin \bar{\tau} \\ \beta \cos \bar{\tau} \end{bmatrix}.$$

The velocity is always perpendicular to the displacement, so

$$\alpha = (c^2 \gamma / \omega) (\bar{t}/\gamma - \bar{\tau} + \beta^2 (\sin \bar{\tau} - \bar{D} \cos \bar{\tau})). \quad (45.9)$$

45.3.2 Electric field

We can now evaluate the derivatives needed in Equation 45.3:

$$\left. \frac{\partial}{\partial \tau} \right|_r \left(\frac{U^0}{\alpha} \right) = \frac{\gamma \omega^2}{c \bar{\alpha}^2} (1 - \beta^2 (\cos \bar{\tau}_r + \bar{D} \sin \bar{\tau}_r)). \quad (45.10)$$

Introduce abbreviations for some unit vectors in the xy plane:

$$\hat{G} = \begin{bmatrix} \cos \bar{\tau}_r \\ \sin \bar{\tau}_r \end{bmatrix}, \quad \hat{H} = \begin{bmatrix} -\sin \bar{\tau}_r \\ \cos \bar{\tau}_r \end{bmatrix}, \quad \hat{K} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Then

$$\frac{\partial}{\partial \tau} \Big|_{\mathbf{r}} \left(\frac{\vec{U}}{\alpha} \right) = -\gamma \beta \omega^2 c^{-1} [\bar{\alpha}^{-1} \hat{G} + \bar{\alpha}^{-2} \hat{H} (-1 + \beta^2 (\cos \bar{\tau}_r + \bar{D} \sin \bar{\tau}_r))]. \quad (45.11)$$

Substitute Equations 45.10–45.11 into Equation 45.3 to find the electric field $\vec{E}_i = c \underline{F}^{0i}$:

$$\begin{aligned} \frac{4\pi}{\mu_0 q c^2} \vec{E} &= \frac{\omega^2 \beta}{c^2} \bar{\alpha}^{-2} \left[(\bar{t}/\gamma - \bar{\tau}_r) (\hat{G} - \bar{\alpha}^{-1} \hat{H} (1 - \beta^2 (\cos \bar{\tau}_r + \bar{D} \sin \bar{\tau}_r))) \right. \\ &\quad \left. + \bar{\alpha}^{-1} (1 - \beta^2 (\cos \bar{\tau}_r + \bar{D} \sin \bar{\tau}_r)) \left[\frac{1 - \cos \bar{\tau}_r}{\bar{D} - \sin \bar{\tau}_r} \right] \right]. \end{aligned}$$

The formulas are getting long, so it is time to simplify by looking only at the far fields. First recall that $t = D/c + \Delta t$, so to leading order in large D , Equation 45.9 reduces to

$$\bar{\alpha} \approx \bar{D} \beta (1 - \beta \cos \bar{\tau}_r).$$

We introduce one last abbreviation by denoting the factor in parentheses by η , a dimensionless, D -independent quantity we will call the “beaming factor.” Thus,

$$\bar{\alpha} \approx \bar{D} \beta \eta. \quad \text{far field}$$

Next, the leading behavior of the electric field is

$$\begin{aligned} \frac{4\pi}{\mu_0 q c^2} \vec{E} &\approx (\omega/c)^2 (\eta^{-2} \bar{D}^{-1} \hat{G} + \eta^{-3} \bar{D}^{-1} \sin \bar{\tau}_r (\hat{H} \beta - \hat{K})) \\ &= (\omega/c)^2 \bar{D}^{-1} \left[\hat{G} \eta^{-2} + (\beta \hat{H} - \hat{K}) \sin \bar{\tau}_r \eta^{-3} \right]. \quad \text{far field} \end{aligned} \quad (45.12)$$

At last, we have found a compact, explicit formula for the electric far field, and we see that indeed it has a radiative component, that is, its leading far field falls as $1/D$. We will soon see that the peak radiation intensity occurs at $\bar{\tau}_r = 0$; here Equation 45.12 is polarized transversely to the line of sight (along \hat{x}), as befits a radiation field.

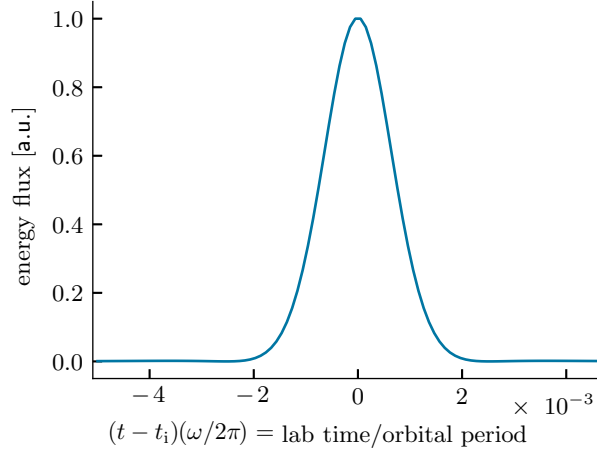
45.3.3 Magnetic fields

We can calculate $\vec{B}_k = \frac{1}{2} \varepsilon_{klm} \underline{F}_{lm}$ by using similar steps. The far fields are:

$$\begin{aligned} \frac{4\pi}{\mu_0 q c} \vec{B}_k &= \varepsilon_{klm} \frac{\omega^2 \beta^2}{c^2} (\bar{D} \beta \eta)^{-2} \left[\frac{1 - \cos \bar{\tau}_r}{\bar{D} - \sin \bar{\tau}_r} \right]_{\ell} \left[\hat{G} + (\bar{D} \beta \eta)^{-1} \hat{H} (-1 + \beta^2 (\cos \bar{\tau}_r + \bar{D} \sin \bar{\tau}_r)) \right]_m \\ &\approx \bar{D}^{-1} \frac{\omega^2}{c^2 \eta^2} \hat{K} \times (\hat{G} + (\beta \eta)^{-1} \hat{H} \beta^2 \sin \bar{\tau}_r) \\ &= \bar{D}^{-1} \frac{\omega^2}{c^2 \eta^2} (-\cos \bar{\tau}_r + \beta \eta^{-1} \sin^2 \bar{\tau}_r) \hat{z}. \end{aligned}$$

Again we find that the leading behavior is $\propto D^{-1}$. The magnetic far field is always polarized perpendicular to the orbital plane.

Figure 45.2: [Mathematical function.] **Time course of the far-field energy flux at a fixed position.** A charged particle with $\beta = 0.95c$ is viewed in the plane of its circular orbit. The graph shows the function Equation 45.13 in a brief window close to the pulse center. There is no appreciable energy flux for the other 99.6% of the orbital period.



45.3.4 Energy flux

The energy flux² observed at \underline{X} is given by the Poynting vector:

$$\vec{S}^i = \underline{T}^{0i} = -\mu_0^{-1} \underline{E}^{0j} \underline{E}_j^i = \mu_0^{-1} (\vec{E}/c \times \vec{B})_i$$

Using our far field formulas,

$$\vec{S} = \mu_0^{-1} \left(\frac{\mu_0 q \omega}{4\pi c} \right)^2 \bar{D}^{-2} [\eta^{-2} \hat{G} + \eta^{-3} (\beta \hat{H} - \hat{K}) \sin \bar{\tau}_r] \times \frac{\omega^2}{\eta^2} \hat{z} (-\cos \bar{\tau}_r + \beta \eta^{-1} \sin^2 \bar{\tau}_r).$$

Use the identities $\hat{G} \times \hat{z} = -\hat{H}$, $\hat{H} \times \hat{z} = \hat{G}$, and $\hat{K} \times \hat{z} = \hat{x}$ to get

$$\vec{S} = \frac{\mu_0 q^2 \omega^4}{(4\pi \bar{D})^2 c} [\hat{H} \eta^{-4} - (\hat{G} \beta - \hat{x}) \eta^{-5} \sin \bar{\tau}_r] (\cos \bar{\tau}_r - \beta \eta^{-1} \sin^2 \bar{\tau}_r).$$

The radial component of energy flux is then

$$\vec{S} \cdot \hat{y} = \text{const} \times \eta^{-4} [\cos \bar{\tau}_r - \beta \eta^{-1} \sin^2 \bar{\tau}_r]^2. \quad (45.13)$$

Equivalently, this formula gives the outward power per solid angle received at a fixed location in the orbital plane. To use Equation 45.13, first compute $\beta = b\omega/c$. Use Equation 45.8 to translate from lab time to $\bar{\tau}_r$ and substitute into $\eta = (1 - \beta \cos \bar{\tau}_r)$ and Equation 45.13.

Figure 45.2 shows 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 45.7. 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 in the relativistic case, energy emission is also tightly peaked in the polar angle (out of the plane of the orbit).

²Beware that many authors instead compute a related but different quantity, the energy loss from the particle per *emission* time t_r .

FURTHER READING

Semipopular:

x

Intermediate:

Davidson, 2019, chap. 17; Zangwill, 2013, §23.3.3; Garg, 2012, §170.5; Heald & Marion, 2012, §8.4; Vanderlinde, 2004; Jackson, 1999, §14.1–3; Schwinger et al., 1998, chap. 38–40.

PROBLEMS

45.1 *Moderately-relativistic case*

Equation 45.7 gave an approximate formula for retarded proper time, valid for observer time very close a radiation pulse. For ultrarelativistic particles, the pulse is very short and this approximation was adequate. Write a short computer code that finds retarded proper time numerically, substitute into Equation 45.13, and plot the time course of field strength squared for $b\omega/c = 0.8$.

45.2 *Polar beaming*

[Not ready yet.]

CHAPTER 46

The Microwave Polarizer

46.1 FRAMING: JONES TENSOR

Media 1 shows 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 explain 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

The wire spacing was smaller than the wavelength, so let's model the microwave polarizer as a thin, planar conducting sheet at $z = 0$. It's highly anisotropic, conducting 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:

$$\vec{j}^{[2D]} = \vec{\kappa}_s \cdot \vec{E}, \quad \text{where} \quad \vec{\kappa}_s = \kappa_s \hat{x} \otimes \hat{x}. \quad (46.1)$$

We approximate the incoming fields far from the source as a plane wave traveling along \hat{z} , and begin by supposing that it is linearly polarized along the conducting direction:

$$\vec{E} = \frac{1}{2} \bar{E} \hat{x} e^{-i(\omega t - kx)} + \text{c.c.}, \quad \text{where} \quad k = \omega/c.$$

We will also simplify by considering a poor conductor, that is, κ_s is small. Then each surface element will have little influence on the others; each just responds to the incoming plane wave \vec{E}_{in} via our ohmic hypothesis. Each surface element responds in phase with the others. Each in turn radiates according to the Green function solution. For example, at a point along the $+\hat{z}$ axis we have a total radiation field from all surface elements given by

$$\vec{A}_{\text{rad}} = \hat{x} \frac{\mu_0}{4\pi} \int d^2 r_* \frac{1}{R} \kappa_s \frac{1}{2} \bar{E} e^{-i\omega(t-R/c)} + \text{c.c.}$$

The integral runs over the whole plane $z = 0$. 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 φ_* just gives 2π and we are left with $r_* dr_*$. The integrand, $R^{-1}e^{i\omega R/c}$, is a messy function of r_* , but there is an amazing trick. At an observation point along the $+\hat{z}$ axis ($z > 0$), we have $R^2 = r_*^2 + z^2$, so $RdR = r_* dr_*$. Thus, we can change variables in the integral to get

$$\int_0^\infty d^2 r_* R^{-1} e^{ikR} = 2\pi \int_z^\infty dR e^{ikR}.$$

That integral is easy! But it's confusing:

$$= \frac{c}{i\omega} \left[e^{i\infty} - e^{ikz} \right].$$

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 consider larger L . But let's compute the magnetic field, a physical quantity:

$$\vec{\nabla} \times \vec{A} = \frac{\mu_0}{4\pi} \frac{2\pi\kappa_s \vec{E}}{2ik} e^{-i\omega t} (ik)(-\hat{x}) \times \hat{z} \left(-e^{ikz} + \frac{z}{\sqrt{L^2 + z^2}} e^{ik\sqrt{L^2 + z^2}} \right) + \text{c.c.}$$

Taking $L \rightarrow \infty$ at fixed z , we see the second term may be dropped:

$$\vec{B} = -\hat{y} \frac{\kappa_s \vec{E} \mu_0}{4} e^{-i(\omega t - kz)} + \text{c.c.}$$

Your Turn 46A

Compute the electric field as usual, obtaining

$$\vec{E}_{\text{rad}} = -\hat{x} \frac{\kappa_s \mu_0 \vec{E} c}{4} e^{-i\omega(t-z/c)} + \text{c.c.}$$

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

$$\vec{E}_{\text{tot}} = \hat{x} \frac{1}{2} \vec{E} \left(1 - \frac{1}{2} \kappa_s \mu_0 c \right) e^{-i\omega(t-z/c)} + \text{c.c.} \quad (46.2)$$

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 $\approx (1 - \kappa_s \mu_0 c)$ (remember that we work only to lowest order in κ_s). You should work out the radiated wave \vec{E}_{rad} in the backward (reflected) direction, along $-\hat{z}$, but clearly its energy flux will be proportional to $(\kappa_s)^2$, and so cannot fully account for the effect that we found. Instead, we must look for the culprit elsewhere.

A conductor with finite conductance *dissipates* energy as heat. The total loss is

$$\int d^2 r_* \vec{E} \cdot \vec{j}^{[2D]}, \quad (46.3)$$

where the 2D charge flux on the surface, $\vec{j}^{[2D]}$, is given by Equation 46.1. The loss per unit area is just the integrand of Equation 46.3.

Your Turn 46B

Add it to the energy flux from Equation 46.2 and compare to the incoming energy flux.

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 κ_s) and \hat{y} (eigenvalue 0). Equation 46.2 says that an incoming wave 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 *unaffected*—as seen in Media 1.

We can succinctly combine those results by saying that the outgoing complex polarization (Jones vector) $\vec{E}_{\perp}^{\text{out}}$ is linearly related to the incoming by a linear transformation:

$$\vec{E}_{\perp}^{\text{out}} = \vec{J} \cdot \vec{E}_{\perp}^{\text{in}},$$

where \vec{J} is called the **Jones tensor**; it acts on the complex 2D space of transverse directions:

$$\vec{J}_{ij} = \begin{bmatrix} 1 - \frac{1}{2}\kappa_s\mu_0c & \\ & 1 \end{bmatrix}_{ij}.$$

46.4 A TILTED POLARIZER CAN REGENERATE A MISSING POLARIZATION

Finally, suppose that the incoming wave polarization is linear but tilted by 45 deg: $\vec{E}_{\text{in}} = \bar{E}(\hat{x} + \hat{y})/\sqrt{2}$. Now we find the forward wave to be

$$\vec{E}_{\text{tot}} = \frac{1}{2}\bar{E}\left(\frac{\hat{x} + \hat{y}}{\sqrt{2}} - \frac{\kappa_s\mu_0c\hat{x}}{2\sqrt{2}}\right)e^{-i\omega(t-z/c)} + \text{c.c.}$$

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

Aligned, conducting polymer chains can create an optical polarizer.

[[E. Land experimented with polyvinyl alcohol chains aligned on plastic substrate. When the material is heated or stretched, the chains become electrically conducting, creating large polarizability in one direction. This comes along with dissipation (Chapter 46), and so Land's “polaroid filter” effectively blocked EM radiation with one linear polarization, much like the microwave polarizer.¹ Polaroid filters are not the only way to obtain polarized light, but they were much cheaper and more convenient than the alternatives available at that time.]]

¹See Media 1.

PROBLEMS

46.1 *Another integral*

Another situation of interest involves a plane wave that impinges on a dielectric (nonconducting but polarizable) sheet. We then need an integral of the form $\int_0^\infty (2\pi\rho d\rho) (1 - \cos^2 \alpha) e^{ikr} / (4\pi r)$. Here $r = \sqrt{\rho^2 + (z_*)^2}$ and $\cos \alpha = \rho/r$. k , z_* are constants.

Following the discussion in the *Feynman Lectures*,² we can wave our hands a bit and argue that this integral is approximately equal to $\frac{i}{2k} e^{ikz_*}$. You may or may not find this argument convincing, but either way, it's good to check. Unfortunately this integral is probably not one 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_* and k . There is only one dimensionless combination of these parameters; call it $M = kz_*$.

- a. Change variables in the integral from ρ to r . Define dimensionless variable $u = kr$, and express the thing that is to be shown in terms of it. Express it in the form (a certain integral) ≈ 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?

²Volume 1, sections (30-7)–(31-2).

CHAPTER 47

Scattering by Free and Bound Charges

47.1 FRAMING: RERADIATION

[Not ready yet.]

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.¹

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

$$\vec{E}(t, \vec{r}) = \frac{1}{2} \vec{E} e^{-i(\omega t - \vec{k} \cdot \vec{r})} + \text{c.c.}$$

(and the associated \vec{B} field). The charge feels an electric force $\vec{f} = q\vec{E}$. The magnetic force is negligible because $E = cB$ so $q\vec{v} \times \vec{B} \sim q(v/c)E$. Our assumption of nonrelativistic motion, $v/c \ll 1$, means that we can neglect this part of the force.²

Write the charge's resulting trajectory as $\vec{\Gamma}(t) = \frac{1}{2} \vec{\Gamma} e^{-i\omega t} + \text{c.c.}$ Then Newton's law gives the amplitude of the shaking motion as $\vec{\Gamma} = -(q\vec{E})/(m\omega^2)$, 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.³

¹Chapter 54 will study plasmas. A situation effectively like this one also holds for some of the electrons in a metal.

²Chapter 20 studied the *longitudinal* force, for which the magnetic part was the leading term and so could not be dropped.

³But not in the free electron laser.

47.3 SCATTERING CROSS SECTION AND THE THOMSON FORMULA

Our shaking charge gives rise to a time-dependent dipole moment $\vec{\mathcal{D}}_E(t) = q\vec{\Gamma}(t)$, so it will radiate at the same frequency. The charge's motion remains confined to a region of size $\|\vec{\Gamma}\|$. The criterion for the electric-dipole approximation⁴ is met by virtue of Equation 47.1:

$$\|\vec{\Gamma}\| \omega/c = (q\vec{E})/(m\omega^2 c) \omega \ll 1.$$

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

$$\vec{S} = \hat{r} \frac{\mu_0}{(4\pi r)^2} \frac{1}{c} \frac{d^2}{dt^2} \mathcal{D}_E^2 \sin^2 \vartheta, \quad [43.15, \text{page 569}]$$

where ϑ is the angle between the dipole moment and the direction of observation \hat{r} . In our case, suppose that the incoming wave is polarized along \hat{x} ; then $\vec{\mathcal{D}}_E(t) = \hat{x} \frac{1}{2} \bar{\mathcal{D}}_E e^{-i\omega t} + \text{c.c.}$, with $\bar{\mathcal{D}}_E = -q^2 \vec{E}/(m\omega^2)$. The power output per solid angle is then $d\mathcal{P}/d\Omega = r^2 \hat{r} \cdot \vec{S}$, and its time average is

$$\left\langle \frac{d\mathcal{P}}{d\Omega} \right\rangle = \frac{1}{(4\pi)^2} \frac{1}{\epsilon_0 c^3} \frac{q^4}{m^2} \frac{1}{2} \|\vec{E}\|^2 \sin^2 \vartheta. \quad (47.2)$$

Remarkably, the angular frequency ω 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 $\vec{\mathcal{D}}_E$. 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 tell 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, absorbed, whatever) 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 $1/(1/\mathbb{L}^2)$, 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:

$$\langle \|\vec{S}_{\text{in}}\| \rangle = \langle \mu_0^{-1} \|\vec{E} \times \vec{B}\| \rangle = \frac{1}{2} \epsilon_0 c \|\vec{E}\|^2.$$

⁴Section 43.2.3 (page 563).

The cross-section is traditionally denoted σ . Extending our original thought experiment, we can subdivide this scattering cross section into bits attributable to scattering into particular angular bins $d\Omega$, or:

$$\frac{d\sigma}{d\Omega} = \left\langle \frac{d\mathcal{P}}{d\Omega} \right\rangle / \left\langle \|\vec{S}_{\text{in}}\| \right\rangle.$$

This quantity is generically called the **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

$$\frac{d\sigma}{d\Omega} = \left(\frac{1}{4\pi\epsilon_0 c^2} \frac{q^2}{m} \right)^2 \sin^2 \vartheta. \quad \text{Thomson scattering cross-section} \quad (47.3)$$

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 **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

$$\int d\varphi d(\cos \vartheta) \sin^2 \vartheta = 8\pi/3.$$

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.

47.4 LIGHT PROPAGATES DIFFUSIVELY IN A STELLAR INTERIOR

Light moves fast, but it still takes a long time to escape from the interior of a star.

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 $1/(r_c^2 c_e)$, where c_e is the volume density of free electrons. The mean free path for light is this quantity times some geometrical constants of order one.

47.5 POLARIZED INCOMING LIGHT RETAINS ITS POLARIZATION UPON SCATTERING

Suppose that the incoming light travels along \hat{z} , with polarization along \hat{x} . Then $\vec{\mathcal{D}}_{\text{E}} \parallel \hat{x}$. 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 327) argued that we can treat unpolarized light as an incoherent superposition of many pure waves. Scattering can *create* polarization from such light. For example, when viewed at 90 deg 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 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 yet.]

Polarization of CMBR gives a diagnostic for early Universe inhomogeneity.

47.7 THE CASE OF BOUND CHARGES

47.7.1 Rayleigh scattering cross section

Next suppose that the charge is bound, for example, to a heavy atomic nucleus. The simplest classical model we can make of that situation is to suppose that the charge gets a linear restoring force with some spring constant k . As usual with harmonic oscillators, it is convenient to introduce $\omega_0 = \sqrt{k/m}$. Then Newton's law becomes

$$-m\omega^2 \vec{\Gamma} = -\omega_0^2 \vec{\Gamma} + q\vec{E}, \quad \text{so} \quad \vec{\Gamma} = \frac{q\vec{E}}{m(\omega_0^2 - \omega^2)}.$$

Substituting this expression into earlier results then gives the Thomson expression for differential and total cross-sections, each now multiplied by $(1 - (\omega_0/\omega)^2)^2$. Either of

these formulas is called the **Rayleigh cross-section** formula. 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

Direct sunlight is white and unpolarized; however, the sky is bluish and partially polarized.

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 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 sunlight itself, 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

Light scattered from a colloidal suspension is also bluish and partially polarized.

It's easy to send a beam of white light from a projector into a dilute suspension of nonfat milk.⁵ Milk contains dissolved lactose, and so on, but that just gives a solution that's homogeneous on the scale of the wavelength of light (it alters the refractive index), and so again is irrelevant for scattering. Nonfat milk is also a colloidal suspension of protein micelles,⁶ which

- are well separated compared to light wavelength;
- are themselves much smaller than wavelength of light (nanometer scale);
- move randomly and independently; and
- Have polarizability 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.

⁵See Media 16.

⁶Whole milk additionally contains fat globules, which are larger than protein micelles and would complicate the discussion.

T₂

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** $2\pi\hbar/(mc)$, 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 415).

PROBLEMS

47.1 *Estimates and approximations*

A red laser gives a 100 mW beam that is approximately a plane wave with cross-sectional area 1 mm^2 .

- Find the electric field strength in this beam.
- Estimate the fractional deformation of a hydrogen atom placed in this beam, due to the electric field. Is it likely that we could make the approximation of working to first order in this deformation when we study polarizability?
- Suppose that this beam encounters a single free electron. The electron responds by oscillating. Justify our use of the nonrelativistic approximation for that motion.

47.2 *Diffusion of light*

Idealize the Sun as a highly ionized plasma with average free electron density about 10^{24} cm^{-3} .

- Use the Thomson formula to find the mean free path for electromagnetic radiation in the Sun, as a function of wavelength.
- 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 $7 \cdot 10^8 \text{ m}$.

CHAPTER 49

Light in Isotropic, Linear 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*

49.1 FRAMING: *CROSS-SUSCEPTIBILITY*

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, the 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. We also suppose the latter to be finely enough divided (compared to the length scales of the disturbances under study) that they can 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: A solution of randomly oriented molecules is completely isotropic, yet nevertheless can rotate polarized light.

Physical idea: Molecular chirality affects light by allowing *cross-susceptibility*.

49.2 POLARIZABLE MEDIA

49.2.1 Induced electric dipole moment can be merged with \vec{E} , yielding the displacement field

First we review the discussion of dielectric materials from Chapter 6. Figure 49.1 recalls the argument for why a bound charge density arises with

$$\rho_{q,b} = -\vec{\nabla} \cdot \vec{P}. \quad (49.1)$$

¹You'll add this complication in Problem 49.3.

²Or equivalently, 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} , \vec{M} below are all averages of this sort.

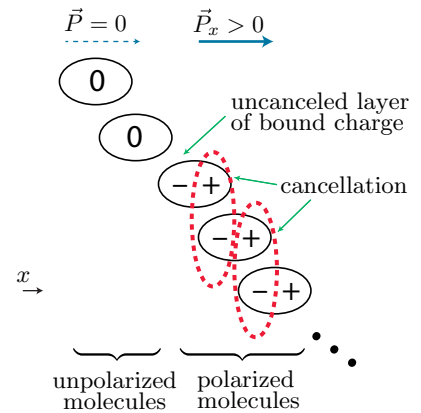


Figure 49.1: [Duplicate of Figure 6.3.] **Origin of bound charge density.** A collection of electrically polarizable “molecules” with nonuniform electric dipole moment density \vec{P} (magnitude increasing from left to right). Net bound charge appears that is minus the divergence of the polarization density, which in this case is $-\partial\vec{P}_x/\partial x < 0$.

If the polarization is time-dependent, then the localized motions of bound charges will also give rise to a **bound charge flux** $\vec{j}_{b,P}$ via the continuity equation: $\partial\rho_{q,b}/\partial t = -\vec{\nabla} \cdot \vec{j}_{b,P}$. Substituting that result into Equation 49.1 gives $0 = \vec{\nabla} \cdot (\vec{j}_{b,P} - \partial\vec{P}/\partial t)$, which will be satisfied if

$$\vec{j}_{b,P} = \partial\vec{P}/\partial t. \quad \text{electric dipole contribution to bound charge flux} \quad (49.2)$$

To understand this result, 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.

The **electric displacement** is defined by Equation 6.7:

$$\vec{D} = \epsilon_0\vec{E} + \vec{P}. \quad (49.3)$$

(We’ll just call it “the \vec{D} field.”) With this definition, the electric Gauss law will take a simple form (see Equation 49.7 below). The only source appearing explicitly in that formula is the free charge density.

[T2] Section 49.2.1’ (page 619) introduces dissipation.

49.2.2 Induced magnetic dipole moment can be combined with \vec{B} to yield the \vec{H} field

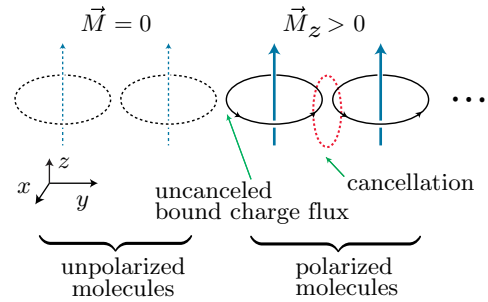
Let \vec{M} denote the net magnetic dipole moment density created by the motions of bound charges in individual polarizable objects. If \vec{M} is spatially nonuniform, it will give rise to a second contribution to the bound charge flux (in addition to Equation 49.2). Figure 49.2 shows a simple example of this effect. The general formula

$$\vec{j}_{b,M} = \vec{\nabla} \times \vec{M} \quad \text{magnetic dipole contribution to bound charge flux} \quad (49.4)$$

is rotationally invariant and agrees with the figure in the special case shown there. Deleting the unpolarized “molecules” at the left of the figure also shows that at the boundary between medium and vacuum, we get a **bound surface charge flux**

$$\vec{j}_b^{[2D]} = \vec{M} \times \hat{n}, \quad (49.5)$$

Figure 49.2: One source of bound charge flux. A collection of magnetically polarizable “molecules” in a nonuniform magnetic field (magnitude increasing as we move to the right). *Black rings* indicate classical currents equivalent to those induced by the applied field. Net bound charge flux appears that is proportional to the curl of magnetic moment density, in this case equal to $\hat{x}(\partial\vec{M}_z/\partial y)$, which is directed out of the page.



where \hat{n} is the outward-directed perpendicular to the interface.

The **magnetic field intensity** is then defined by

$$\vec{H} = \mu_0^{-1}\vec{B} - \vec{M}. \quad (49.6)$$

(We’ll just call it “the \vec{H} field.”) With this definition, the Ampère law will take a simple form (see Equation 49.8 below). The only source appearing explicitly in that formula is the free charge flux.

49.2.3 The Maxwell equations look simple in terms of the new fields

We wish to eliminate explicit mention of the bound charges and currents, a job begun in Chapter 6. The remaining (not bound) charges and currents are called “free”: $\rho_{q,f}$, \vec{j}_f . Excess static charges, which macroscopically violate charge neutrality, are considered free, for example, the charge placed on a capacitor. Currents that transport net charge over macroscopic lengths are also considered free, for example, those in a coil of wire surrounding an inductor.

Your Turn 49A

Using Equations 49.1, 49.2, and 49.4, show that

$$\vec{\nabla} \cdot \vec{D} = \rho_{q,f} \quad \text{Gauss} \quad (49.7)$$

$$\vec{\nabla} \times \vec{H} - \frac{\partial \vec{D}}{\partial t} = \vec{j}_f. \quad \text{Ampère} \quad (49.8)$$

Equation 49.7 extends the validity of Equation 6.8 (page 78) 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

We have already seen that the perpendicular component of the \vec{B} field must be continuous across a boundary between media:

$$\Delta\vec{B}_\perp = 0. \quad \text{always} \quad [15.23, \text{page 224}]$$

We also saw that at a dielectric/vacuum interface, with no free surface charge or current,

$$\hat{n} \cdot (\vec{E}^{[\text{vac}]} - \vec{E}^{[1]}) = \sigma_b / \epsilon_0, \quad [6.19, \text{page } 87]$$

$$\Delta \vec{E}_{\parallel} = 0, \quad \text{and} \quad [6.21, \text{page } 87]$$

$$\Delta \vec{B}_{\parallel} = \mu_0 \vec{j}_b^{[2D]} \times \hat{n}, \quad [15.24, \text{page } 224]$$

where \hat{n} points outward from medium 1 (toward the vacuum). At an interface between two magnetic media, or one such medium and vacuum, the contribution from bound currents can be incorporated into \vec{H} .

Your Turn 49B

Allow for free surface charge density and flux. Use Equations 6.19, 15.24, 6.4, and 49.5 to show that

$$\Delta \vec{D}_{\perp} = \sigma_f; \quad \Delta \vec{H}_{\parallel} = \vec{j}_f^{[2D]} \times \hat{n}.$$

Here $\Delta \vec{D}_{\perp} = (\vec{D}^{[2]} - \vec{D}^{[1]}) \cdot \hat{n}$, where \hat{n} is the unit perpendicular vector pointing from medium 1 to medium 2; similarly for $\Delta \vec{H}_{\parallel}$.

These results are particularly useful when we have reason to believe that an interface has zero free surface charge density and zero free surface current. The other boundary conditions are the same as always:³

$$\Delta(\vec{B}_{\perp}) = 0 \quad \text{and} \quad \Delta(\vec{E}_{\parallel}) = 0.$$

Section 49.2 has reviewed the useful separation of charge density into free and bound components, and then extended it to cover charge flux as well.

[T2] Section 49.2' (page 619) mentions more sophisticated ways to think about bound charge and current.

49.3 LINEAR REGIME

Our goal was to eliminate explicit mention of bound charges and currents from the Maxwell equations, but Equations 49.7 and 49.8 didn't really succeed: Together with the remaining unmodified Maxwell equations, they 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 eliminating them (Equations 49.3 and 49.6). We now introduce a further level of approximation that, when justified, finishes our job in a simple way.

49.3.1 Induced electric dipole moment effectively modifies ϵ_0

Many dielectric media are approximately linear:⁴ That is, \vec{P} is a linear function of \vec{E} , described by the **dielectric susceptibility**⁵ $\vec{\chi}_e$ via the response function $\vec{P} = \epsilon_0 \vec{\chi}_e \cdot \vec{E}$.

³See Sections 6.10 and 15.7.

⁴**[T2]** For nonlinear media, see Section 49.3' (page 620).

⁵Susceptibility is a tensor because in general a medium's polarizability need not be isotropic (page 193). This chapter restricts to the isotropic case, but Chapter 50 will relax that assumption.

The dielectric susceptibility describes how much induced electric dipole moment density you get (deformation times charge per volume) per applied electric field (force per charge). That is, $\tilde{\chi}_e$ is essentially a spring constant tensor, multiplied by density and charge squared. Like any spring constant tensor, it is symmetric.

For simplicity, this chapter assumes that the medium is **isotropic** (χ_e is a scalar constant⁶). 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**⁷ $\epsilon = \epsilon_0(1 + \chi_e)$. Then

$$\vec{D} = \epsilon \vec{E}. \quad \text{constitutive relation for uniform, linear, isotropic, lossless, nonchiral dielectric} \quad [6.9, \text{page 78}]$$

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 ϵ), anisotropy ($\vec{\epsilon}$ with tensor structure), and chirality.⁸

49.3.2 Induced magnetic dipole moment effectively modifies μ_0

Many magnetic media are also approximately linear; that is, \vec{M} is a linear function of \vec{B} . For isotropic media, it can be described by another material property, the **magnetic susceptibility**⁹ $\tilde{\chi}_m$, via the response function $\vec{M} = \mu_0^{-1} \tilde{\chi}_m \vec{B}$. Define the **permeability** $\mu = \mu_0 / (1 - \tilde{\chi}_m)$. Then

$$\vec{H} = \mu^{-1} \vec{B}. \quad \text{constitutive relation for uniform, linear, isotropic, lossless nonchiral magnetic material} \quad (49.9)$$

More general forms of the constitutive relation include dissipation (complex μ), anisotropy ($\vec{\mu}$ with tensor structure), and chirality.

49.3.3 The Maxwell equations then only involve free charge and charge flux

Equations 49.7 and 49.8 are general. For the special case of linear media, they can be combined with Equations 6.9 and 49.9, and the boundary conditions, to form a closed system that can be solved to give all fields in terms of free charges and currents.

⁶ [T2] Problem 14.2 (page 212) showed that a rotationally invariant rank-2 tensor must be a constant times the identity.

⁷ Section 6.5.1 (page 76).

⁸ Chapter 50 studies anisotropy. Section 49.6 studies chirality. [T2] Chapter 54 discusses the meaning of complex response functions.

⁹ Again the assumption of isotropy implies that the susceptibility is a 3-scalar. Equation 49.9 follows a convention in Feynman et al., 2010a. Some books instead define a different quantity χ_m by $\vec{M} = \chi_m \vec{H}$. The two descriptions are equivalent: The relation between the susceptibilities is $\tilde{\chi}_m = \chi_m / (1 + \chi_m)$.

That is, we can forget about the medium if it's linear; the 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 49B that the same is true for the boundary conditions.

In particular, in a bulk isotropic medium, there will be the same wave solutions as in vacuum (two transverse polarizations), except that the velocity is $(\epsilon\mu)^{-1/2}$ instead of c . For example, dielectric polarizability ($\epsilon > \epsilon_0$) leads to a slowdown, that is, to a value of the refraction index that is larger¹⁰ 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.
- A jumble of circular rings of conductor, oriented randomly. This medium is magnetically polarizable and isotropic.

[Not ready yet.]

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. Although 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.

Note that \vec{P} and \vec{M} may arise due to processes that are not instantaneous. Nevertheless, linearity and time-translation invariance of the Maxwell equations imply the existence of single-frequency solutions. But ϵ and μ will in general be frequency dependent, leading to **dispersion**, that is, the dependence of wave velocity on frequency (Figure 49.3). Dispersion, in turn, implies that refraction of light will be wavelength dependent. Thus, a glass lens will have slightly different focal lengths for each wavelength, making it unable to simultaneously focus them all (**chromatic aberration**).

Here are some examples:¹¹

- $\epsilon \approx 81\epsilon_0$ for water at $\omega \rightarrow 0$; it's highly polarizable. But $\epsilon \approx (4/3)^2\epsilon_0$ for water at visible frequencies; the alignment of permanent dipoles is sluggish.
- For a jumble of *split* rings, each ring can act as an RC circuit and will display resonance.¹²

¹⁰Exotic “metamaterials” exist with ϵ that is not positive in certain frequency ranges, requiring special interpretation (Chapter 55). *Anisotropic* polarizability, for example in a crystalline material, leads to birefringence (Chapter 50).

¹¹We also saw examples of dispersion in Thomson and Rayleigh scattering (Chapter 47).

¹²[\[T2\]](#) See Zangwill, 2013, §18.5.6.

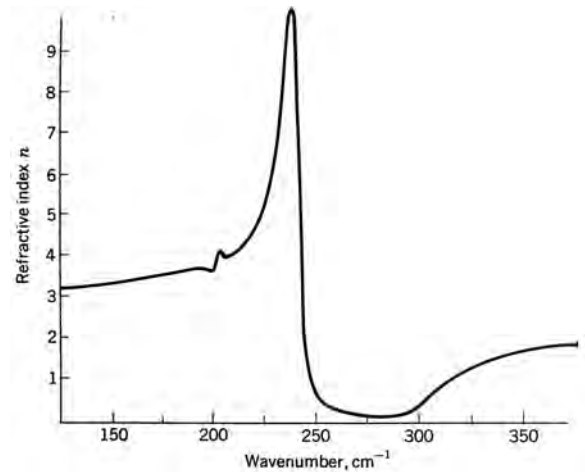


Figure 49.3: Index of refraction of a solid medium (cadmium sulfide) as a function of wavelength, showing complex behavior near a resonance.

- Section 54.3.2 will discuss dispersion in a cold plasma: $\epsilon = \epsilon_0(1 - (\frac{\omega_P}{\omega})^2)$.

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.

[T2] Section 49.3' (page 620) will discuss more general response functions.

49.4 "TOTAL" INTERNAL REFLECTION AND THE EVANESCENT WAVE

"Total" internal reflection is not quite total.

Our discussion has justified the approach to optics used in Chapter 21 and has extended it to magnetically responsive media.

[Not ready yet.]

49.5 CIRCULAR BIREFRINGENCE SEEMS TO PRESENT A PARADOX

Section 49.3.3 argued that waves propagate in isotropic, linear, uniform media in much the same way as in vacuum: Changing the value of ϵ , μ , or both slows the waves down, but cannot alter polarization.

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,¹³ effectively creating an isotropic medium. So we again predict no effect on the polarization of light.

Circular birefringence, also called optical activity, in an isotropic solution.

The prediction just made *fails* spectacularly, however, even for everyday material

¹³ **[T2]** See Problem 14.2.

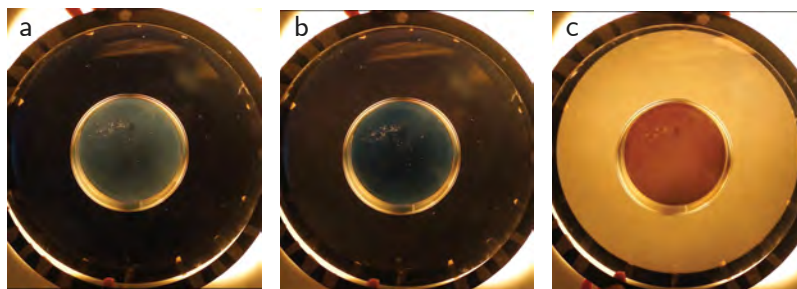


Figure 49.4: [Photos.] **Circular birefringence.** A dish containing a layer of corn syrup 0.5 cm deep was placed between two polarizing filters and illuminated from below with white light. (a) When the two polarizers are oriented at 90 deg, the region outside the dish is darkest, but light passing through the syrup is partly transmitted. (b) When the angle between polarizers is increased to about 92 deg, 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 deg, blue is maximally blocked, giving a reddish tint to the light transmitted through syrup. [Photos courtesy Le-Qi Tang.]

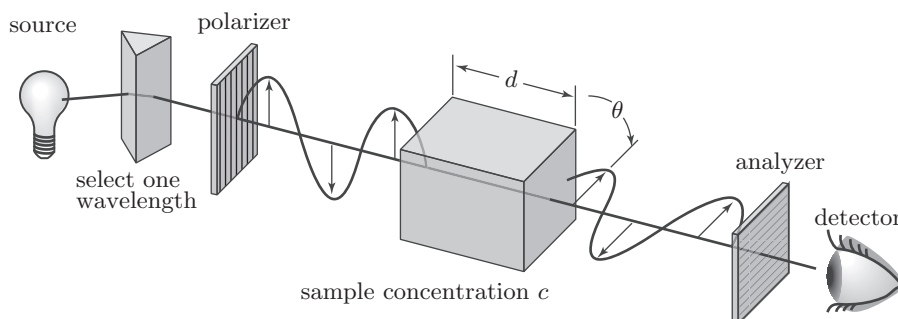


Figure 49.5: [Schematic.] **Measuring circular birefringence with a polarimeter.** The *arrows* represent the electric field vector in a beam of light. They are shown rotating by an angle θ as the light passes through the sample; the rotation shown corresponds to the positive value $\theta = +\pi/2$. 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 optical rotation changes sign.

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 counterclockwise direction when viewed along \vec{k} (Figure 49.4). This Electromagnetic Phenomenon is called **circular birefringence**.¹⁴

What property could determine this direction of rotation, a choice that breaks spatial inversion invariance (Figure 49.5)? Because the Maxwell equations are themselves invariant under inversions,¹⁵ the only source of optical rotation must be a property

¹⁴Some books use the synonym **optical activity**; the medium is sometimes said to possess **optical rotatory power**. Circular birefringence is different from ordinary birefringence, which can happen even in a nonchiral crystal of nonchiral objects (Chapter 50).

¹⁵Problems 15.4 and 18.1.

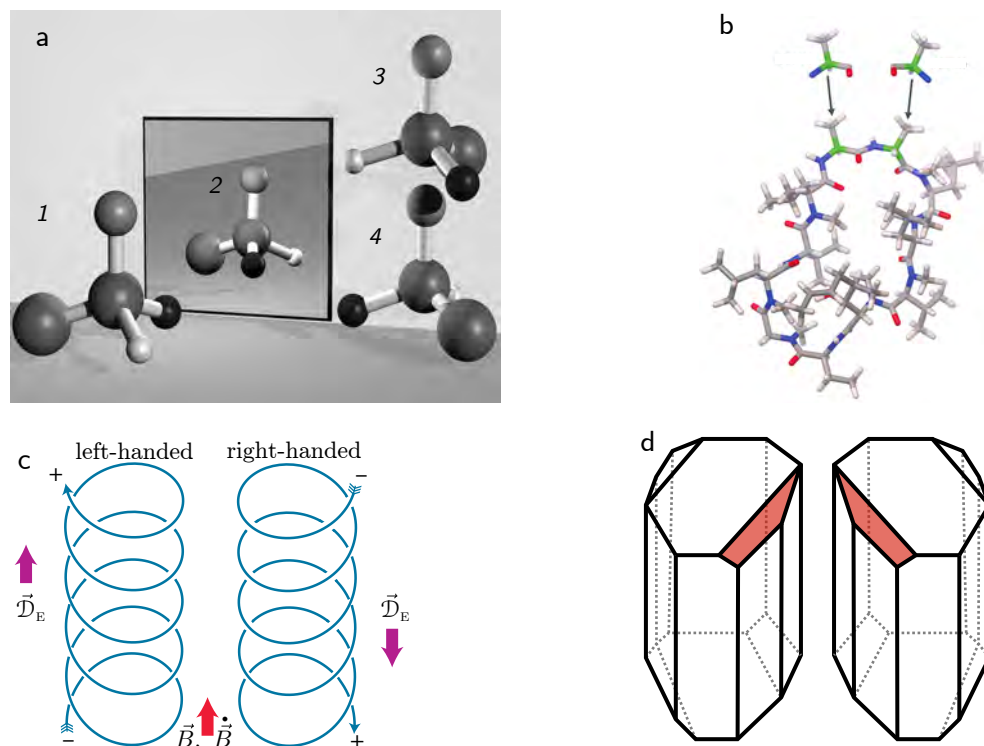


Figure 49.6: Chirality. (a) A simple chiral molecule can be obtained by bonding four different atoms to a central carbon. The molecule *1* cannot be brought into coincidence with its mirror image (shown as *2*) by any rotation; *3,4* show some failed attempts. (b) Chirality of amino acids. Cyclosporin, a cyclic peptide made by fungi, contains a pair of alanines 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 crystals of tartaric acid that cannot be rotated into each other's mirror image. [(a) Art by Sarina Bromberg; (b) from Goodsell, 2016.]

of the sugar molecules themselves—one not shared by, say, water molecules.

Indeed, glucose differs from H_2O by a property called **chirality**. An object that cannot be superimposed on its mirror image by any rotation or translation is called **chiral** (Figure 49.6a).¹⁶ 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 sky¹⁷ 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 solution of any kind of molecule—chiral or not.

We must be missing something crucial. Since we calculated that an effect is zero

¹⁶Objects that are not chiral are called “nonchiral” or “achiral.” The two mirror images of a chiral object are called each other's **enantiomers**.

¹⁷Section 47.7.2 (page 598).

and observed that it's not, maybe we made a bad approximation. A typical impulse is to wonder: Maybe we truncated a power series to an order at which the effect does not yet arise. But that's not the answer.

49.6 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:

$$\begin{bmatrix} \vec{P} \\ \vec{M} \end{bmatrix} = \begin{bmatrix} \epsilon_0 \chi_e \mathbf{\overset{\leftrightarrow}{\mathbb{I}}} & ? \\ ? & \frac{1}{\mu_0} \tilde{\chi}_m \mathbf{\overset{\leftrightarrow}{\mathbb{I}}} \end{bmatrix} \begin{bmatrix} \vec{E} \\ \vec{B} \end{bmatrix}. \quad (49.10)$$

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 6.9 and 49.9) 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).

49.6.1 Macroscopic physical realizations give intuition about the reality of a new effect

Are the cross-susceptibilities imagined in Equation 49.10 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 49.6c). 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 time-dependent but spatially uniform \vec{E} field directed along the helical axis direction, with magnitude $\|\vec{E}\|$ increasing in time, so that $\partial\vec{E}/\partial t$ is parallel to \vec{E} . This applied field leads to an electric dipole moment \vec{D}_E as usual. Because it's time-dependent, there must be a net current in the wire. The helical structure then forces the charge flux to have an azimuthal component, so it generates a magnetic dipole moment: $\vec{D}_M = (c\eta')(\partial\vec{E}/\partial t)$, where η' is a constant.

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 η' , depend on the handedness of the helix:

Your Turn 49C

Show that η' is positive for a right-handed helix and negative for a left-handed one.

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 49.6c). 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.¹⁸ But the helical shape also imposes an *axial* motion of charge, and hence an *electric* dipole moment:

Your Turn 49D

- Show that $\vec{\mathcal{D}}_{\text{E}} = -c\eta(\partial\vec{B}/\partial t)$, where the constant η is positive for the right-handed helix, or negative for the left-handed one.
- Check that the units of η and η' match.
- 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. Even if the medium contains randomly oriented helices, some fraction of them will have their axes along \vec{E} or \vec{B} .

49.6.2 The general constitutive relation has new, frequency-dependent terms

The preceding discussion suggested that in general, a uniform, linear, isotropic, lossless, medium will have¹⁹

$$\begin{bmatrix} \vec{P} \\ \vec{M} \end{bmatrix} = \begin{bmatrix} \epsilon_0\chi_e & -\eta\partial/\partial t \\ \eta'\partial/\partial t & (\mu_0c^2)^{-1}\tilde{\chi}_m \end{bmatrix} \begin{bmatrix} \vec{E} \\ \vec{B} \end{bmatrix}, \quad (49.11)$$

where η and η' 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.)

Generally, the susceptibilities χ_e , η , η' , and $\tilde{\chi}_m$ are 3-tensors, but in an isotropic medium such as aqueous solution, they get replaced by their averages over orientation, that is, by 3-scalars times the identity tensor.²⁰

As mentioned before, χ_e and $\tilde{\chi}_m$ may be frequency dependent. Similarly, for disturbances at a specific angular frequency ω , the cross-terms will also be functions of frequency, due both to possible frequency dependence of η and η' and to the explicit time derivatives in the formula. By time-reversal invariance, they must be odd functions of frequency.

We conclude that the generalized response function proposed in Equation 49.10 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 49.11 odd in frequency was needed to join quantities with different time-reversal characters.

Previously you showed that η and η' always have the same sign in a macroscopic physical realization of cross-polarization.²¹ Replacing the helices by their mirror images

¹⁸Lenz's law; see Section 18.3.1.

¹⁹There is a slight change of notation here: Now η, η' include the density of the polarizable molecules.

²⁰Equation 49.11 abbreviates by omitting the factors of $\vec{1}$.

²¹Your Turns 49C and 49Da. $\boxed{\mathcal{T}2}$ 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 η and η' . If molecular chirality were 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 this prediction is confirmed. Let's now make the connection explicit.

49.7 THE ORIGIN OF CIRCULAR BIREFRINGENCE

Your Turn 49E

- Formulate a plane-wave trial solution for the medium described by Equation 49.11.
- To keep things simpler, you may (unrealistically) set $\chi_e = \tilde{\chi}_m = 0$, that is, neglect the ordinary susceptibilities and focus only on the cross-susceptibilities. Show that the condition for a plane-wave solution simplifies if we expand the polarization vector in the circular polarization basis (helicity basis) $\vec{\zeta}_{(\pm)} = (\hat{x} \pm i\hat{y})/\sqrt{2}$ (Equation 18.32, page 272).
- Show that each circular polarization propagates with a different phase velocity if the medium is chiral.

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 49E). 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 49F

- 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.
- Show that the angle of rotation is proportional both to $n_+ - n_-$ and to the thickness of the slab.
- 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 49.5, page 611). 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 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 49.6a).
- Alternatively, nonchiral molecules may be arranged in a chiral crystal structure (such as in quartz).
- However, air (O_2, N_2), liquid water (H_2O), and so on don't display this phenomenon—they are all disordered arrangements of nonchiral (inversion-invariant) objects.
- The time derivatives in Equation 49.11 predict that the effect will be strongly dependent on frequency.

49.8 HOW TO OBSERVE CIRCULAR BIREFRINGENCE

To follow up on those questions, Figure 49.4 shows²² a dish with a layer of corn syrup (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.

The figure also shows that blue (higher frequency) light rotates more than red. Had we diluted the solution by adding more H_2O to it, the total optical thickness would have gone up, but the total projected density of sugar molecules/area would not; empirically, one indeed finds that the total polarization rotation doesn't change.

49.9 MORE 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

²²See also Media 17.

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 49.6c). 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.²³ 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).²⁴ 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 murder mysteries.
- One could now try to formulate a quantum mechanical calculation that leads from molecular structure to a prediction of the value of η . 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 η characterizing the effect of chirality of an isotropic medium on light (to leading nontrivial order in frequency).
- The math predicted that circular birefringence goes to zero at zero frequency, due to the time derivative in Equation 49.11 (page 614). More generally, the entire spectrum 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.²⁵
- There can also be chiral dissipation (“friction”) terms, leading to different absorption lengths for each helicity (each choice of polarization vector $\hat{\zeta}_{(\pm)}$). The entire spectrum of the differential absorption is called the material’s **circular dichroism** spectrum, yet another fingerprint of a molecule that can be observed in solution. An unexpected structural form of the DNA molecule called “Z-DNA” was first discovered via its nonstandard CD spectrum.
- Chapter 54 will show that a similar phenomenon can occur in an astrophysical plasma, if a uniform \vec{B} field is present. Although this medium is very different from sugar solution, nevertheless it breaks spatial inversion symmetry in a way

Optical rotatory dispersion.

Circular dichroism is used for nondestructive molecular structure assays.

²³Pasteur’s breakthrough is all the more remarkable in that the very existence of molecules, let alone their definite shapes, was controversial in 1848.

²⁴Even mass spectrometry cannot separate them, because they have the same charge/mass ratio.

²⁵Although the polarization rotation angle is ambiguous by 180 deg, 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. $\boxed{\mathcal{T}}$ 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.

that is mathematically similar to what we have studied, again leading to different phase velocities for the two circular polarizations.

49.10 PLUS ULTRA

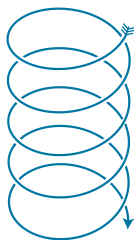


Fig. 49.6c (page 612)

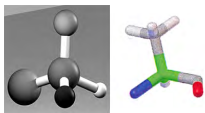


Fig. 49.6ab (page 612)

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 49.9 described some observations.

Although we argued that the imagined structure in Figure 49.6c 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 are the same: They all 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 33.3.4 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.)

Physical realization of a cross-polarizable material: See Feynman et al., 2010b, §33-5; Hecht, 2017, §8.10.1; Nieves & Pal, 1994.

[T2] Relativistic treatment of media: Landau et al., 1984, §76.

Technical:

Experimental observation of double diffraction from optically active liquids: Ghosh et al., 2007.

Liquid crystals: de Vries, 1951.

[T2] Quantum mechanical treatment of circular birefringence of molecules: Craig & Thirunamachandran, 1998, chap. 8.

T₂

49.2' 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.

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 the main text motivated general formulas allowed by principles such as rotational, time reversal, and (when appropriate) spatial inversion invariance.

T₂

49.2.1' 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 we will have

$$\vec{D} = \epsilon(\omega)\vec{E},$$

which defines the frequency-dependent permittivity function. We have tacitly assumed that ϵ is real, but this need not be the case if there is dissipation.²⁶

To understand complex permittivity, imagine a material composed of polarizable “molecules” with density ρ_{mol} , consisting of a pair of charges $\pm q$ that can separate by Δx . Let $\chi_e(\omega) = (\epsilon(\omega)/\epsilon_0) - 1$, as usual. In response to the field, charge will separate by $\Delta x = \frac{1}{2}\Delta x e^{-i\omega t} + \text{c.c.}$

The density of induced dipole moment is then $P = \rho_{\text{mol}}q\Delta x$. That result lets us find the velocity $v(t) = \frac{1}{2}\bar{v}e^{-i\omega t} + \text{c.c.}$, where

$$\bar{v} = \frac{-i\omega\epsilon_0\chi_e(\omega)\bar{E}}{q\rho_{\text{mol}}}.$$

The rate at which the field does work on the particle is qE times v , or

$$qEv = q\left(\frac{1}{2}\bar{E}e^{-i\omega t} + \text{c.c.}\right)\left(\frac{-i\omega\epsilon_0\chi_e\bar{E}}{2\rho_{\text{dip}}}e^{-i\omega t} + \text{c.c.}\right).$$

The time average of that power, per volume, is thus

$$\frac{1}{4}(-i)\omega\epsilon_0\chi_e|\bar{E}|^2 + \frac{1}{4}(i)\omega\epsilon_0\chi_e^*|\bar{E}|^2 = \frac{1}{2}\omega\epsilon_0|\bar{E}|^2 \text{Im } \chi_e.$$

As claimed, if the permittivity function is complex, then the material dissipates energy (into heat). Similar remarks apply for the magnetic permeability.

²⁶See also Section 54.3.1 (page 643).

T₂

49.3' 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 89). Also, any medium will be linear only in some regime of sufficiently weak applied fields. For example, the orientational ordering of water molecules (Section 6.8, page 84) 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 yet.]).

Similarly, ferromagnets have nonzero \vec{M} at zero applied magnetic field (Section 17.5.2, page 249). Also, again any medium is only magnetically linear for sufficiently weak applied fields.

T₂

49.6'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.

49.6'b Relativistic formulation

Equation 49.11 (page 614) 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 $\underline{F}^{\mu\nu}$, as

$$\underline{R}^{\mu\nu} = \begin{bmatrix} 0 & \vec{P}_x & \vec{P}_y & \vec{P}_z \\ -\vec{P}_x & 0 & -\vec{M}_z & \vec{M}_y \\ -\vec{P}_y & \vec{M}_z & 0 & -\vec{M}_x \\ -\vec{P}_z & -\vec{M}_y & \vec{M}_x & 0 \end{bmatrix}^{\mu\nu}, \quad (49.12)$$

where $\vec{M}_i = c^{-1}\vec{M}_i$. This big formula can be summarized in the usual way by $\underline{R}^{0i} = -\underline{R}^{i0} = \vec{P}_i$, and $\underline{R}^{ij} = -\varepsilon_{ijk}\vec{M}_k/c$. Also, let \underline{J}_f denote the free charge flux 4-vector field.

In terms of these definitions, four of the Maxwell equations take the form

$$\partial_\mu \underline{H}^{\nu\mu} = c^{-1} \underline{J}_f^\nu, \quad (49.13)$$

where

$$\underline{H}^{\nu\mu} = c\varepsilon_0 \underline{F}^{\nu\mu} + \underline{R}^{\nu\mu}. \quad (49.14)$$

Thus, $\underline{H}^{0m} = \vec{D}_m$ and $\underline{H}^{nm} = c^{-1}\varepsilon_{nml}\vec{H}_l$, in parallel to the naming of elements of \underline{F} . So \underline{R} must be a 4-tensor, because the world is Lorentz invariant, and Equations 49.13–49.14 are only invariant if \underline{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 \underline{R} is a linear function of \underline{F} :

$$\underline{R}^{\mu\nu} = \underline{K}^{\mu\nu}{}_{\lambda\sigma} \underline{F}^{\lambda\sigma}, \quad (49.15)$$

where the **susceptibility operator** \underline{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 4-tensor. We know that \underline{R} and \underline{F} are 4-tensors, so Equation 49.15 implies that \underline{K} is a 4-tensor operator. 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 \underline{U} . Hence, it must be possible to express \underline{K} as a combination of \underline{U} 's and 4-scalar quantities describing the medium. \underline{K} must also be a symmetric operator in the sense that exchanging $\mu\nu$ with $\lambda\sigma$, and $\underline{\partial} \rightarrow -\underline{\partial}$, must leave it unchanged. Playing around shows that there are only three possible forms permitted by the symmetries:²⁷

$$\begin{aligned} \underline{K}^{\mu\nu}{}_{\lambda\sigma} = & \frac{\alpha}{2} (\delta_{\lambda}^{\mu} \delta_{\sigma}^{\nu} - \delta_{\lambda}^{\nu} \delta_{\sigma}^{\mu}) + \frac{\tau}{2} (\underline{U}^{\mu} \underline{U}_{\sigma} \delta_{\lambda}^{\nu} - \underline{U}^{\nu} \underline{U}_{\sigma} \delta_{\lambda}^{\mu} - \underline{U}^{\mu} \underline{U}_{\lambda} \delta_{\sigma}^{\nu} + \underline{U}^{\nu} \underline{U}_{\lambda} \delta_{\sigma}^{\mu}) \\ & + \frac{\gamma}{2} (\underline{\epsilon}^{\mu\nu}{}_{\tau\lambda} \underline{U}^{\tau} \underline{U}_{\sigma} - \underline{\epsilon}^{\mu\nu}{}_{\tau\sigma} \underline{U}^{\tau} \underline{U}_{\lambda} - \underline{\epsilon}_{\lambda\sigma\tau}{}^{\mu} \underline{U}^{\tau} \underline{U}^{\nu} + \underline{\epsilon}_{\lambda\sigma\tau}{}^{\nu} \underline{U}^{\tau} \underline{U}^{\mu}) \underline{U}^{\rho} \underline{\partial}_{\rho}. \end{aligned} \quad (49.16)$$

Here the components of the 4-dimensional Levi-Civita tensor are $\underline{\epsilon}_{0123} = +1$ and so on.²⁸

Your Turn 49G

Specialize this formula to an inertial coordinate system in which the medium is at rest. Show that the constants α , τ , and γ can be chosen so that Equation 49.16 reproduces Equation 49.11 (which also has three phenomenological parameters χ_e , $\tilde{\chi}_m$, and η).

Then substituting an arbitrary 4-velocity at once tells us the appropriate form of the susceptibility tensor in a *moving* medium.²⁹

Every term in Equation 49.16 must be time-reversal invariant, because a static collection of molecules does not break time-reversal invariance.³⁰ (This is why the γ term needs a derivative.) Also, the α and τ terms are invariant under spatial inversions—but not the γ term. Thus, γ must equal zero for a non-chiral medium, as noted for liquid water in Section 49.8 (page 616).

²⁷More 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 483). Some terms that may seem to be missing from our list are in fact redundant by the Maxwell equations and the constraint that $\underline{U}_{\mu} \underline{U}^{\mu} = -c^2$.

²⁸Section 34.7'a (page 469).

²⁹You previously used similar logic in Problem 34.2 (page 476).

³⁰A more general approach would be needed to include ferromagnetism.

PROBLEMS

49.1 *Electrorotation of living cells*

[Not ready yet.]

49.2

Repeat Your Turn 49E, but this time without the unrealistic simplifying assumptions $\chi_e = \tilde{\chi}_m = 0$.

49.3 *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 a scalar multiplier times $-\vec{\nabla}\psi$, but that coefficient (the conductivity κ) may not be spatially uniform. However, you may assume that the dielectric constant ϵ/ϵ_0 is uniform and isotropic.

- a. Show that the medium will in general acquire a nonzero free electric charge density $\rho_{q,f}(\vec{r})$. Show that this charge density may be written as the dot product of $\vec{\nabla}\psi$ with a certain vector field, and find that vector field.
- b. Repeat for the case where ϵ is also nonuniform, though isotropic.

49.4 *Polarization of evanescent wave*

Polarized total internal reflection microscopy.

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 610). Let’s look into the second point more closely.

A linearly polarized, monochromatic wave of angular frequency ω enters a sample chamber filled with water (refractive index $n_2 \approx 1.33$) from a medium with larger index n_1 (typically quartz, ≈ 1.46). For this problem, you may assume that the permeabilities are equal: $\mu_1 = \mu_2$.

The interface between media is the yz plane. The incoming wave (in the region $x < 0$) has wavevector \vec{k} lying in the xy plane; all fields are independent of z . The incoming \vec{k} makes angle θ with the perpendicular to the interface, that is, $\vec{k} \cdot \hat{x} = \cos\theta$. We’ll eventually consider the case where the angle of incidence θ is large, but you should first work out the answers for arbitrary θ , then specialize to large θ .

It’s convenient to choose the following basis vectors for the incoming polarization:

- “TE” polarization (also called “s-wave”): \vec{E} is parallel to \hat{z} .
- “TM” polarization (also called “p-wave”): \vec{B} is parallel to \hat{z} .

Review Section 49.4 for the definition of the critical angle θ_c , the transmitted wavevec-

tor \vec{k}' , and the reflected wavevector \vec{k}'' . Write the incident wave as

$$\vec{E}(t, x, y, z) = \frac{1}{2} \left[\vec{E} e^{i(\vec{k} \cdot \vec{r} - \omega t)} + \text{c.c.} \right] \quad x < 0.$$

Here \vec{E} is the incoming polarization vector. The transmitted and reflected waves are given by similar expressions with \vec{E}' , \vec{k}' , and so on; they all have the *same* value of ω .³¹ Section 49.4 discussed how to find the transmitted and reflected waves.

The 3-vector \vec{E}' describes the amplitude, phase, and polarization of the transmitted wave. We want to know the polarization, particularly in the case where the transmitted wave is nonpropagating.

Problem:

- Consider a quartz–water interface and incoming light with vacuum wavelength 514 nm. Find the critical angle. Find the exponential amplitude falloff length scale, assuming $\theta = 70$ deg.
- Find the amplitude and direction of the electric field³² for $x > 0$, in the case of TE incident polarization. That is, suppose that $\vec{E}' = \bar{E} \hat{z}$, where \bar{E} is a real constant. Then specialize to the case with $\theta > \theta_c$. Characterize in words the *type* of polarization you get for the evanescent electric field. Then substitute the numbers in (a) to get a quantitative characterization.
- Repeat for the TM polarization. Again characterize in words the *type* of polarization obtained, then substitute the numbers in (a) to get a quantitative characterization.

49.5 T₂ Relativistic formulation

- Use Equations 49.13–49.16 to derive the plane-wave solutions for light in flowing water, relevant to the Fizeau experiment.
- Find solutions corresponding to light propagating in an isotropic, chiral medium (such as sugar water) at rest.

³¹One prime for transmitted, two primes for reflected.

³²We are not interested in any overall phase shift.

CHAPTER 51

Čerenkov Radiation

51.1 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 41. 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: Unlike in vacuum, in this situation, a stationary observer passes suddenly from an early-time regime, with no causal contact to the charge, to a late-time regime that does see it, leading to a sharp transition like the *bow shock* in the water behind a speedboat.

51.2 IN VACUUM, A CHARGED PARTICLE HAS ONE SOURCE POINT IN THE OBSERVER'S PAST LIGHT CONE

To begin to answer, recall the derivation of the fields created by a charge in uniform, straight-line motion in vacuum via from the Liénard–Weichert formula:¹ Reassuringly, we found that there is no radiation (\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. We must now revisit that conclusion in the presence of a transparent dielectric medium, such as water.

51.3 IN A DIELECTRIC MEDIUM, A THERE CAN BE TWO SOURCE POINTS, OR NONE, IN THE OBSERVER'S PAST LIGHT CONE

Section 6.5 argued that in the presence of a medium we may *summarize* the medium's effect simply by modifying the Maxwell equations, replacing ϵ_0 by a larger permittivity ϵ . But now an interesting possibility arises: What if the particle moves faster than the speed of light *in medium*, that is, $\beta c > c/n$ where $n = \sqrt{\epsilon/\epsilon_0}$? It is true that the modified Maxwell equations have a Lorentz-like invariance, with $c_m = c/n$ playing the role of light speed, and we can use that invariance to find the fields if $\beta < c_m$. 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.

¹Section 41.5.2 (page 545)

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 51.2. However, the geometry is different when $v/c_m > 1$. In the language of Figure 41.5b, in this case the stick held fixed on the z axis is *longer* than the pivoting stick. You'll explore the consequences of this difference in Problem 51.1, but the upshot is 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.
- 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 approximation.

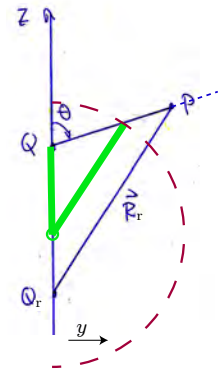


Fig. 41.5b (page 546)

51.4 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 β 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 that we added 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 that molecule a momentary jolt. The sum of the resulting fields from all of the molecules can and does include a radiation component, if $v > c_m$.

Čerenkov radiation is concentrated on a traveling cone with apex on the source particle.

51.5 APPLICATION: PARTICLE IDENTIFICATION IN UNDERGROUND DETECTORS

[Not ready yet.]Figure 51.1.

²Or Vavilov–Čerenkov radiation (named after S. Vavilov and P. Čerenkov, who observed it experimentally). But it was predicted theoretically by Oliver Heaviside, in papers published in 1888–89.

³Legend 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 from individual cosmic ray particles.

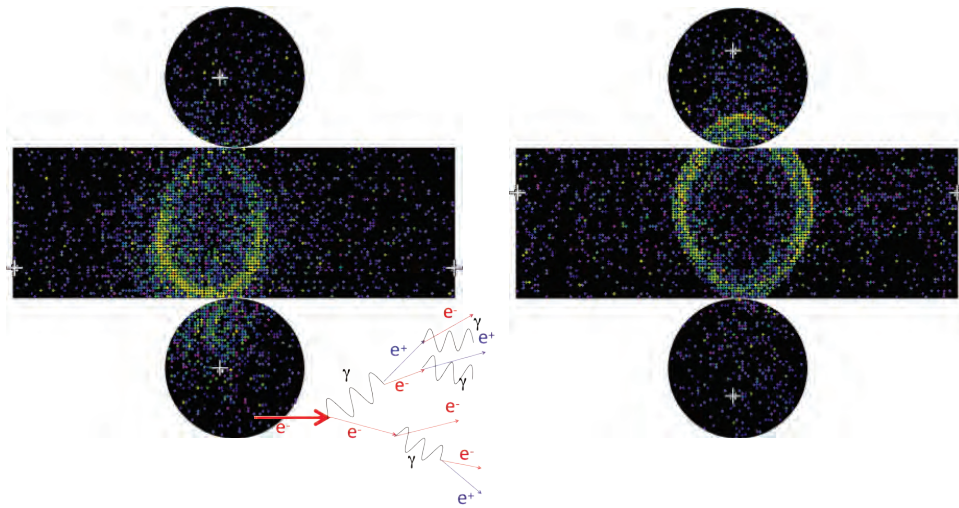


Figure 51.1: [Computer simulation.] **Particle-ID discrimination.** *Left:* Electrons undergo multiple scattering and bremsstrahlung, which then produce electron-positron pairs and a diffuse Čerenkov cone. *Right:* Muons scatter little, leading to a sharp Čerenkov cone. [Images by T2K collaboration, courtesy Atsuko K. Ichikawa.]

FURTHER READING

Intermediate:

Smith, 1997; Ginzburg, 1989.

Technical:

Historical: Jelley, 1958.

PROBLEMS

51.1 *What a shock*

Background: Chapter 41 worked out the potentials created by a point charge in vacuum, in uniform, straight-line motion, by using the Green function solution. 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).

In this problem, you'll consider fields in a dielectric medium, perhaps water. There is an approximate regime (fields not too strong, time variation not too fast) in which we may forget the medium and just replace ϵ_0 by some larger value ϵ , the “permittivity” of the medium. (You may neglect the analogous possibility for magnetic fields because $\mu \approx \mu_0$ for many dielectric media). Also neglect dispersion (assume ϵ is constant). Then we just get Maxwell's equations, and in particular the wave equation, in their usual form apart from a reduced value of the speed of light $c_m = (\mu_0\epsilon)^{-1/2}$. Hence, the Green function is the same apart from that one change.

We can now consider the problem of a charged particle that cruises through this medium at uniform speed $\tilde{\beta}c_m$. If $\tilde{\beta} < 1$, then everything is exactly the same as before, and we find that (in this approximate treatment of the medium) the charged particle just carries a blob of field energy along with it, and in particular there is no energy radiated out to infinity.

The interesting new Electromagnetic Phenomenon concerns the possibility that now $\tilde{\beta}$ may *exceed 1*. No physical law forbids a particle from moving through water at, say $0.9c$, which is $\approx 1.2c_m$. Now, however, we are on new territory. The modified Maxwell equations have a Lorentz-type invariance, but no transformation of this form can bring a particle from rest to faster than c_m , so we may not obtain the fields in the easy way used in Chapter 33. Nevertheless, the proof that the Green function solves the equations is still valid, so we can still use the method of Chapter 41 with appropriate changes.

Do: The main text argued that, for $\tilde{\beta} < 1$, there was always exactly one source point in the past light-cone of any observation point.

- a. Show that, for $\tilde{\beta} > 1$, at any time t some observation points have *no* source point in their past light-cone. The fields at such points, at time t , must equal *zero*. Characterize the set of all such points. [*Hint:* Start by finding the appropriate modification of Figure 41.4, that is, in the $(c_m t)(z)$ plane. Then generalize to two space dimensions (modify Figures 41.5a,b), adapting the “two sticks” argument to show that some angles θ cannot be attained.]

[*Remark:* Your diagrams should be accurate enough to be convincing. You could get some software to help you with this. Alternatively, even a straightedge and some bottle-tops of various sizes can give you nice lines and circles, better than freehand drawing.]

- b. Make the needed changes to the “third proof” in Section 41.5.2, page 546. Show that outside the forbidden region you found in (a), all observation points have *two* source points in their past light cone.⁴

⁴Right on the edge of the forbidden region, those two points merge into one.

- c. Then get expressions for the scalar and vector potentials.
- d. Compute appropriate derivatives to find what direction \vec{E} and \vec{B} , and hence the Poynting vector, poynt. 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 , φ , and z) may be useful:

$$\vec{\nabla} \times \vec{A} = \hat{r} \left(r^{-1} \frac{\partial \vec{A}_z}{\partial \varphi} - \frac{\partial \vec{A}_\varphi}{\partial z} \right) + \hat{\varphi} \left(\frac{\partial \vec{A}_r}{\partial z} - \frac{\partial \vec{A}_z}{\partial r} \right) + \hat{z} \left(r^{-1} \frac{\partial}{\partial r} (r \vec{A}_\varphi) - r^{-1} \frac{\partial \vec{A}_r}{\partial \varphi} \right).$$

Here r is distance from the z axis; \hat{r} , $\hat{\varphi}$, and \hat{z} are all unit vectors; and $\vec{A}_\rho = \hat{\rho} \cdot \vec{A}$ and so on.]

CHAPTER 54

Waves in a Cold Plasma and the Faraday Effect

Who knows then, but there may be, as the Antients thought, a region of this fire above our atmosphere, prevented by our air, and its own too great distance for attraction, from joining our earth? . . . 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 EFFECT

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.¹ 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 effect.” (Unlike the Zeeman effect (Problem 18.8, page 281), which involves one atom at a time and requires very large magnetic fields to be visible, we’ll see that the Faraday effect accumulates over many atoms and hence can be seen with field strengths available in Faraday’s day.) Far from being a historical curiosity, the *Faraday effect* is used today to give us evidence of strong magnetic fields in astrophysical objects, and in other research areas as well.

One reason that electromagnetic effects on light are hard to observe is that Maxwell’s 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. Accordingly, we must look for *nonlinear* effects, which can arise when light interacts with matter. This chapter will mainly study the Faraday effect (and other wave phenomena), in the context of plasmas.

[\[\[. . . Synchrotron radiation from an accretion disk is polarized, and so this rotation can be used to disclose strong magnetic fields in plasma regions that intervene between the disk and an observer.\]\]](#)

Electromagnetic phenomenon: Light from a black-hole accretion disk has a pattern of polarization.

Physical idea: Strong magnetic fields give rise both to polarized synchrotron radiation, and to Faraday rotation of the resulting light.

¹The first electro-optical nonlinear effect was ultimately seen by J. Kerr in 1875 (Section 50.4, page 628).

54.2 APPROXIMATIONS

A plasma is a partially (or fully) ionized gas, or more generally 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, the discharge inside a fluorescent light bulb remains cool to the touch, and certainly the Earth's ionosphere is very cold; in each case, some agency other than thermal collisions keeps the atoms ionized.² If thermal motion is negligible, then there is no gas pressure, no Debye screening, no frictional drag on particles, and so on.

As further approximations, Section 54.3 will consider only low-amplitude electric and magnetic fields; 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:

$$\vec{E}(t, \vec{r}) = \frac{1}{2} \vec{E} e^{-i\omega t + i\vec{k} \cdot \vec{r}} + \text{c.c.}, \quad \vec{B}(t, \vec{r}) = \frac{1}{2} \vec{B} e^{-i\omega t + i\vec{k} \cdot \vec{r}} + \text{c.c.}$$

With perhaps less justification, 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 stating what the electrons are doing. We focus on one representative electron, subject to the Lorentz force law, which again is linear in the electric and magnetic fields. So we may again suppose its response to contain a single angular frequency:

$$\vec{r}(t) = \frac{1}{2} \vec{r} e^{-i\omega t} + \text{c.c.}$$

The electron is assumed to feel others only via a mean field that they create, which is included in \vec{E} and \vec{B} . Its motion is nonrelativistic because fields are weak, so we may use newtonian mechanics and also neglect the magnetic force. Thus, \vec{r} satisfies

$$(-i\omega)^2 m_e \vec{r} = q \vec{E}, \quad (54.1)$$

²In a fluorescent tube, the passage of an electric arc; in the ionosphere, bombardment by solar wind and hard ultraviolet.

where $q = -e$ is the electron charge. Solving for \vec{r} and taking a time derivative yields the velocity:

$$\vec{v}(t) = \frac{1}{2}(-i\omega)\left(\frac{-q}{m_e\omega^2}\right)\vec{E}e^{-i\omega t} + \text{c.c.}$$

So the electron executes motion in a plane perpendicular to \vec{k} , that is, $\vec{r} \cdot \vec{k} = 0$. Conveniently, the fields of our assumed plane wave are all constant over any plane perpendicular to \vec{k} .

We now assume that many electrons, at number density c_e , are all doing this dance, each with the phase appropriate to its plane of motion. Averaged over space, their motion sets up a charge flux $\vec{j} = q\vec{v}c_e$, or

$$\vec{j} = \frac{1}{2}ic_e\left(\frac{q^2}{m_e\omega}\right)\vec{E}e^{-i\omega t+i\vec{k}\cdot\vec{r}} + \text{c.c.} \quad (54.2)$$

This relation superficially resembles that in an ohmic material (Equation 8.7, page 115), but there is a crucial difference:

- We did not find that charge flux (a real vector) is a (real) constant times electric field (a real vector).
- Rather, we found that the complex quantity \vec{j} is a complex constant times the complex quantity \vec{E} .

Had the constant of proportionality in the second statement been real, that would have implied the first statement; but in our case, the constant is *purely imaginary*. You may hear people say, “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:³

Your Turn 54A

- Show that, unlike in an ohmic material, the time-averaged power dissipation per volume equals zero.
- Show that $\vec{\nabla} \cdot \vec{j} = 0$, and hence that it was self-consistent for us to have assumed that electron density is constant in our trial solution.

(Other waves, involving compression, may also exist. Our transverse trial solution is sometimes called “electromagnetic” to distinguish it from these “acoustic” modes.)

54.3.2 The dispersion relation has a cutoff

The next steps are familiar.⁴ We already know that our trial solution satisfies the electric Gauss law.⁵ Imposing Faraday’s law as usual says that

$$\vec{B} = (\vec{k}/\omega) \times \vec{E}, \quad (54.3)$$

³ [T2] Conversely, when a quantity that normally represents a *nondissipative* effect, for example permittivity, is assigned an imaginary part, that can represent a dissipative process (Section 49.2.1’, page 619).

⁴ See for example Problem 21.2 (page 307).

⁵ The spatially averaged charge density is zero, because the electrons are neutralized by the ions that liberated them.

which automatically follows the magnetic Gauss law. Ampère's law in a conductive medium gives

$$i\vec{k} \times \vec{B} = \mu_0 \left(i \frac{q^2 c_e}{m_e \omega} + \epsilon_0 (-i\omega) \right) \vec{E}.$$

Substituting gives

$$\vec{k} \times \left((\vec{k}/\omega) \times \vec{E} \right) = -\mu_0 \omega \left(\epsilon_0 - \frac{q^2 c_e}{m_e \omega^2} \right) \vec{E}. \quad (54.4)$$

Expanding the triple cross product and using the assumed transversality gives the dispersion relation: The trial solution indeed solves the Newton+Maxwell equations if

$$\|\vec{k}\|^2 / \omega^2 = c^{-2} \left(1 - \frac{q^2 c_e}{\epsilon_0 m_e \omega^2} \right). \quad (54.5)$$

Some abbreviations and comments are in order. First, define the **plasma frequency** as

$$\omega_p = |q| \sqrt{c_e / (\epsilon_0 m_e)}. \quad (54.6)$$

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 ϵ_0 :

$$\epsilon = \epsilon_0 (1 - (\omega_p / \omega)^2).$$

Hence, the phase velocity of our waves is:

$$v_{\text{ph}} = \omega / k = c / \sqrt{1 - (\omega_p / \omega)^2}.$$

For frequencies above the plasma frequency, this is greater than c ; that is, the index of refraction v_{ph}/c is less than one, unlike any ordinary dielectric. However, in that regime the index is at least real: Waves propagate without loss.⁶ The dependence on frequency means that *propagation is highly dispersive* for frequencies close to ω_p .

The situation gets more interesting at frequencies below the plasma frequency: Here the index of refraction is *pure imaginary*, and so $e^{i\vec{k}\cdot\vec{r}}$ is exponentially damped. When a wave in this regime, traveling in vacuum, impinges on the plasma, it cannot penetrate far. Nor is it converted to heat; instead, it must *reflect*. Section 21.4.6 (page 303) 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:

Your Turn 54B

- Check how the units work in Equation 54.6.
- Work out the group velocity $v_g = (dk/d\omega)^{-1}$ from the dispersion relation as a function of angular frequency and comment.

Electromagnetic waves in cold plasma are highly dispersive.

Electromagnetic waves in cold plasma have a cutoff frequency

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 ultraviolet light from the Sun, with density $c_e \approx 10^{11} \text{ m}^{-3}$ 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 ω_p , so we can observe it from Earth.

In addition,

- $c_e^{-1/3}$ is much smaller than wavelength of radio-frequency radiation, supporting our continuum treatment of current in Maxwell's equations.
- Although the Debye length λ_D is not zero, as it would be at zero temperature, at $\approx 2\text{ mm}$ it too is small compared to wavelength. Also, there are many electrons in a sphere of radius λ_D .

The ionosphere reflects waves with frequencies below its cutoff.

Pulsar chirp is a diagnostic for optical thickness in the intervening space.

54.3.4 Pulsar chirp results from 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 nearly delta-function pulses of energy. Instead, we hear a "chirp," with lower Fourier components arriving first, followed by the higher ones. This dispersion of the signal also delayed discovery of pulsars until long after the initial deployment of radio telescopes. Today, however, it serves as a useful diagnostic of the medium intervening between the radiation source and us, the observers. Typical plasma frequencies in space are in the kilohertz range, so at radio frequency the inverse group velocity is $\approx c^{-1}(1 + \frac{1}{2}\frac{\omega_p^2}{\omega^2})$. Then total transit time at angular frequency ω for an object at distance L is

$$\frac{L}{c} + \frac{1}{2c\omega^2} \int_0^L dx \frac{c_e(x)e^2}{m_e\epsilon_0},$$

whose frequency dependence is related to the optical thickness $\int dx c_e(x)$.

54.3.5 Metals

Some metals become transparent to light beyond a cutoff frequency.

Although the conduction of electricity through metals is quantum-mechanical in character, qualitatively they do have the property of completely reflecting light at low frequencies, while becoming partially transparent above a cutoff frequency. For the simplest metals, such as lithium or sodium, the cutoff is around $2\pi c/\omega \approx 200\text{ nm}$ (hard ultraviolet).

54.4 FARADAY'S MAGNETO-OPTICAL EFFECT

54.4.1 A plasma becomes a chiral medium in the presence of a steady magnetic field

Each circular polarization has its own velocity in a magnetized cold plasma.

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 keep the math simple, we consider only waves propagating along (or opposite to) \vec{B}_0 , for example,

$$\vec{k} = k\hat{z}, \quad \vec{B}_0 = B_0\hat{z},$$

⁶Recall Your Turn 54Aa.

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 \vec{B}_0 itself may not be small. Hence, although we continue to neglect the effect of \vec{B} on the non-relativistic electron motion, we must keep \vec{B}_0 in the Lorentz force law: Equation 54.1 becomes

$$(-\omega^2)m_e\vec{r} = q(\vec{E} + \vec{v} \times \vec{B}_0),$$

so

$$\vec{r} = -\frac{q}{m_e\omega^2}(-i\omega B_0\vec{r} \times \hat{z} + \vec{E}). \quad (54.7)$$

We are familiar with this equation in the *absence* of any wave: Electrons undergo cyclotron motion.⁷ 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 suppose a trial solution in which the complex polarization \vec{E} points along one of the two circular basis vectors,⁸ for example:

$$\vec{E} = a\hat{\zeta}_{(+)}, \text{ where } \hat{\zeta}_{(\pm)} = (\hat{x} \pm i\hat{y})/\sqrt{2}. \quad [18.32, \text{ page 272}]$$

Equivalently, we may state the components of \vec{E} : $\vec{E}_+ = a$, $\vec{E}_- = 0$.

In Problem 18.7, you showed the useful identity that the circular basis vectors are eigenvectors of the operation $\hat{z} \times$, that is,

$$\hat{z} \times \hat{\zeta}_{(\pm)} = \mp i\hat{\zeta}_{(\pm)}. \quad (54.8)$$

Equation 54.7 then separates into two decoupled equations:

$$-\frac{m_e\omega^2}{q}\vec{r}_+ = a + i\omega B_0(-i)\vec{r}_+ \quad (54.9)$$

$$-\frac{m_e\omega^2}{q}\vec{r}_- = i\omega B_0(-i)\vec{r}_-. \quad (54.10)$$

The second of these equations clearly has $\vec{r}_- = 0$ as its solution. The other is more interesting:

$$\vec{r}_+ = -a\left(\frac{m_e\omega^2}{q} + \omega B_0\right)^{-1}. \quad (54.11)$$

Proceeding as before gives the charge flux

$$\vec{j} = \frac{1}{2}(i)c_e\left(\frac{q^2}{m_e\omega + B_0q}a\hat{\zeta}_{(+)}\right)e^{-i\omega t + i\vec{k}\cdot\vec{r}} + \text{c.c.}, \quad (54.12)$$

which reassuringly reduces to Equation 54.2 when $B_0 = 0$.

Again using Equation 54.8, Equation 54.3 becomes

$$\vec{B} = -i(k/\omega)a\hat{\zeta}_{(+)}, \quad (54.13)$$

⁷Section 33.3.5 (page 445).

⁸You'll investigate the other circular polarization in Your Turn 54C.

and hence Equation 54.4 is modified to

$$(-ik)^2 \omega^{-1} a_{\hat{\zeta}(+)}) = -\mu_0 \omega \left(\epsilon_0 - \frac{q^2 c_e}{m_e \omega^2 + B_0 q \omega} \right) a_{\hat{\zeta}(+)}. \quad (54.14)$$

The dispersion relation is thus

$$k^2 = \omega^2 \mu_0 \epsilon_0 \left(1 - \frac{q^2 c_e / \epsilon_0}{m_e \omega^2 + B_0 q \omega} \right), \quad (54.15)$$

which again reduces to Equation 54.5 when $B_0 = 0$.

Your Turn 54C

- Check how the units work in Equation 54.15.
- Redo the derivation for the other helicity (circular polarization) of light and note the difference in dispersion relation.
- Explain how one might have expected the result in (b) on symmetry grounds.

In short, *the originally isotropic plasma has acquired circular birefringence*. From here, the analysis is much the same as in Section 49.7 (page 615); see Problem 54.2.

54.4.2 A one-way light valve

[Not ready yet.] en.wikipedia.org/wiki/Optical_isolator

Faraday rotation combined with polarizing filters creates a one-way valve for light.

54.4.3 The accretion disk of M87*

[Not ready yet.] ... the front cover of this book (see also page iv).

Light from a black-hole accretion disk has a pattern of polarization.

54.4.4 The Faraday effect also appears in condensed matter

Looking back, we may interpret our result by saying that a background magnetic field induces cross-susceptibility.⁹ The imposed electric field 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 crystals break reflection invariance in a particular way (imposing a handedness), and so does a uniform magnetic field.

[[Not ready yet.]...Kerr... "FitzGerald pointed out that since different indices of refraction imply different intensities of reflection, the left- and right-handed components of a polarized beam should have different amplitudes after reflection from a magnetized surface and so should recombine into an elliptically polarized beam of the kind Kerr had observed." – Hunt p15]

Although a calculation of the magneto-optical rotation from first principles is daunting, we may nevertheless expect that the rotation of linearly polarized light will be proportional to B_0 (to leading order) and also to path length traversed. The

⁹See Section 49.6.

“constant” of proportionality, called the **Verdet constant**, is actually a function of wavelength, as we saw for a cold plasma. Its spectrum characterizes the substance under study. Crystals of terbium gallium garnet have an unusually large value, around $-134 \text{ rad}/(\text{T m})$. Organic materials have smaller Verdet constants, in the visible wavelength region typically on the order of a several tens of $\text{rad}/(\text{T m})$.

FURTHER READING

Semipopular:

Faraday effect: Johnson, 2008; Media 18; en.wikipedia.org/wiki/Faraday_effect.

M87: scitechdaily.com/astromers-polarized-image-shows-magnetic-fields-at-the-edge-of-m87s-black-hole/.

“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/.

Intermediate:

Waves in cold plasma: Thorne & Blandford, 2017, chap. 21; Lifshitz & Pitaevskiĭ, 1981; Rybicki & Lightman, 2004.

Magneto-optical effects in condensed matter: Landau et al., 1984, §101.

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 .

[Not ready yet.]

54.2 *Rotation measure*

Suppose that a linearly polarized plane wave with some angular frequency ω enters a cold plasma, propagating along the direction of a uniform background magnetic field of strength B_0 .

Magnetic fields span intergalactic space.

- 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.
- 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. 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 54.3.4.]

CHAPTER 56

Vista: Field Quantization

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*

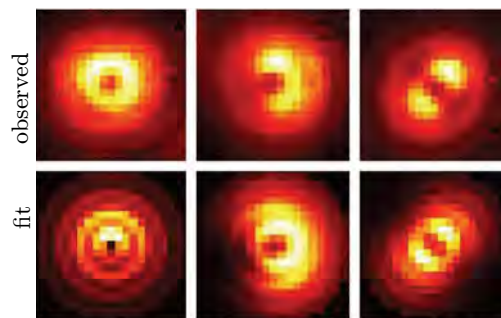
56.1 FRAMING: *EX NIHIL*

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 56.1). 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 nihil* 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 56.1: [Experimental data and fits.] **De-focused 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 deg, 60 deg, and 90 deg. The in-plane orientation (azimuth) was also obtained by fitting. [From Toprak et al., 2006.]



56.2 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 56.3 will find the photon concept as a consequence of field quantization.

As usual, we can represent electric and magnetic fields via a scalar potential field, $\psi(t, \vec{r})$, and a vector potential field, $\vec{A}(t, \vec{r})$:

$$\vec{E} = -\frac{\partial}{\partial t}\vec{A} - \vec{\nabla}\psi; \quad \vec{B} = \vec{\nabla} \times \vec{A}. \quad [18.26, \text{page 268}]$$

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 ψ 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

$$\vec{A}(t, \vec{r}) = \frac{1}{2} \sum'_{\vec{k}} (\vec{A}_{\vec{k}}(t) e^{i\vec{k} \cdot \vec{r}} + \text{c.c.}). \quad (56.1)$$

In this formula, each coefficient $\vec{A}_{\vec{k}}$ 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 $2\pi\eta_i/L$; the η_i 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 $\vec{k} \cdot \vec{A}_{\vec{k}} = 0$, or in other words that the component of each $\vec{A}_{\vec{k}}$ 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 $\hat{\zeta}_{(\alpha;\vec{k})}$, where the index α runs from 1 to 2. Then Equation 56.1 becomes

$$\vec{A}(t, \vec{r}) = \frac{1}{2} \sum'_{\alpha, \vec{k}} (A_{\alpha, \vec{k}}(t) \hat{\zeta}_{(\alpha;\vec{k})} e^{i\vec{k} \cdot \vec{r}} + \text{c.c.}). \quad (56.2)$$

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 $A_{\alpha, \vec{k}}(t)$.

¹Section 18.8.2 (page 268).

Your Turn 56A

Show that, with these definitions, the Maxwell equations in Coulomb gauge become simple:

$$\frac{d^2}{dt^2} A_{\alpha, \vec{k}} = -\|c\vec{k}\|^2 A_{\alpha, \vec{k}}. \quad (56.3)$$

Here α runs over 1,2 and \vec{k} runs over the nonredundant set described earlier.

Equation 56.3 shows that every distinct combination of polarization α 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 $A_{\alpha, \vec{k}}$:

$$A_{\alpha, \vec{k}} = (\epsilon_0 L^3 / 2)^{-1/2} (X_{\alpha, \vec{k}} + iY_{\alpha, \vec{k}}). \quad (56.4)$$

The overall rescaling chosen in the definitions of X and Y will simplify some later formulas.

The real scalar quantities $X_{\alpha, \vec{k}}$ and $Y_{\alpha, \vec{k}}$ separately obey Equation 56.3, so we see that

The vacuum Maxwell equations are mathematically equivalent to a set of decoupled harmonic oscillators. (56.5)

The harmonic oscillator has a well known quantum-mechanical formulation, so Idea 56.5 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} 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 35Ca (page 484) gives

$$\begin{aligned} \mathcal{E} &= \frac{\epsilon_0}{2} \int d^3r (\vec{E}^2 + c^2 \vec{B}^2) = \frac{\epsilon_0}{2} \int d^3r ((-\dot{\vec{A}})^2 + c^2 (\vec{\nabla} \times \vec{A})^2) \\ &= \frac{\epsilon_0}{2} \sum'_{\alpha, \vec{k}_1} \sum'_{\beta, \vec{k}_2} \int d^3r \left[\frac{1}{2} (-\dot{A}_{\alpha, \vec{k}_1} \hat{\zeta}_{(\alpha; \vec{k}_1)} e^{i\vec{k}_1 \cdot \vec{r}} + \text{c.c.}) \cdot \frac{1}{2} (-\dot{A}_{\beta, \vec{k}_2} \hat{\zeta}_{(\beta; \vec{k}_2)} e^{i\vec{k}_2 \cdot \vec{r}} + \text{c.c.}) \right. \\ &\quad \left. + c^2 \frac{1}{2} (A_{\alpha, \vec{k}_1} i\vec{k}_1 \times \hat{\zeta}_{(\alpha; \vec{k}_1)} e^{i\vec{k}_1 \cdot \vec{r}} + \text{c.c.}) \cdot \frac{1}{2} (A_{\beta, \vec{k}_2} i\vec{k}_2 \times \hat{\zeta}_{(\beta; \vec{k}_2)} e^{i\vec{k}_2 \cdot \vec{r}} + \text{c.c.}) \right]. \end{aligned} \quad (56.6)$$

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 $e^{i\vec{k}_1 \cdot \vec{r}} e^{-i\vec{k}_1 \cdot \vec{r}} = 1$, survive. Moreover, we have $\hat{\zeta}_{(\alpha; \vec{k})} \cdot \hat{\zeta}_{(\beta; \vec{k})} = \delta_{\alpha\beta}$, leaving

$$\begin{aligned} \mathcal{E} &= \frac{\epsilon_0 L^3}{4} \sum'_{\alpha, \vec{k}} (|\dot{A}_{\alpha, \vec{k}}|^2 + \|c\vec{k}\|^2 |A_{\alpha, \vec{k}}|^2) \\ &= \frac{1}{2} \sum'_{\alpha, \vec{k}} (\dot{X}_{\vec{k}, \alpha}^2 + \|c\vec{k}\|^2 X_{\alpha, \vec{k}}^2 + \dot{Y}_{\vec{k}, \alpha}^2 + \|c\vec{k}\|^2 Y_{\alpha, \vec{k}}^2). \end{aligned} \quad (56.7)$$

The field momentum is given by a similar calculation, starting with the Poynting

vector (Your Turn 35C, page 484b):

$$\begin{aligned}
\vec{P} &= \epsilon_0 \int d^3r \vec{E} \times \vec{B} & (56.8) \\
&= \epsilon_0 \sum'_{\alpha, \vec{k}_1} \sum'_{\beta, \vec{k}_2} \int d^3r \frac{1}{2} (-\dot{A}_{\alpha, \vec{k}_1} \hat{\zeta}_{(\alpha; \vec{k}_1)} e^{i\vec{k}_1 \cdot \vec{r}} + \text{c.c.}) \times \left(\vec{\nabla} \times \frac{1}{2} (A_{\beta, \vec{k}_2} \hat{\zeta}_{(\beta; \vec{k}_2)} e^{i\vec{k}_2 \cdot \vec{r}} + \text{c.c.}) \right) \\
&= -\frac{\epsilon_0 L^3}{4} \sum'_{\alpha, \vec{k}} \sum'_{\beta} (\dot{A}_{\alpha, \vec{k}} A_{\beta, \vec{k}}^* \hat{\zeta}_{(\alpha; \vec{k})} \times (-i\vec{k} \times \hat{\zeta}_{(\beta; \vec{k})}) + \text{c.c.}) \\
&= \frac{\epsilon_0 L^3}{4} \sum'_{\alpha, \vec{k}} (i\vec{k} \dot{A}_{\alpha, \vec{k}} A_{\alpha, \vec{k}}^* + \text{c.c.}) \\
&= \frac{1}{2} \sum'_{\alpha, \vec{k}} \vec{k} ((i\dot{X}_{\alpha, \vec{k}} - \dot{Y}_{\alpha, \vec{k}})(X_{\alpha, \vec{k}} - iY_{\alpha, \vec{k}}) + \text{c.c.}) \\
&= \sum'_{\alpha, \vec{k}} \vec{k} (\dot{X}_{\alpha, \vec{k}} Y_{\alpha, \vec{k}} - \dot{Y}_{\alpha, \vec{k}} X_{\alpha, \vec{k}}). & (56.9)
\end{aligned}$$

We now have compact formulas for the energy and momentum of the electromagnetic field in terms of the harmonic-oscillator representation (Equation 56.7 and 56.9). The interpretation is that every mode of the field, labeled by α and \vec{k} , makes an independent contribution to \mathcal{E} , and also to each component of \vec{P} . Note, however, that the momentum gets mixed contributions from the X and Y oscillators. We will soon remove this remaining inconvenience.

56.3 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 X and Y variables by a set of quantum operators called \mathbf{Q} and their hermitian conjugates (Equation 56.21). 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 α, \vec{k} . We introduce two hermitian operators² \mathbf{X} and \mathbf{U} , with the property that their commutator is $[\mathbf{X}, \mathbf{U}] = i\hbar$. In the energy function, Equation 56.7, we substitute $X \rightarrow \mathbf{X}$ and $\dot{X} \rightarrow \mathbf{U}$ to obtain the hamiltonian operator for X :

$$\mathbf{H}_X = \frac{1}{2} (\mathbf{U}^2 + \|c\vec{k}\|^2 \mathbf{X}^2). \quad (56.10)$$

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 $\exp(-i\mathbf{H}_X t/\hbar)|\Psi\rangle$.

²In 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} .

It implies that

$$\begin{aligned}\frac{d^2}{dt^2}\langle\Psi_1|\mathbf{X}|\Psi_2\rangle &= \frac{d}{dt}\langle\Psi_1|\frac{i}{\hbar}[\mathbf{H}_X, \mathbf{X}]\Psi_2\rangle = \frac{d}{dt}\langle\Psi_1|\mathbf{U}|\Psi_2\rangle = \langle\Psi_1|\frac{i}{\hbar}[\mathbf{H}_X, \mathbf{U}]\Psi_2\rangle \\ &= -\|c\vec{k}\|^2\langle\Psi_1|\mathbf{X}|\Psi_2\rangle,\end{aligned}\quad (56.11)$$

which implements the classical equation of motion for the harmonic oscillator in Equation 56.3.

We proceed in the same way with the other oscillator family, introducing hermitian operators \mathbf{Y} and \mathbf{V} analogous to \mathbf{X} and \mathbf{U} . Then the operator corresponding to $A_{\alpha, \vec{k}}$ in Equation 56.4 is

$$\mathbf{A} = (\epsilon_0 L^3/2)^{-1/2}(\mathbf{X} + i\mathbf{Y}). \quad (56.12)$$

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 (\mathbf{X}, \mathbf{U}) and (\mathbf{Y}, \mathbf{V}) . However, the harmonic oscillator problem has an elegant reformulation that simplifies the math. Change variables once again by defining new operators

$$\mathbf{S} = (2\hbar\|c\vec{k}\|)^{-1/2}(\|c\vec{k}\|\mathbf{X} + i\mathbf{U}) \text{ and } \mathbf{R} = (2\hbar\|c\vec{k}\|)^{-1/2}(\|c\vec{k}\|\mathbf{Y} + i\mathbf{V}). \quad (56.13)$$

\mathbf{S} and \mathbf{R} are not hermitian; indeed, it is straightforward to verify that

$$[\mathbf{S}, \mathbf{S}^\dagger] = 1, \quad [\mathbf{R}, \mathbf{R}^\dagger] = 1, \quad [\mathbf{S}, \mathbf{R}] = [\mathbf{S}, \mathbf{R}^\dagger] = 0, \quad (56.14)$$

$$\mathbf{H} = \mathbf{H}_X + \mathbf{H}_Y = \hbar\|c\vec{k}\|(\mathbf{S}^\dagger\mathbf{S} + \mathbf{R}^\dagger\mathbf{R} + 1), \text{ and} \quad (56.15)$$

$$\vec{\mathbf{P}} = i\hbar\vec{k}(\mathbf{S}^\dagger\mathbf{R} - \text{h.c.}). \quad (56.16)$$

In the last formula, “h.c.” denotes the hermitian conjugate, that is, $\mathbf{R}^\dagger\mathbf{S}$.

Step 3: Diagonalize momentum

The hamiltonian operator has the nice property that \mathbf{S} and \mathbf{R} make independent, additive contributions to it (Equation 56.15). The momentum operator still mixes \mathbf{S} and \mathbf{R} , but we can diagonalize it, without spoiling \mathbf{H} , by a unitary transformation. Define two new **lowering operators** by

$$\mathbf{Q} = (\mathbf{S} + i\mathbf{R})/\sqrt{2}, \quad \tilde{\mathbf{Q}} = (\mathbf{S} - i\mathbf{R})/\sqrt{2}. \quad (56.17)$$

Your Turn 56B

Show that

$$[\mathbf{Q}, \mathbf{Q}^\dagger] = 1, \quad [\tilde{\mathbf{Q}}, \tilde{\mathbf{Q}}^\dagger] = 1, \quad [\mathbf{Q}, \tilde{\mathbf{Q}}] = [\mathbf{Q}, \tilde{\mathbf{Q}}^\dagger] = 0, \quad (56.18)$$

$$\mathbf{H} = \hbar\|c\vec{k}\|(\mathbf{Q}^\dagger\mathbf{Q} + \tilde{\mathbf{Q}}^\dagger\tilde{\mathbf{Q}} + 1), \text{ and} \quad (56.19)$$

$$\vec{\mathbf{P}} = \hbar\vec{k}(\mathbf{Q}^\dagger\tilde{\mathbf{Q}} - \tilde{\mathbf{Q}}^\dagger\mathbf{Q}). \quad (56.20)$$

We now have new field operators Q and \tilde{Q} that, unlike S and R , diagonalize both the field energy and momentum.

Step 4: Relabel

We now reinstate the mode indices α 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 $\tilde{Q}_{\alpha, \vec{k}}$ as $Q_{\alpha, -\vec{k}}$. Then

$$[Q_{\alpha, \vec{k}_1}, Q_{\beta, \vec{k}_2}^\dagger] = \delta_{\alpha\beta} \delta_{\vec{k}_1, \vec{k}_2}, \quad [Q_{\alpha, \vec{k}_1}, Q_{\beta, \vec{k}_2}] = 0, \quad \text{for all nonzero } \vec{k}_1 \text{ and } \vec{k}_2. \quad (56.21)$$

Our final formulas then become unrestricted sums:

$$H = \sum_{\alpha, \vec{k}} \hbar \|c\vec{k}\| \left(Q_{\alpha, \vec{k}}^\dagger Q_{\alpha, \vec{k}} + \frac{1}{2} \right), \text{ and} \quad (56.22)$$

$$\vec{P} = \sum_{\alpha, \vec{k}} \hbar \vec{k} \left(Q_{\alpha, \vec{k}}^\dagger Q_{\alpha, \vec{k}} \right). \quad (56.23)$$

We now have a set of operators in terms of which the energy and momentum of light will have simple interpretations.

56.4 PHOTON STATES

56.4.1 Basis states can be formed by applying creation operators to the vacuum state

Particles can be created from, or annihilated into, energy.

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.

Your Turn 56C

Show that

$$[H, Q_{\alpha, \vec{k}}] = -\hbar \|c\vec{k}\| Q_{\alpha, \vec{k}} \quad \text{and} \quad [\vec{P}, Q_{\alpha, \vec{k}}] = -\hbar \vec{k} Q_{\alpha, \vec{k}}. \quad (56.24)$$

Equations 56.24 justify the term “lowering operator”:

Applying the lowering operator $Q_{\alpha, \vec{k}}$ to a state lowers its energy by $\hbar \|c\vec{k}\|$, and changes its momentum by $-\hbar \vec{k}$. Conversely, applying the raising operator $Q_{\alpha, \vec{k}}^\dagger$ has the opposite effects. (56.25)

Next, note that both of the terms in the classical electromagnetic energy function (Equation 56.6) are nonnegative. So it must not be possible to lower that energy indefinitely; there must be a state for which any lowering 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 raising operators, each of which may be applied any number of times, always raising the energy by $\hbar \|c\vec{k}\|$ 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*,³ each carrying energy $\hbar\|\vec{c}\vec{k}\|$ and momentum $\hbar\vec{k}$.

Your Turn 56D

Show that when a raising operator acts n times, we can obtain a normalized state as follows:

$$|n_{\alpha,\vec{k}}\rangle = \sqrt{\frac{1}{n!}} (\mathbf{Q}_{\alpha,\vec{k}}^\dagger)^n |0\rangle. \quad (56.26)$$

More generally, we can define $|n_{\alpha_1,\vec{k}_1}; n_{\alpha_2,\vec{k}_2}; \dots\rangle$ as a state obtained by applying several different raising 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”). (56.27)

That is, each one-photon basis state is labeled by a wavevector and a polarization, and carries energy and momentum related by the Einstein and de Broglie relations Equation 56.24:

$$\mathcal{E}_{\alpha,\vec{k}} = \hbar\|\vec{c}\vec{k}\|; \quad \vec{p}_{\alpha,\vec{k}} = \hbar\vec{k}; \quad \text{so } \mathcal{E}_{\alpha,\vec{k}} = \|c\vec{p}_{\alpha,\vec{k}}\|, \quad (56.28)$$

implying that photons are massless (Equation 31.15, page 415). 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 raising and lowering operators: Because they can be interpreted as raising and lowering the *number of photons* in a state, they are also called **creation and destruction 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 436).

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 α, \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 raising operators, because those operators all commute with one another.⁴

³See Section 31.4.1 (page 415) and Section 32.7.2 (page 436).

⁴More precisely, a class of particles that are indistinguishable in this way is called “bosonic.” Another possibility, called “fermionic” particles, has raising operators that mutually *anticommute*.

56.4.2 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 $\vec{A}(\vec{r})$ involves both raising and lowering operators:

Your Turn 56E

Use Equations 56.2, 56.12, 56.13, and 56.17 to show that

$$\vec{A}(\vec{r}) = \sum_{\alpha, \vec{k}} \sqrt{\frac{\hbar}{2L^3 \epsilon_0 \|c\vec{k}\|}} \hat{\zeta}_{(\alpha; \vec{k})} (\mathcal{Q}_{\alpha, \vec{k}} e^{i\vec{k} \cdot \vec{r}} + \text{h.c.}). \quad (56.29)$$

However, we can find eigenvectors of $\mathcal{Q}_{\alpha, \vec{k}}$, called **coherent states**: For any complex number u , define

$$|u, \alpha, \vec{k}\rangle = \exp(-\frac{1}{2}|u|^2) \sum_{n=0}^{\infty} (n!)^{-1/2} (u)^n |n_{\alpha, \vec{k}}\rangle. \quad (56.30)$$

Your Turn 56F

- Show that the states $|u, \alpha, \vec{k}\rangle$ just defined are all properly normalized for any complex number u .
- Show that $\mathcal{Q}_{\alpha, \vec{k}} |u, \alpha, \vec{k}\rangle = u |u, \alpha, \vec{k}\rangle$, and hence also $\langle u, \alpha, \vec{k} | \mathcal{Q}_{\alpha, \vec{k}}^\dagger = u^* \langle u, \alpha, \vec{k} |$.
- Then show that Equation 56.29 implies

$$\langle u, \alpha, \vec{k} | \vec{A}(\vec{r}) |u, \alpha, \vec{k}\rangle = (2L^3 \epsilon_0 \|c\vec{k}\| / \hbar)^{-1/2} \hat{\zeta}_{(\alpha; \vec{k})} u e^{i\vec{k} \cdot \vec{r}} + \text{c.c.}$$

Your Turn 56G

The coherent states are superpositions of states with different numbers of photons. Find the probabilities of getting exactly ℓ photons in a measurement on a coherent state by computing the modulus squared of each individual term of Equation 56.30. Is this a distribution you have seen previously?

Your result in Your Turn 56F shows that the coherent state based on a particular wavevector and polarization is the quantum analog of a classical single-mode state (Equation 56.2, page 656). Moreover, Your Turn 56G 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.⁵

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.

56.5 INTERACTION WITH ELECTRONS

56.5.1 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 ψ , though we can still impose $\vec{\nabla} \cdot \vec{A} = 0$. The electric Gauss law then says

$$\vec{\nabla} \cdot \vec{E} = -\nabla^2 \psi = \rho_q / \epsilon_0, \quad [2.4, \text{page } 28]$$

where ρ_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 $\vec{j}(t, \vec{r})$:

$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{j} + \mu_0 \epsilon_0 \frac{\partial \vec{E}}{\partial t}. \quad [18.17, \text{page } 262]$$

Casting everything into plane wave mode expansions as before gives the full Maxwell equations as

$$k^2 \psi_{\vec{k}} = \frac{1}{\epsilon_0} \rho_{q, \vec{k}} \quad \text{and} \quad (56.31)$$

$$\frac{d^2}{dt^2} \vec{A}_{\vec{k}} + \|c\vec{k}\|^2 \vec{A}_{\vec{k}} = -i\vec{k} \frac{d\psi_{\vec{k}}}{dt} + \frac{1}{\epsilon_0} \vec{j}_{\vec{k}}, \quad (56.32)$$

where $c = (\mu_0 \epsilon_0)^{-1/2}$ and $\psi_{\vec{k}}$, $\rho_{q, \vec{k}}$, and $\vec{j}_{\vec{k}}$ are the plane-wave components of ψ , ρ_q , and \vec{j} , respectively. We now take the dot product of both sides of Equation 56.32 with the two transverse basis vectors $\hat{\zeta}_{(\alpha; \vec{k})}$ to find the desired generalization of Equation 56.3:

$$\frac{d^2}{dt^2} A_{\alpha, \vec{k}} = -\|c\vec{k}\|^2 A_{\alpha, \vec{k}} + \frac{1}{\epsilon_0} \vec{j}_{\vec{k}} \cdot \hat{\zeta}_{(\alpha; \vec{k})} \quad \text{for each } \vec{k}, \alpha. \quad (56.33)$$

The scalar potential ψ has dropped out of this equation of motion.

56.5.2 Electromagnetic interactions can be treated perturbatively

There is no need to quantize the scalar potential ψ , 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.

⁵Books 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).

The last term of Equation 56.33 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 ψ as usual. There is another term describing the free electromagnetic field (Equation 56.22). To these terms we then add the perturbation

$$- \int d^3r \vec{j}(\vec{r}) \cdot \vec{A}(\vec{r}), \quad (56.34)$$

where $\vec{j}(\vec{r})$ is the operator version of the charge flux and $\vec{A}(\vec{r})$ is given by Equation 56.29. This term modifies the quantum equations of motion, introducing the last part of Equation 56.33.

Each electron in the atom or molecule of interest contributes a delta function to \vec{j} that is localized at the electron’s position \vec{r}_e , with strength equal to its charge, $-e$, times its velocity,⁶ \vec{p}_e/m_e . Thus, each electron makes a contribution to the integral in Equation 56.34 equal to

$$- \sum_{\alpha, \vec{k}} \sqrt{\frac{\hbar}{2L^3 \epsilon_0 \|c\vec{k}\|}} \hat{\zeta}_{(\alpha; \vec{k})} \cdot (-e)(\vec{p}_e/m_e) (\mathbf{Q}_{\alpha, \vec{k}} e^{i\vec{k} \cdot \vec{r}_e} + \text{h.c.}). \quad (56.35)$$

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 56.35 sandwiched between the initial and final states.⁷ The hermitian conjugate term, involving $\mathbf{Q}_{\alpha, \vec{k}}^\dagger$, 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, $k \approx 10^{-2} \text{ nm}^{-1}$. But r_e cannot exceed the size of the atom or molecule, typically $\approx 1 \text{ nm}$, so $\vec{k} \cdot \vec{r}_e$ 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.⁸

56.5.3 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

$$\left| \langle \text{ground}; \alpha, \vec{k} | \sum_{\beta, \vec{k}'} \mathbf{Q}_{\beta, \vec{k}'}^\dagger \hat{\zeta}_{(\beta; \vec{k}')} \cdot \vec{p}_e | \text{excited} \rangle \right|^2$$

⁶See Section 34.9.2 (page 465).

⁷Quantum mechanics textbooks call this scheme the “Golden Rule” of time-dependent perturbation theory.

⁸See Chapter 43.

$$= \left| \langle \text{ground} | \vec{p}_e | \text{excited} \rangle \cdot \hat{\zeta}_{(\alpha; \vec{k})} \right|^2. \quad (56.36)$$

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

$$[H_e, \vec{r}_e] = \frac{-i\hbar}{m} \vec{p}_e.$$

Sandwich this relation between the ground and excited states to find

$$\langle \text{ground} | (E_0 \vec{r}_e - \vec{r}_e E_*) | \text{excited} \rangle = \frac{-i\hbar}{m} \langle \text{ground} | \vec{p}_e | \text{excited} \rangle.$$

The right-hand side of this formula is a constant times the quantity needed in Equation 56.36. The left-hand side is can be written in terms of the electric dipole moment operator, $\vec{D}_E = -e\vec{r}_e$, 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.

As in the classical version, radiation is dominated by dipole emission if the transition dipole is nonzero.

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:

$$\langle \text{ground} | \vec{D}_E | \text{excited} \rangle = \mathcal{D}_E \hat{z}. \quad (56.37)$$

Suppose that, as is the case in many experiments, we record every photon received regardless of its polarization. The sum of Equation 56.36 over α includes the factor⁹

$$\sum_{\alpha} \hat{z} \cdot \hat{\zeta}_{(\alpha; \vec{k})} \hat{\zeta}_{(\alpha; \vec{k})} \cdot \hat{z}. \quad (56.38)$$

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 $\vec{\mathbf{1}} - \hat{k} \otimes \hat{k}$, so we get

$$\hat{z} \cdot (\vec{\mathbf{1}} - \hat{k} \otimes \hat{k}) \cdot \hat{z} = \hat{z} \cdot \hat{z} - (\hat{z} \cdot \hat{k})^2 = 1 - \cos^2 \theta = \sin^2 \theta, \quad (56.39)$$

where θ is the polar angle between the direction of observation, \hat{k} , and the transition dipole.

Equations 56.39 and 56.36 show that the probability density function for the angles at which photons are emitted has a “dipole doughnut” pattern:¹⁰ 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.

As in the classical version, dipole radiation is anisotropic with a “doughnut” pattern.

The mean rate at which photons are emitted is determined by the transition dipole \mathcal{D}_E defined by Equation 56.37, 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 $\wp(\theta, \varphi) \propto \sin^2 \theta$. (56.40)

⁹We chose $\hat{\zeta}_{(\alpha; \vec{k})}$ to be real vectors in Section 56.2.

¹⁰See Section 38.2.2.

This section has resolved the puzzle posed at the start of this chapter: The pattern of photon emission observed in defocused orientation imaging (Figure 56.1) agrees with the dipole radiation pattern in classical electrodynamics because the same angular factors enter each calculation.

56.6 VISTAS

56.6.1 Many invertebrates can detect the polarization of light

Many invertebrates can discriminate 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.¹¹ 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.¹² 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.¹³ 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.¹⁴ 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.¹⁵

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.

¹¹Many other invertebrates, including spiders, crustaceans, and cephalopods, have this ability as well.

¹²Polarization vision can also be used for communication: Some insects create polarized light, presumably to identify to others of their species (Figure 50.3, page 628).

¹³Distance 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.

¹⁴If no blue sky was visible, either by design or on an overcast day, then communication failed.

¹⁵Section 47.7.2 (page 598).

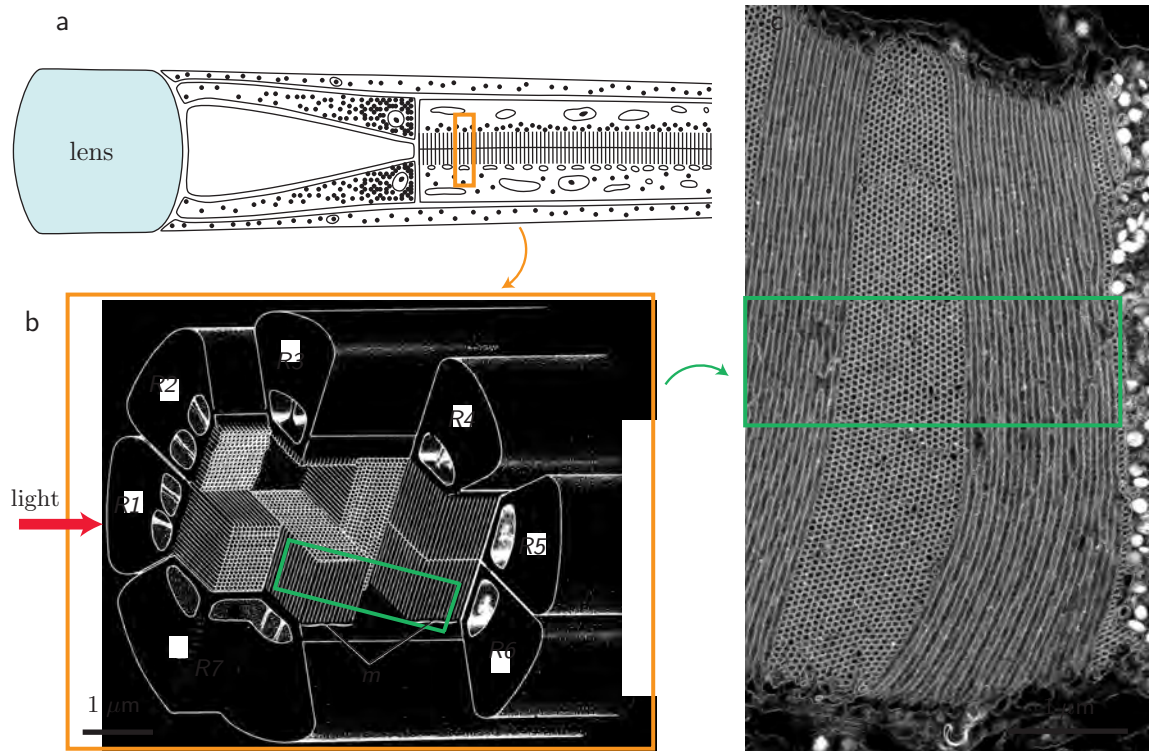


Figure 56.2: [Sketches; electron micrograph.] **Invertebrate photoreceptors.** (a) Cross-section through one facet (ommatidium) of a generic insect eye. Light passes through the lens (*left*) to the rhabdom (*right*). (b) Enlargement of the *orange box* in (a). Several long photoreceptor cells (*R1–R7*) 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.]

56.6.2 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 (retinal) 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 not.¹⁶

Insect and crustacean eyes have a different morphology (Figure 56.2). 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 optical fiber.¹⁷ 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**.¹⁸ 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 56.5.2 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.

56.6.3 Some transitions are far more probable than others

Section 56.5 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 56.2 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 56.5.2, 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 $(kr_e)^2$.

¹⁶Actually 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).

¹⁷One way to understand this guidance is via total internal reflection (Section 21.3.4, page 298).

¹⁸All the rhabdomeres together form the rhabdom in Figure 56.2.

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.

56.6.4 Lasers exploit a preference for emission into an already occupied state

Sections 56.5.2–56.5.3 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 56.26 (page 661) implies

$$\langle n+1 | Q^\dagger | n \rangle = \langle 0 | \frac{1}{\sqrt{(n+1)!}} Q^{n+1} (Q^\dagger)^{n+1} \frac{1}{\sqrt{n!}} | 0 \rangle = \langle 0 | \sqrt{\frac{(n+1)!}{n!}} | 0 \rangle = \sqrt{n+1}.$$

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. (56.41)

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**.

Light is preferentially emitted into states that already contain photons.

FURTHER READING

Semipopular:

Walmsley, 2015.

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.

Radiation; forbidden transitions: van der Straten & Metcalf, 2016.

Technical:

General: Berman & Malinovsky, 2011; Mandel & Wolf, 1995.

Defocused orientation imaging: Toprak et al., 2006; Böhmer & Enderlein, 2003; Bartko & Dickson, 1999a; Bartko & Dickson, 1999b.

APPENDIX A

Units and Dimensional Analysis

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 International*, 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.¹ 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**).² The system also creates related units via the prefixes giga ($=10^9$), mega ($=10^6$), kilo ($=10^3$), deci ($=10^{-1}$), 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 μ **g** is a microgram, and so on.

A symbol like μm^2 means $(\mu\text{m})^2 = 10^{-12} \text{m}^2$, not “ $\mu(\text{m}^2)$.”

¹One exception involves temperatures expressed using the Celsius and Fahrenheit scales, each of which differ from the absolute (Kelvin) scale by an offset.

²The standard abbreviation is **C**, but this risks confusion with the speed of light, a concentration or capacitance variable, or a generic constant.

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 \mathbb{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 \mathbb{L}^2 and \mathbb{L}^3 , respectively.
- The symbol \mathbb{M} denotes the dimension of mass. Its SI base unit is the kilogram.
- The symbol \mathbb{T} denotes the dimension of time. Its SI base unit is the second.
- The symbol \mathbb{Q} denotes the dimension of electric charge.³ Its SI base unit is the coulomb.
- Electric current has dimensions $\mathbb{Q}\mathbb{T}^{-1}$. The SI assigns it a standard unit coul/s , also called “ampere” and abbreviated A.
- Energy has dimensions $\mathbb{M}\mathbb{L}^2\mathbb{T}^{-2}$. The SI assigns it a standard unit $\text{kg m}^2/\text{s}^2$, also called “joule” and abbreviated J.
- Power (energy per unit time) has dimensions $\mathbb{M}\mathbb{L}^2\mathbb{T}^{-3}$. The SI assigns it a standard unit $\text{kg m}^2/\text{s}^3$, 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 $\mathbb{Q}\mathbb{T}^{-1}$. 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: Photon flux irradiance ($\mathbb{T}^{-1}\mathbb{L}^{-2}$) times area (\mathbb{L}^2) has dimensions appropriate for a rate (\mathbb{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 that both have the same dimensions (\mathbb{L}^3). We can automate unit conversions, and reduce errors, if we restate the conversion $1 \text{ US fluid ounce} \approx 29.6 \text{ cm}^3$ 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

³Some books use $\mathbb{I} = \mathbb{Q}/\mathbb{T}$, a “current” dimension, instead of \mathbb{Q} .

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.

Functions applied to dimensional quantities

If $x = 1 \text{ m}$, then we understand expressions like $2\pi x$ (with dimensions \mathbb{L}), and even x^3 (with dimensions \mathbb{L}^3). But what about $\sin(x)$ or $\log_{10} x$? These expressions are meaningless;⁴ more precisely, they don’t transform in any simple multiplicative way when we change units, unlike say $x/26$ or x^2 .

Additional SI units

circular frequency: One hertz (Hz) equals one complete cycle per second. Expressed as an angular frequency, $1 \text{ Hz} = 2\pi \text{ rad/s}$.

temperature: One kelvin (K) can be defined by saying that the atoms of an ideal monoatomic gas have mean kinetic energy $(3/2)k_{\text{B}}T$, where $k_{\text{B}} = 1.38 \cdot 10^{-23} \text{ J K}^{-1}$.

resistance and conductance: One ohm (Ω) equals one volt per ampere. One siemens is an inverse ohm: $1 \text{ S} = 1 \Omega^{-1}$.

electric potential: One volt (volt) equals 1 J/coul.

Traditional but non-SI units

mass: One dalton (also called “unified atomic mass unit,” and abbreviated u) is $1 \text{ Da} = 931.5 \text{ MeV}/c^2$.

time: One minute is 60 s, and so on.

length: One Ångstrom unit (Å) equals 0.1 nm.

volume: One liter (L) equals 10^{-3} m^3 . Thus, $1 \text{ mL} = 1 \text{ cm}^3$.

number density: A 1 M solution has a number density of $1 \text{ mole/L} = 1000 \text{ mole m}^{-3}$, where “mole” represents the number $\approx 6.02 \cdot 10^{23}$.

energy: An electron volt (eV) equals $e \times (1 \text{ volt}) = 1.60 \cdot 10^{-19} \text{ J} = 96 \text{ kJ/mole}$. Here e is the electric charge on a proton. An erg (erg) equals 10^{-7} J . Thus, $1 \text{ kcal mole}^{-1} = 0.043 \text{ eV} = 6.9 \cdot 10^{-21} \text{ J} = 6.9 \cdot 10^{-14} \text{ erg} = 4.2 \text{ kJ mole}^{-1}$.

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.

⁴One 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.

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 [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 [photons/ μm^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/ μm^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 $\log_{10}(x/(1 \text{ m}^2))$ or $\log_{10}(x \text{ [a.u.]})$, because the logarithm of a quantity with dimensions has no meaning.

A.3.1 Arbitrary units

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.3.2 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 \mathbb{L}) by r (with dimensions \mathbb{L}). Another clue is that if θ 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 2π 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 unit⁵ 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 2π .

Other traditional units of angle include **arcmin** = $(1/60)$ deg, and **arcsec** = $(1/60)$ arcmin. 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 $''$.

⁵We will use this abbreviation instead of the shorter $^\circ$, to avoid potential conflict with the temperature unit.

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.

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_{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.

APPENDIX B

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 in which it’s *harder to write down wrong formulas than it is to write correct formulas*.

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 332.

Operations

$\|\vec{b}\|$ Length of a real 3-vector, $=\sqrt{\vec{b}\cdot\vec{b}}$. For a complex vector, it means $\sqrt{\vec{b}^*\cdot\vec{b}}$.

z^* Complex conjugate of a complex number z .

$|z|$ Absolute value of a complex number, $=\sqrt{z^*z}$.

$\|\underline{X}\|^2$ Invariant norm-squared of a 4-vector.

∇^2 Laplace operator.

\square d’Alembert operator. (Some books write \square^2 instead of \square , to parallel the symbol ∇^2 .)

$\underline{T}\star$ Hodge dual operation (Section 15.9’b, page 226 and Section 34.9’, page 470).

$\vec{a}\cdot\vec{b}$ scalar product of two 3-vectors, itself a scalar.

$\vec{a}\cdot\vec{b}$ cross (vector) product of two 3-vectors, itself a (pseudo)vector.

$\vec{a}\otimes\vec{b}$ dyad (tensor) product of two vectors, itself a rank-2 tensor. (Other authors call it “outer product.”) It’s a special case of “tensor product.” (Some books omit the symbol \otimes and just write $\vec{a}\vec{b}$.)

¹A mathematician might therefore say “rank-1 tensor” wherever this book says “vector.”

$\vec{T}^{[S]}$ and $\vec{T}^{[A]}$ symmetric and antisymmetric parts of a rank-2 tensor (Equation 32.8, page 425).

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 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 $f(t) = \frac{1}{2}\bar{f}e^{-i\omega t} + \text{c.c.}$

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 ε_{ijk} .

The components of a vector or tensor in some coordinate system are always denoted with subscripts.² 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, φ , 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 $i, j, k \dots$ indices for components with respect to the original coordinate system and $a, b, c \dots$ for the transformed system. Sometimes \vec{v}_x is used as a synonym for \vec{v}_1 , and so on for \vec{v}_y and \vec{v}_z .

²To describe general tensors in curvilinear coordinates or on curved space, one must distinguish up-from down-indices, but this book will never do that.

$\vec{\nabla}$ Spatial gradient operator. Its cartesian components $\vec{\nabla}_i$ are the partial derivatives $\partial/\partial\vec{r}_i$.

$\vec{\mathbb{1}}$ Identity tensor (identity matrix regarded as a 3-tensor), also called “unit tensor”. Its cartesian components $\vec{\mathbb{1}}_{ij}$ 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, r indicates the length of \vec{r} . However, d^3r denotes $dx dy dz$ (which is not a vector).

A differential element of surface has area denoted generically by $d^2\Sigma$, or $d^2r = dx dy$ if specific 2D cartesian coordinates are used. When multiplied by a perpendicular unit vector, it becomes the vector $d^2\vec{\Sigma}$. 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 vectors³ \hat{x} , \hat{y} , and \hat{z} . These are constant unit vectors, but the radial unit vector $\hat{r} = \vec{r}/r$ 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, $\vec{r}_{(\ell)}$ is the position of particle ℓ ; its x component is then $\vec{r}_{(\ell)1}$ 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 49.7, page 615).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 α , shorthand for $\alpha\vec{\mathbb{1}}$.

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,” rotation. We may use the same symbol for each set to emphasize the correspondence, but distinguish the modified ones with a tilde: $\vec{\tilde{V}}$ in place of \vec{V} , or even $\vec{\tilde{r}}_{(\ell)}$ in place of $\vec{r}_{(\ell)}$.

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,

³Some books use the symbols \hat{i} , \hat{j} , \hat{k} , or simply \mathbf{i} , \mathbf{j} , \mathbf{k} , to represent the unit vectors that this book calls \hat{x} , \hat{y} , \hat{z} .

⁴This elaborate notation is not needed for a scalar quantity like index of refraction: n_{\pm} unambiguously indicates two cases.

which we will write as \vec{V}_i , $i = 1, 2, 3$ and \vec{V}'_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.

Occasionally, prime will instead be used to mean a derivative with respect to a spatial coordinate.

Similar remarks apply to higher-rank 3-tensors.

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 \underline{p}^μ . However, d^4X denotes $cdtdxdydz$ (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 μ, ν, λ, \dots indices for components with respect to the original coordinate system and $\alpha, \beta, \gamma, \dots$ 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 \underline{p}^1 or \vec{p}_1 .


The usage of tilde (active) and prime (passive) is the same as for three-dimensional objects.

∂ Spacetime gradient operator [dimension \mathbb{L}^{-1}].

$\binom{p}{q}$ Denotes the rank of a tensor with p upper and q lower indices.

\underline{g} Denotes the tensor whose components in any inertial coordinate system form the matrix $\text{diag}(-1, +1, +1, +1)$.

Spinors

 See Section 34.11' (page 471).

Matrices

Matrices are set in sans-serif type, M . They are arrays of numbers that do not necessarily transform in the specific manner of tensors upon coordinate change.

$\mathbf{1}$ Unit matrix.

S 3D rotation matrix.

Λ 4D Lorentz transformation matrix.

M^t Matrix transpose (exchange rows and columns). If the transpose equals the inverse, then M is called an orthogonal matrix.

M^* Complex conjugate each entry. If it equals M , then the matrix is real.

M^\dagger Hermitian conjugate, same as M^{*t} . If the hermitian conjugate equals the inverse, then M is called a unitary matrix.

Relations

\sim Has the same dimensions as.

\approx Is approximately equal to.

Miscellaneous

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 UNITS

See Chapter 16.

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.

\vec{A} Three-dimensional magnetic vector potential.

\underline{A} Four-vector potential.

b Generic name for a constant. \bar{b} , generic name for the amplitude of a sinusoidally-varying quantity.

B_{ij} Shape operator for a 2D surface in 3-space.

\vec{B} Magnetic induction (often called “magnetic field”) (a pseudovector); \check{B} , modified form, $= c\vec{B}$ [same dimensions as electric field, MLT^{-2}Q].

- c Speed of light in vacuum. c_s , speed of vibrations in a medium, e.g. a spring.
- c_e, c_{ion} , etc. Number density of electrons, ions, etc. [dimensions 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 d'Alembert operator.
- \vec{D} Electric displacement (analog of $\epsilon_0 \vec{E}$ in a medium).
- $\vec{\mathcal{D}}_E$ Electric dipole moment. $\vec{\mathcal{D}}_E$, its quantum version.
- $\vec{\mathcal{D}}_M$ Magnetic dipole moment (a pseudovector); $\check{\mathcal{D}}_M = \vec{\mathcal{D}}_M/c$, modified form (with same units as electric dipole moment).
- $\hat{e}_{(i)}$ Basis of three mutually perpendicular, unit 3-vectors defined by a cartesian coordinate system.
- e Charge on a proton.
- \vec{E} Electric field [dimensions MLT^{-2}Q].
- \mathcal{E} Energy, usually the relativistic (correct) form. $\check{\mathcal{E}}$, 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 L] (Section 39.5.2, page 516).
- F Linear tension, for example in a spring or along a 1D interface [dimensions of force] (Chapters 7, 27).
- \mathcal{F} Helmholtz free energy.
- \underline{F} Faraday 4-tensor.
- g Conductance per area.
- G Conductance.
- G Gauss curvature of a surface in space (Chapter 7).
- G_N Newton gravitation constant.
- \underline{g} Metric 4-tensor. In special relativity, this is a rank- $\binom{0}{2}$ tensor whose 16 components in any E-inertial coordinate system, $g_{\mu\nu}$, are always the same numerical constants. The same letter \underline{g} can also be used to refer to the dual metric tensor, a rank- $\binom{2}{0}$ tensor whose 16 components in any E-inertial coordinate system, $\underline{g}^{\mu\nu}$, are the *same* numerical constants as those of $\underline{g}_{\mu\nu}$. The notation is unambiguous because applying the index-raising operation to the first version does yield the second one.
- \vec{h} Displacement (position) of an object relative to the origin of coordinates or other reference point. h , generic symbol for a distance.
- H Mean curvature of a surface in space (Chapter 7).
- \vec{H} Magnetic intensity (analogous to \vec{B}/μ_0 but includes a medium) (a pseudovector).
- H Hamiltonian operator (Chapter 56) [dimensions ML^2T^{-2}].
- I Electric current [dimensions QT^{-1}]. I_x axial current in a cable. I_r , radial (“leak”) current in a cable.

- J Linear density of a line current source (Section 8.7.2, page 118) [dimensions $\mathbb{Q}\mathbb{T}^{-1}\mathbb{L}^{-1}$].
- \overleftrightarrow{J} Moment of inertia tensor of a rigid body.
- \vec{j} Electric charge flux [dimensions $\mathbb{Q}\mathbb{L}^{-2}\mathbb{T}^{-1}$]; $j^{(1D)}$, one-dimensional version. \vec{j} , its quantum version.
- \vec{j}_E Flux of energy.
- j_{ion} Number flux of ions of some species [dimensions $\mathbb{L}^{-2}\mathbb{T}^{-1}$].
- $\vec{j}^{[2D]}$ 2D charge flux in a surface (sometimes called “surface current density”); $\vec{j}_f^{[2D]}$, free surface charge flux.
- \overleftrightarrow{J} Matrix form of the Stokes parameters (Chapter 24) [dimensions $(\text{volt}/\text{m})^2$].
- \underline{J} Electric charge 4-flux (sometimes called “4-current”); \underline{j} , scalar analog sometimes used in simplified formulas.
- \mathcal{J} Source term for scalar wave equation (Equation 25.6, page 332).
- $\underline{\mathcal{J}}$ Generic conserved 4-flux arising from a continuous symmetry (Equation 40.14, page 531).
- 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 \underline{p}^0 .
- \overleftrightarrow{K} Hooke-law spring constant tensor.
- $\underline{K}^{\mu\nu}_{\lambda\sigma}$ Susceptibility operator (Section 49.6'b, page 620).
- ℓ 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.
- ℓ_B Bjerrum length (Equation 10.29).
- $\vec{\ell}$ Parametric representation of a generic curve in space; $d\vec{\ell}$, small element.
- L Inductance.
- \vec{L} Angular momentum (a pseudovector).
- \mathcal{L} Lagrangian density (Section 40.2, page 525–40.3).
- m Mass [dimensions \mathbb{M}]; m_e , mass of electron.
- m Generic 3-space index.
- \vec{M} Volume density of magnetic dipole moment (a pseudovector); $\check{M} = \vec{M}/c$, modified form (same units as \vec{P}).
- n refractive index.
- $\underline{M}^{\mu\nu\lambda}$ Angular momentum flux tensor (Section 35.5, page 483).
- 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.
- \vec{p} A particle’s 3-momentum, usually the relativistic (correct) form. \vec{p} , specifically the relativistic momentum when it is necessary to distinguish it from the newtonian quantity \vec{p}^N .
- \underline{p} A particle’s 4-momentum.
- \vec{p}_e electron momentum operator (Section 56.5.2, page 663) [dimensions $\mathbb{M}\mathbb{L}\mathbb{T}^{-1}$].
- \vec{P} Volume density of electric dipole moment (“polarization density”).

- \vec{P} momentum of electromagnetic field (Equation 56.8, page 658) [dimensions MLT^{-1}]. \vec{P} , corresponding quantum operator.
- \mathcal{P} Power [dimensions ML^2T^{-3}].
- Prob Probability (a real, dimensionless quantity between 0 and 1). $\varphi(x)$, Probability density function for a continuous random variable x [dimensions match those of $1/x$].
- q Electric charge.
- Q, Q^\dagger lowering (destruction) and raising (creation) operators, respectively, for electromagnetic field (Equation 56.17, page 659) [dimensionless].
- \vec{Q}_E Electric quadrupole 3-tensor. \vec{Q}_M Magnetic quadrupole 3-tensor.
- r Radial coordinate in spherical polar system (distance from origin).
- r_c Classical electron radius.
- \vec{r} Three-dimensional position vector, with cartesian components $\vec{r}_i = (x, y, z)^t$. Sometimes specifically the field point (observer location); then \vec{r}_* denotes source point.
- \vec{r}_e Electron position [dimensions \mathbb{L}]; \vec{r}_e , corresponding quantum operator (Equation 56.35, page 664).
- \vec{R} Displacement between source point and field point; R_{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.
- s Arc length parameter along a curve in 3-space.
- s_i Individual Stokes parameters (Chapter 24) [dimensions $(\text{volt/m})^2$].
- S Action functional (Section 40.2, page 525–40.3).
- S A 3D rotation, or the 3×3 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_* 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); T , temperature.
- \vec{T} Momentum flux 3-tensor (called “stress tensor” in some books).
- \underline{T} energy–momentum flux tensor (called “stress-energy tensor” in some books).
- T generator of a rotation (Equation 3.12, page 44).
- u, v Light-cone coordinates.
- u Displacement of a continuous spring.
- U Potential energy of a particle.
- \underline{U} Four-velocity. Its three spatial components are not equal to the components of ordinary velocity \vec{v} .
- v 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). v_* , velocity of a Galilean or Lorentz boost. v_m , velocity of a

- material medium that supports waves (spring, water, æther, ...). v depolarization ($\Delta\psi$ shifted by ψ^0); v_1 and v_2 , special fixed-point values (Figure 12.4); $\bar{v}(t)$, depolarization waveform of a traveling wave; \bar{v} , dimensionless rescaled form.
- V A region in 3-space, or its volume; ∂V , the boundary of V , that is a closed surface. An area element $d^2\vec{\Sigma}$ of that surface is conventionally taken to point outward.
- w Generic length variable, for example, thickness of a layer.
- x, y, z Right-handed cartesian coordinates of 3-space, or spatial components of a right-handed E-inertial coordinate system on spacetime.
- \underline{X} Four-vector coordinates of an event. Sometimes specifically the field (observation) point; then \underline{X}_* is the source event.

Greek alphabet

- α Electric polarizability of a molecule or other small object; α_m , magnetic polarizability. $\vec{\alpha}$, polarizability tensor of an anisotropic object.
- β Cross-polarizability of a single chiral molecule.
- $\vec{\beta}$ Velocity of a particle divided by c .
- γ Abbreviation for $1/\sqrt{1-\beta^2}$ (Section 30.3.1, page 389).
- γ Another generic quantity name.
- $\underline{\Gamma}(\xi)$ 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.
- $\vec{\vec{\Gamma}}$ Alternate representation of the magnetic dipole moment as an antisymmetric 3-tensor of rank 2.
- $\delta^{(n)}$ Product of n Dirac delta functions [dimensions of argument to power $-n$].
- ϵ Dielectric permittivity of a medium [dimensions $\text{Q}^2\text{T}^2\text{M}^{-1}\text{L}^{-3}$]; ϵ_0 , permittivity of vacuum. The dimensionless ratio ϵ/ϵ_0 is called the “dielectric constant,” but we don’t assign any symbol to it.
- ϵ_{ijk} Components of the 3D Levi-Civita tensor in a particular cartesian coordinate system. In a right-handed system, $\epsilon_{123} = +1$. $\underline{\epsilon}_{\mu\nu\kappa\lambda}$, components of the 4D Levi-Civita tensor in a particular E-inertial coordinate system. In a right-handed system, $\underline{\epsilon}_{0123} = +1$.
- ϵ_{multi} Multipole parameter (Equation 43.8) [dimensionless].
- $\epsilon_{\alpha\beta}$, $\tilde{\epsilon}^{\alpha\beta}$, $\epsilon_{\dot{\alpha}\dot{\beta}}$, and $\tilde{\epsilon}^{\dot{\alpha}\dot{\beta}}$, spinor metrics (Equations 34.27 and 34.30, page 474).
- $\vec{\zeta}$ Polarization 3-vector for a plane EM wave; $\hat{\zeta}_{(1)}$, $\hat{\zeta}_{(2)}$, linear polarization basis (real); $\hat{\zeta}_{(+)}$, $\hat{\zeta}_{(-)}$, circular polarization basis (complex). $\hat{\zeta}_{(\alpha;\vec{k})}$ basis of unit polarization vectors ($\alpha = 1, 2$) for plane waves traveling along \vec{k} (Equation 56.2, page 656) [dimensionless].
- $\underline{\zeta}$ Polarization 4-vector.
- η Viscous drag coefficient for a particle in fluid.
- η Bulk cross-polarizability of a chiral material.
- η_i integers specifying a mode in a cavity (Section 56.2, page 656) [dimensionless].

- θ Polar angle in spherical polar coordinates [dimensionless].
- ϑ Angle between an incoming wave's linear polarization and the line of sight to an observer.
- ϑ Velocity of neural action potential.
- Θ Step function [dimensionless].
- κ Electric conductivity of a medium; κ , elastic stretch modulus of a continuous spring; κ , curvature of a curve in a plane (Chapter 7).
- λ Wavelength of a plane or spherical wave.
- λ_D Debye length.
- λ_{cable} Space constant of a nerve axon or other cable.
- Λ A Lorentz transformation linking two E-inertial coordinate systems, or the 4×4 matrix representing it; Λ^μ_ν , its explicit components.
- μ Magnetic permeability of a medium [dimensions MLQ^{-2}]; μ_0 , permeability of vacuum.
- ν Circular frequency of a sinusoidally varying quantity (cycles per unit time) [dimensions T^{-1}].
- ξ 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.
- Ξ Gauge-transformation parameter.
- ρ Radial coordinate in cylindrical coordinate system.
- ρ Generic symbol for volume density of a quantity that has dimensions; ρ_q , electric charge density [dimensions QL^{-3}]; ρ_E , energy density; ρ_m , mass density.
- $\rho_q^{(1D)}$, linear electric charge density (coul/m); $\rho_E^{(1D)}$, linear energy density; $\rho_m^{(1D)}$, linear mass density (kg/m).
- σ Generic symbol for areal density of a scalar quantity; σ_q , surface charge density; σ_f , free surface charge density; σ_b , bound surface charge density.
- σ Scattering cross section.
- Σ A 2D surface, or its area; $d\vec{\Sigma}$, 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 Σ , that is, a closed curve with a direction chosen by applying the right-hand rule to the chosen perpendicular.
- τ 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.
- τ_{cable} Time constant of a nerve axon or other cable.
- Υ Rapidity parameter of a Lorentz boost.
- φ Azimuthal angle in either cylindrical or spherical polar coordinate system [dimensionless].
- ϕ Phase shift of one sine function relative to another.

- ϕ_N Newtonian gravitational potential.
- $\Phi_{\vec{k},\omega}$ The complex function $e^{i(\vec{k}\cdot\vec{r}-\omega t)}$ (dimensionless).
- Φ_B Integral of $\vec{B} \cdot d\vec{\Sigma}$ over an area. $\check{\Phi}_B = \Phi_B/c$, modified version.
- χ_e Dielectric susceptibility (polarizability of an isotropic medium); χ_m , magnetic susceptibility (polarizability of an isotropic medium); $\check{\chi}_m$, modified form. For anisotropic media, these are replaced by tensors.
- ψ Scalar potential field, also called electric potential [dimensions $\text{ML}^2\text{T}^{-2}\text{Q}^{-1}$]. In electrostatics, also called the electrostatic potential. $\bar{\psi}$, its dimensionless form (in static or quasi-static situations), $\bar{\psi}$, amplitude of a potential varying sinusoidally in time. $\psi^{[p]}$, standard 2^p -pole potentials. ψ_{in} , potential inside a neuron; ψ_{out} , potential outside (often taken to be zero). ψ^{Nernst} , Nernst potential; ψ^0 , quasisteady resting potential; v , membrane potential relative to ψ^0 .
- ω Angular frequency (radians per unit time).
- ω_p Plasma frequency.
- $\vec{\omega}$ Angular frequency of rigid body rotation, with direction corresponding to its axis of rotation via the right-hand rule (a pseudovector).
- $\vec{\omega}$ Alternate representation of \vec{B} as an antisymmetric, 3-tensor of rank 2.
- Ω Solid angle (sometimes called angular area).
-
- \vec{D}_E electric dipole moment operator (Section 56.5.2, page 663) [dimensions QL]. D_e , the length of its matrix element (the “transition dipole”).
- \hbar reduced Planck constant [dimensions ML^2T^{-1}].
- H Hamiltonian operator (Chapter 4) [dimensions ML^2T^{-2}].
- $n_{\vec{k},\alpha}$ photon occupation number (Equation 56.26, page 661) [dimensionless].
- \vec{p}_e electron momentum operator (Section 56.5.2, page 663) [dimensions MLT^{-1}].
- \vec{P} momentum of electromagnetic field (Equation 56.8, page 658) [dimensions MLT^{-1}]. \vec{P} , corresponding quantum operator.
- Q, Q^\dagger lowering (destruction) and raising (creation) operators, respectively, for electromagnetic field (Equation 56.17, page 659) [dimensionless].
- r_F Förster radius (Section 4.3.2, page 60) [dimensions L].
- \vec{r}_e electron position [dimensions L]; \vec{r}_e , corresponding quantum operator (Equation 56.35, page 664).
-
- Δ amount by which some quantity changes. Usually used as a prefix: Δx denotes a small change in x .
- $\hat{\zeta}_{(\alpha;\vec{k})}$ basis of unit polarization vectors ($\alpha = 1, 2$) for plane waves traveling along \vec{k} (Equation 56.2, page 656) [dimensionless].
- η_i integers specifying a mode in a cavity (Section 56.2, page 656) [dimensionless].
- $|\Phi\rangle$, vector in quantum-mechanical state space.

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

For salt solution at concentration 100 mM, $\kappa \approx 0.1 \Omega^{-1} \text{m}^{-1}$.
Liquid water: $\epsilon \approx 80\epsilon_0$ at low frequency.

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: $e^2/(4\pi\epsilon_0) \approx 1.44 \text{eV nm} = 1.44 \text{MeV fm}$.

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, $k_B \approx 1.38 \cdot 10^{-23} \text{J K}^{-1}$. Typical thermal energy at room temperature $k_B T_r \approx 4.1 \text{pN nm} \approx 4.1 \cdot 10^{-21} \text{J} \approx 2.5 \text{kJ mole}^{-1} \approx 0.59 \text{kcal mole}^{-1} \approx 0.025 \text{eV}$.

Permittivity of vacuum, $\epsilon_0 \approx 8.85 \cdot 10^{-12} \text{coul}^2 \text{N}^{-1} \text{m}^{-2}$. Permeability of vacuum, $\mu_0 \approx 4\pi \cdot 10^{-7} \text{m kg coul}^{-2}$.

C.2 OPTICS

C.2.1 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: $n_w \approx 1.33$.

Glass: 1.5–1.7. This book uses the illustrative value 1.52.

C.2.2 Miscellaneous

Earth's magnetic field strength at surface, approx $5 \cdot 10^{-5} \text{T}$.

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.

APPENDIX D

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. 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/jwkvam/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 may be able to do a video screen capture of your animation as it runs in Python.
 - You can directly generate an `mp4` file, which in turn is 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.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 is to create a folder containing many still images (individual video frames). You can then:
 - 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%05d.png -pix_fmt yuv420p mymovie.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.

¹See Pine, 2019, Kinder & Nelson, 2021, Hill, 2020.

²Users of Google Colab will find that `ffmpeg` or something equivalent is automatically available.

Bibliography

Looking through this volume... was like roaming through an exquisite palace while its inhabitants slept.

— Orhan Pamuk

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.

- AKIYAMA, K, & OTHERS (EVENT HORIZON TELESCOPE COLLABORATION). 2021a. First M87 event horizon telescope results. VII. Polarization of the ring. *Astrophysical J. Lett.*, **910**(1), L12.
- AKIYAMA, K, & OTHERS (EVENT HORIZON TELESCOPE COLLABORATION). 2021b. First M87 event horizon telescope results. VIII. Magnetic field structure near the event horizon. *Astrophysical J. Lett.*, **910**(1), L13.
- ALLEN, L, PADGETT, M J, & BABIKER, M. 1999. The orbital angular momentum of light. *Pages 294–272 of: WOLF, E. (Ed.), Progress in Optics XXXIX*. Elsevier.
- ALVÄGER, T, FARLEY, FJM, KJELLMAN, J, & WALLIN, L. 1964. Test of the second postulate of special relativity in the GeV region. *Physics Letters*, **12**(3), 260–262.
- ANDREWS, D L, & BRADSHAW, D S. 2022. *Optical nanomanipulation*. 2nd ed. IOP Publishing.
- ARFKEN, G B, WEBER, H J, & HARRIS, F E. 2013. *Mathematical methods for physicists: A comprehensive guide*. 7th ed. Amsterdam: Elsevier.
- ARIYARATNE, A, & ZOCCHI, G. 2016. Toward a minimal artificial axon. *J Phys. Chem. B*, **120**(26), 6255–6263.
- BACKLUND, M P, ARBABI, A, PETROV, P N, ARBABI, E, SAURABH, S, FARAON, A, & MOERNER, W E. 2016. Removing orientation-induced localization biases in single-molecule microscopy using a broadband metasurface mask. *Nat. Photon.*, **10**(7), 459–463.
- BAGOTSKII, V S. 2006. *Fundamentals of electrochemistry*. Hoboken NJ: Wiley Interscience.
- BAHAR, I, JERNIGAN, R L, & DILL, K A. 2017. *Protein actions: Principles and modeling*. New York: Garland Science.
- BAKER, B B, & COPSON, E T. 1950. *The mathematical theory of Huygens' principle*. Oxford UK: Oxford Univ. Press.
- BAKER, P F, HODGKIN, A L, & SHAW, T I. 1962. Replacement of the axoplasm of giant axon fibres with artificial solutions. *J. Physiol. Lond.*, **164**, 330–354.
- BAND, Y B. 2006. *Light and matter: Electromagnetism, optics, spectroscopy and lasers*. Chichester UK: John Wiley.
- BARGMANN, V. 1954. On unitary ray representations of continuous groups. *Ann. Mathematics*, **59**(1), 1.
- BARTKO, A, & DICKSON, R. 1999a. Imaging three-dimensional single molecule orientations. *J. Phys. Chem. B*, **103**, 11237–11241.
- BARTKO, A, & DICKSON, R. 1999b. Three-dimensional orientations of polymer-bound single molecules. *J. Phys. Chem. B*, **103**, 3053–3056.

- BASANO, L, & OTTONELLO, P. 2005. Demonstration experiments on nondiffracting beams generated by thermal light. *Am. J. Phys.*, **73**(9), 826–830.
- BAYLOR, S M. 2020. *Computational cell physiology: With examples In Python*. amazon.com: Kindle Direct Publishing.
- BECHHOEFER, J, & WILSON, S. 2002. Faster, cheaper, safer optical tweezers for the undergraduate laboratory. *Am. J. Phys.*, **70**(4), 393–400.
- BENEDEK, G B, & VILLARS, F M H. 2000. *Physics with illustrative examples from medicine and biology*. 2nd ed. Vol. 3. New York: AIP Press.
- BERG, H C, & TURNER, L. 1993. Torque generated by the flagellar motor of Escherichia coli. *Biophysical Journal*, **65**(5), 2201–2216.
- BERMAN, P R, & MALINOVSKY, V S. 2011. *Principles of laser spectroscopy and quantum optics*. Princeton NJ: Princeton Univ. Press.
- BERRY, M V. 1981. Singularities in waves and rays. *Pages 453–543 of: BALIAN, R D, KLEMAN, M, & POIRIER, J-P (Eds.), Physics of defects*. Amsterdam: North-Holland.
- BERRY, M V. 2015. Nature's optics and our understanding of light. *Contemp. Physics*, **56**(1), 2–16.
- BERRY, M V. 2017. *A half-century of physical asymptotics and other diversions: Selected works by Michael Berry*. World Scientific.
- BERRY, M V, & GEIM, A K. 1997. Of flying frogs and levitrons. *Eur. J. Phys.*, **18**(4), 307–313.
- BETH, R A. 1935. Direct detection of the angular momentum of light. *Phys. Rev.*, **48**, 471–471.
- BETH, R A. 1936. Mechanical detection and measurement of the angular momentum of light. *Phys. Rev.*, **50**, 115–125.
- BODANIS, D. 2005. *Electric universe: How electricity switched on the modern world*. New York: Three Rivers Press.
- BÖHMER, M, & ENDERLEIN, J. 2003. Orientation imaging of single molecules by wide-field epifluorescence microscopy. *J. Opt. Soc. Am. B*, **20**, 554–559.
- BOHREN, C F, & CLOTHIAUX, E E. 2006. *Fundamentals of atmospheric radiation*. Weinheim: Wiley-VCH.
- BOHREN, C F, HUFFMAN, D R, & CLOTHIAUX, E E. 2015. *Absorption and scattering of light by small particles*. 2nd ed. New York: Wiley-VCH.
- BORK, A M. 2005. Maxwell, displacement current, and symmetry. *Am. J. Phys.*, **31**(11), 854–859.
- BORN, M, & WOLF, E. 1999. *Principles of optics*. 7th ed. Cambridge UK: Cambridge Univ. Press.
- BRECHER, K. 1977. Is the speed of light independent of the velocity of the source? *Physical Review Letters*, **39**(17), 1051–1054.
- BRESSLOFF, P C. 2014. *Waves in neural media*. New York: Springer.
- BUCHWALD, J Z. 1985. *From Maxwell to microphysics: Aspects of electromagnetic theory in the last quarter of the nineteenth century*. Chicago: Univ. Chicago Press.
- BURKE, W L. 1985. *Applied differential geometry*. Cambridge UK: Cambridge Univ. Press.
- BUTT, H-J, & KAPPL, M. 2018. *Surface and interfacial forces*. 2d ed. Wiley-VCH.
- CAGNET, M, FRANCON, M, & THRIERR, J C. 1962. *Atlas optischer Erscheinungen. Atlas de phénomènes d'optique. Atlas of optical phenomena*. Berlin: Springer.
- CAHILL, K. 2013. *Physical mathematics*. Cambridge MA: Cambridge Univ. Press.
- CAI, J, WANG, L, & WU, P. 2007. Oxygen enrichment from air by using the interception effect of gradient magnetic field on oxygen molecules. *Elsevier*.
- CAMPBELL, L, & GARNETT, W. 1882. *The life of James Clerk Maxwell with a selection from his correspondence and occasional writings and a sketch of his contributions to science*. London: Macmillan. <https://www.sonnetsoftware.com/resources/maxwell-bio.html>.
- CHALMERS, A F. 1975. Maxwell and the displacement current. *Phys. Educ.*, **10**, 45–49.

- CLEGG, R M. 2006. The history of FRET. *Pages 1–45 of: GEDDES, C, & LAKOWICZ, J (Eds.), Reviews in fluorescence 2006.* Reviews in Fluorescence, Vol. 3. New York: Springer US.
- COHEN, I B. 1990. *Benjamin Franklin's science.* Cambridge MA: Harvard Univ. Press.
- COLE, K S. 1972. *Membranes, ions, and impulses: A chapter of classical biophysics.* Berkeley CA: Univ. California Press.
- COLEMAN, S. 2019. *Quantum field theory: Lectures of Sidney Coleman.* Singapore: World Scientific. Eds. Chen, B G-G, D Derbes, D Griffiths, B Hill, R Sohn, and Y-S Ting.
- COOPERSMITH, J. 2017. *The lazy Universe: An introduction to the principle of least action.* Oxford UK: Oxford Univ. Press.
- COWLEY, A C, FULLER, N L, RAND, R P, & PARSEGIAN, V A. 1978. Measurements of repulsive forces between charged phospholipid bilayers. *Biochemistry*, **17**, 3163–3168.
- CRAIG, D P, & THIRUNAMACHANDRAN, T. 1998. *Molecular quantum electrodynamics: An introduction to radiation-molecule interactions.* New York: Dover Publications.
- CURTIS, J, & GRIER, D. 2003. Structure of optical vortices. *Phys. Rev. Lett.*, **90**(13), 133901.
- CUSHING, J T. 1981. Electromagnetic mass, relativity, and the Kaufmann experiments. *American Journal of Physics*, **49**(12), 1133–1149.
- DARRIGOL, O. 2022. *Relativity principles and theories from Galileo to Einstein.* Oxford UK: Oxford Univ. Press.
- DAVIDSON, P A. 2019. *An introduction to electrodynamics.* Oxford UK: Oxford Univ. Press.
- DE BROGLIE, L. 1923a. Ondes et quanta. *CR Acad. Sci. Paris.*
- DE BROGLIE, L. 1923b. Quanta de lumière, diffraction et interférences. *CR Acad. Sci. Paris.*
- DE VRIES, H. 1951. Rotatory power and other optical properties of certain liquid crystals. *Acta Crystallographica*, **4**(3), 219–226.
- DILL, K A, & BROMBERG, S. 2011. *Molecular driving forces: Statistical thermodynamics in biology, chemistry, physics, and nanoscience.* 2d ed. New York: Garland Science.
- DODELSON, S, & SCHMIDT, F. 2021. *Modern cosmology.* London: Academic Press.
- DREINER, H K, HABER, H E, & MARTIN, S P. 2010. Two-component spinor techniques and Feynman rules for quantum field theory and supersymmetry. *Physics Reports*, **494**(1-2), 1–196.
- DUBROVIN, B A, FOMENKO, A T, & NOVIKOV, S P. 1992. *Modern geometry—methods and applications.* 2nd ed. Vol. 1. New York: Springer-Verlag.
- DURMUS, N G, TEKIN, H C, GUVEN, S, SRIDHAR, K, ARSLAN YILDIZ, A, CALIBASI, G, GHIRAN, I, DAVIS, R W, STEINMETZ, L M, & DEMIRCI, U. 2015. Magnetic levitation of single cells. *Proc. Natl. Acad. Sci. USA*, **112**(28), E3661–8.
- DURNIN, J. 1987. Exact solutions for nondiffracting beams 1. The scalar theory. *J. Opt. Soc. Am. A*, **4**(4), 651–654.
- DURNIN, J, MICELI, J J, & EBERLY, J H. 1986. Diffraction-free beams. *J. Opt. Soc. Am. A*, **3**(13), P128.
- DURNIN, J, MICELI, J, & EBERLY, J. 1987. Diffraction-free beams. *Phys. Rev. Lett.*, **58**(15), 1499–1501.
- DURNIN, J, EBERLY, J, & MICELI, J. 1988. Comparison of Bessel and Gaussian beams. *Opt. Lett.*, **13**, 79–80.
- EID, J, & OTHERS. 2009. Real-time DNA sequencing from single polymerase molecules. *Science*, **323**(5910), 133–138.
- EINSTEIN, A. 1998. On the electrodynamics of moving bodies. In: STACHEL, J (Ed.), *Einstein's miraculous year: Five papers that changed the face of physics.* Princeton NJ: Princeton Univ. Press. German original: Ann. Phys. **17** (1905)891–921.
- ELMORE, W C, & HEALD, M A. 1969. *Physics of waves.* New York: Dover Publications.
- FEYNMAN, R P, LEIGHTON, R, & SANDS, M. 2010a. *The Feynman lectures on physics.* New millennium ed. Vol. 2. New York: Basic Books. Free online: <http://www.feynmanlectures.caltech.edu/>.

- FEYNMAN, R P, LEIGHTON, R, & SANDS, M. 2010b. *The Feynman lectures on physics*. New millennium ed. Vol. 1. New York: Basic Books. Free online: <http://www.feynmanlectures.caltech.edu/>.
- FIZEAU, M. 1859. Sur les hypothèses relatives à l'éther lumineux, et sur une expérience qui paraît démontrer que le mouvement des corps change la vitesse avec laquelle la lumière se propage dans leur intérieur. *Ann. de Chim. et de Phys. 3e série*, **LVII**, 385–404.
- FLEISCH, D. 2012. *A student's guide to vectors and tensors*. Cambridge Univ. Press.
- FRANKLIN, B. 1941. *Experiments and observations on electricity*. Cambridge MA: Harvard Univ. Press.
- FRANKLIN, W S, & NICHOLS, E L. 1894. On the condition of the ether surrounding a moving body. *Phys. Rev. (Series I)*, **1**(6), 426–441.
- FRASER, A B. 1983a. Chasing rainbows. *Weatherwise*, **36**(6), 280–289.
- FRASER, A B. 1983b. Why can the supernumerary bows be seen in a rain shower? *J. Opt. Soc. Am.*, **73**(12), 1626–1628.
- FREEMAN, R, KING, J, & LAFYATIS, G. 2019. *Electromagnetic radiation*. Oxford UK: Oxford Univ. Press.
- GALISON, P. 2003. *Einstein's clocks, Poincaré's maps: Empires of time*. W.W. Norton.
- GALVEZ, E J. 2013. Vector beams in free space. *Chap. 3, pages 51–70 of: ANDREWS, D L, & BABIKER, M (Eds.), The angular momentum of light*. Cambridge UK: Cambridge Univ. Press.
- GAO, L, SHAO, L, CHEN, B-C, & BETZIG, E. 2014. 3D live fluorescence imaging of cellular dynamics using Bessel beam plane illumination microscopy. *Nat. Protoc.*, **9**(5), 1083–1101.
- GARG, A. 2012. *Classical electromagnetism in a nutshell*. Princeton NJ: Princeton Univ. Press.
- GHOSH, A, FAZAL, F M, & FISCHER, P. 2007. Circular differential double diffraction in chiral media. *Opt. Lett.*, **32**(13), 1836–1838.
- GILLESPIE, D T. 1998. The mathematics of Brownian motion and Johnson noise. *Am. J. Phys.*, **64**(3), 225–240.
- GINZBURG, V. L. 1989. *Applications of electrodynamics in theoretical physics and astrophysics*. Gordon and Breach.
- GODDI, C, & OTHERS (EVENT HORIZON TELESCOPE COLLABORATION). 2021. Polarimetric properties of event horizon telescope targets from ALMA. *Astrophysical J. Lett.*, **910**(1), L14.
- GOEDECKE, G H, WOOD, R C, & NACHMAN, P. 1999. Magnetic dipole orientation energetics. *Am. J. Phys.*, **67**, 45–51.
- GOODSELL, D S. 2016. *Atomic evidence: Seeing the molecular basis of life*. Springer.
- GÖTTE, J B, & BARNETT, S M. 2013. Light beams carrying orbital angular momentum. *Chap. 1, pages 1–30 of: ANDREWS, D L, & BABIKER, M (Eds.), The angular momentum of light*. Cambridge UK: Cambridge Univ. Press.
- GOULD, R J. 2006. *Electromagnetic processes*. Princeton NJ: Princeton Univ. Press.
- GRATIY, S L, HALNES, G, DENMAN, D, HAWRYLYCZ, M J, KOCH, C, EINEVOLL, G T, & ANASTASSIOU, C A. 2017. From Maxwell's equations to the theory of current-source density analysis. *Eur. J. Neurosci.*, **45**(8), 1013–1023.
- GRAY, N. 2022. *A student's guide to special relativity*. Cambridge UK: Cambridge Univ. Press.
- GRIER, D G. 2003. A revolution in optical manipulation. *Nature*, **424**(6950), 810–816.
- GRIFFITHS, D J, DERBES, D, & SOHN, R B (Eds.). 2022. *Sidney Coleman's lectures on relativity*. Cambridge UK: Cambridge.
- GRODZINSKY, A J. 2011. *Fields, forces, and flows in biological systems*. London UK: Garland Science.
- GRYNBERG, G, ASPECT, A, & FABRE, C. 2010. *Introduction to quantum optics*. Cambridge UK: Cambridge Univ. Press.
- GUTTAG, J V. 2021. *Introduction to computation and programming using Python: With application to computational modeling and understanding data*. 3rd ed. Cambridge MA: MIT Press.

- HA, T, ENDERLE, T, OGLETREE, D F, CHEMLA, D S, SELVIN, P R, & WEISS, S. 1996. Probing the interaction between two single molecules: Fluorescence resonance energy transfer between a single donor and a single acceptor. *Proc. Natl. Acad. Sci. USA*, **93**(13), 6264–6268.
- HAJDUK, M, ROTH, E, & GRANITE, E. 2013. Magnetic separation of oxygen from air. *30th Annual International Pittsburgh Coal Conference 2013, PCC 2013*, **5**, 3769–3777.
- HARMAN, P M. 1998. *The natural philosophy of James Clerk Maxwell*. Cambridge UK: Cambridge Univ. Press.
- HASEGAWA, A, & KODAMA, Y. 1995. *Solitons in optical communications*. Oxford UK: Oxford Univ. Press.
- HEALD, M A, & MARION, J B. 2012. *Classical electromagnetic radiation*. 3rd ed. New York: Dover Publications.
- HECHT, E. 2017. *Optics*. 5th ed. Boston MA: Pearson.
- HEDRICH, R, & NEHER, E. 2018. Venus Flytrap: How an Excitable, Carnivorous Plant Works. *Trends Plant Sci*, **23**(3), 220–234. Hedrich, Rainer Neher, Erwin eng Research Support, Non-U.S. Gov't Review England 2018/01/18 Trends Plant Sci. 2018 Mar;23(3):220-234. doi: 10.1016/j.tplants.2017.12.004. Epub 2018 Jan 11.
- HEHL, F W, & OBUKHOV, Y N. 2003. *Foundations of classical electrodynamics: Charge, flux, and metric*. Boston: Birkhäuser.
- HILL, C. 2020. *Learning scientific programming with Python*. 2nd ed. Cambridge UK: Cambridge Univ. Press. scipython.com/book2/.
- HOBBIE, R K, & ROTH, B J. 2015. *Intermediate physics for medicine and biology*. 5th ed. Cham CH: Springer International Publishing.
- HODGKIN, A. 1992. *Chance and design: Reminiscences of science in peace and war*. Cambridge UK: Cambridge Univ. Press.
- HODGKIN, A L, & KATZ, B. 1949. Effect of sodium ions on the electrical activity of the giant axon of the squid. *J. Physiol. Lond.*, **108**, 37–77.
- HORVÁTH, G (Ed.). 2014. *Polarized light and polarization vision in animal sciences*. 2nd ed. New York: Springer.
- HUBBARD, J H, & HUBBARD, B B. 2007. *Vector calculus, linear algebra, and differential forms: A unified approach*. 2nd ed. Ithaca NY: Matrix Editions.
- HULST, H C VAN DE. 1957. *Light scattering by small particles*. New York: Wiley.
- HUNT, B J. 1991. *The maxwellians*. Ithaca NY: Cornell Univ. Press.
- HWANG, L C, HOHLBEIN, J, HOLDEN, S J, & KAPANIDIS, A N. 2009. Single-molecule FRET: Methods and biological applications. *Chap. 5, pages 129–164 of: HINTERDORFER, P, & VAN OIJEN, A (Eds.), Handbook of single-molecule biophysics*. New York: Springer.
- IQBAL, A, ARSLAN, S, OKUMUS, B, WILSON, T J, GIRAUD, G, NORMAN, D G, HA, T, & LILLEY, D M J. 2008. Orientation dependence in fluorescent energy transfer between Cy3 and Cy5 terminally attached to double-stranded nucleic acids. *Proc. Natl. Acad. Sci. USA*, **105**(32), 11176–11181.
- ISRAELACHVILI, J N. 2011. *Intermolecular and surface forces*. 3rd ed. Burlington MA: Academic Press.
- IVANOV, D T, & NIKOLOV, S N. 2016. A new way to demonstrate the rainbow. *Physics Teach.*, **54**(8), 460–463.
- JACKSON, J D. 1999. *Classical electrodynamics*. 3rd ed. New York: John Wiley.
- JAGGER, W S, & SANDS, P J. 1999. A wide-angle gradient index optical model of the crystalline lens and eye of the octopus. *Vision Res.*, **39**(17), 2841–2852.
- JELLEY, J V. 1958. *Cerenkov radiation and its applications*.
- JI, Z, LIU, W, KRYLYUK, S, FAN, X, ZHANG, Z, PAN, A, FENG, L, DAVYDOV, A, & AGARWAL, R. 2020. Photocurrent detection of the orbital angular momentum of light. *Science*, **368**(6492), 763–767.
- JOANNOPOULOS, J D, MEADE, R D, & WINN, J N. 2008. *Photonic crystals: Molding the flow of light*. 2d ed. Princeton NJ: Princeton Univ. Press.
- JOHNSON, D A, LEATHERS, V L, MARTINEZ, A M, WALSH, D A, & FLETCHER, W H. 1993. Fluorescence resonance energy transfer within a heterochromatic cAMP-dependent protein kinase holoenzyme under equilibrium conditions: New insights into the conformational changes that result in cAMP-dependent activation. *Biochemistry*, **32**(25), 6402–6410.

- JOHNSON, G. 2008. *The ten most beautiful experiments*. New York: Alfred A. Knopf.
- JONES, P H, MARAGÒ, O M, & VOLPE, G. 2015. *Optical tweezers: Principles and applications*. Cambridge UK: Cambridge Univ. Press.
- JORGENSEN, T J. 2021. *Spark: The life of electricity and the electricity of life*. Princeton NJ: Princeton Univ. Press.
- KEENER, J, & SNEYD, J. 2009. *Mathematical physiology I: Cellular physiology*. 2d ed. New York: Springer.
- KENNEDY, R E. 2012. *A student's guide to Einstein's major papers*. Oxford UK: Oxford Univ. Press.
- KINDER, J M, & NELSON, P. 2021. *A student's guide to Python for physical modeling*. 2nd ed. Princeton NJ: Princeton Univ. Press.
- LA PORTA, A, & WANG, M D. 2004. Optical torque wrench: angular trapping, rotation, and torque detection of quartz microparticles. *Phys. Rev. Lett.*, **92**(19), 190801.
- LAHAYE, T, LABASTIE, P, & MATHEVET, R. 2012. Fizeau's "æther-drag" experiment in the undergraduate laboratory. *Am. J. Phys.*, **80**, 497–505.
- LAKSHMINARAYANAN, V, & ENOCH, J M. 2011. Biological waveguides. *Chap. 9 of: BASS, M, DECUSATIS, C M, ENOCH, J, LAKSHMINARAYANAN, V, LI, G, MACDONALD, C, MAHAJAN, V N, & VAN STRYLAND, E (Eds.), Handbook of optics*, Vol. 3. OSA/McGrawHill.
- LANDAU, L D, & LIFSHITZ, E M. 1977. *Quantum mechanics: Non-relativistic theory*. 3rd ed. New York: Pergamon Press.
- LANDAU, L D, & LIFSHITZ, E M. 1979. *The classical theory of fields*. 4th rev. ed. Oxford UK: Pergamon Press.
- LANDAU, L D, & LIFSHITZ, E M. 1981. *Mechanics*. 3rd ed. New York: Butterworth-Heinemann.
- LANDAU, L D, LIFSHITZ, E M, & PITAEVSKII, L P. 1984. *Electrodynamics of continuous media*. 2d ed. Oxford UK: Pergamon Press.
- LARSON, E. 2006. *Thunderstruck*. New York: Crown Publishers.
- LEE, N K, KAPANIDIS, A N, WANG, Y, MICHALET, X, MUKHOPADHYAY, J, EBRIGHT, R H, & WEISS, S. 2005. Accurate FRET measurements within single diffusing biomolecules using alternating-laser excitation. *Biophys. J.*, **88**(4), 2939–2953.
- LEE, SH, ROICHMAN, Y, & GRIER, DG. 2010. Optical solenoid beams. *Opt. Express*, **18**(7), 6988–6993.
- LEE, JR., R L, & FRASER, A B. 2001. *The rainbow bridge: Rainbows in art, myth, and science*. University Park PA and Bellingham WA: Pennsylvania State University Press and SPIE Press.
- LEONHARDT, U. 2010. *Essential quantum optics*. Cambridge UK: Cambridge Univ. Press.
- LERCHE, I. 1977. The Fizeau effect: Theory, experiment, and Zeeman's measurements. *Am. J. Phys.*, **45**, 1154–1163.
- LEVENE, M J, KORLACH, J, TURNER, S W, FOQUET, M, CRAIGHEAD, H G, & WEBB, W W. 2003. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, **299**(5607), 682–686.
- LEW, M D, & MOERNER, W E. 2014. Azimuthal polarization filtering for accurate, precise, and robust single-molecule localization microscopy. *Nano Lett.*, **14**(11), 6407–6413.
- LIFSHITZ, E M, & PITAEVSKIĬ, LP. 1981. *Physical kinetics*. Oxford UK: Pergamon.
- LIONNET, T, ALLEMAND, J-F, REVYAKIN, A, STRICK, T R, SALEH, O A, BENSIMON, D, & CROQUETTE, V. 2012a. Magnetic trap construction. *Cold Spring Harb. Protoc.*, **2012**(1), 133–138.
- LIONNET, T, ALLEMAND, J-F, REVYAKIN, A, STRICK, T R, SALEH, O A, BENSIMON, D, & CROQUETTE, V. 2012b. Single-molecule studies using magnetic traps. *Cold Spring Harb. Protoc.*, **2012**(1), 34–49.
- LIPSON, A, LIPSON, S G, & LIPSON, H. 2011. *Optical physics*. 4th ed. Cambridge UK: Cambridge Univ. Press.
- LIU, S, HUH, H, LEE, S H, & HUANG, F. 2020. Three-dimensional single-molecule localization microscopy in whole-cell and tissue specimens. *Annu. Rev. Biomed. Eng.*, **22**, 155–184.
- LONG, Y, WEI, H, LI, J, YAO, G, YU, B, NI, D, GIBSON, A L F, LAN, X, JIANG, Y, CAI, W, & WANG, X. 2018. Effective Wound Healing Enabled by Discrete Alternative Electric Fields from Wearable Nanogenerators. *ACS Nano*, **12**(12), 12533–12540. PMID: 30488695.

- LONGAIR, M S. 2020. *Theoretical concepts in physics*. 3rd ed. Cambridge UK: Cambridge Univ. Press.
- LOUDON, R. 2000. *The quantum theory of light*. 3d ed. Oxford UK: Oxford Univ. Press.
- LOUDON, R. 2003. Theory of the forces exerted by Laguerre-Gaussian light beams on dielectrics. *Phys. Rev. A*, **68**(1), 013806.
- LYNCH, D K, & LIVINGSTON, W. 2001. *Color and light in nature*. 2nd ed. Cambridge Univ. Press.
- LYNCH, D K, & SCHWARTZ, P. 1991. Rainbows and fogbows. *Appl. Opt.*, **30**(24), 3415–3420.
- MADAENI, S S, ENAYATI, E, & VATANPOUR, V. 2011. Separation of nitrogen and oxygen gases by polymeric membrane embedded with magnetic nano-particle. *Polymers for Advanced Technologies*, **22**(12), 2556–2563.
- MAHAJAN, S. 2014. *The art of insight in science and engineering: Mastering complexity*. Cambridge MA: MIT Press.
- MALMIVUO, J, & PLONSEY, R. 1995. *Bioelectromagnetism - Principles and Applications of Bioelectric and Biomagnetic Fields*. New York: Oxford University Press. <http://www.bem.fi/book/>.
- MANDEL, L, & WOLF, E. 1995. *Optical coherence and quantum optics*. Cambridge UK: Cambridge Univ. Press.
- MARTIN, A C, KASCHUBE, M, & WIESCHAUS, E F. 2009. Pulsed contractions of an actin–myosin network drive apical constriction. *Nature*, **457**(7228), 495–499.
- MARTINOT, Z E, AGUIRRE, J E, KOHN, S A, & WASHINGTON, I Q. 2018. Ionospheric Attenuation of Polarized Foregrounds in 21 cm Epoch of Reionization Measurements: A Demonstration for the HERA Experiment. *Astrophys. J.*, **869**(1).
- MAXWELL, J, & JENKIN, F. 1865. LXI. On the elementary relations between electrical measurements, part I: Introductory. *Philosophical Magazine Series 4*, **29**(198), 436–460.
- MCQUEEN, C A, ARLT, J, & DHOLAKIA, K. 1999. An experiment to study a “nondiffracting” light beam. *Am. J. Phys.*, **67**(10), 912–915.
- MELIA, F. 2001. *Electrodynamics*. Chicago IL: Univ. Chicago Press.
- MERMIN, N D. 2005. *It’s about time: Understanding Einstein’s relativity*. Princeton NJ: Princeton Univ. Press.
- MICHELSON, A, & MORLEY, E. 1886. Influence of motion of the medium on the velocity of light. *American Journal of Science*, **31**, 377–386.
- MICHELSON, A A, & MORLEY, E W. 1887. On the relative motion of the Earth and the luminiferous ether. *Amer. J. Sci.*, **XXXIV**(Nov.), 333–341.
- MICHILLI, D, & OTHERS. 2018. An extreme magneto-ionic environment associated with the fast radio burst source FRB 121102. *Nature*, **553**(7687), 182–185.
- MILONNI, P W, & EBERLY, J H. 2010. *Laser physics*. New York: Wiley.
- MINNAERT, M G J. 1993. *Light and color in the outdoors*. New York: Springer.
- MIRICA, K A, PHILLIPS, S T, MACE, C R, & WHITESIDES, G M. 2010. Magnetic levitation in the analysis of foods and water. *J. Agric. Food Chem.*, **58**(11), 6565–6569.
- MORSE, P M, & FESHBACK, H. 1953. *Methods of theoretical physics*. New York: McGraw-Hill.
- MÜLLER, H, HERRMANN, S, BRAXMAIER, C, SCHILLER, S, & PETERS, A. 2003. Modern Michelson-Morley experiment using cryogenic optical resonators. *Phys. Rev. Lett.*, **91**(2), 020401. Muller, Holger Herrmann, Sven Braxmaier, Claus Schiller, Stephan Peters, Achim eng 2003/08/09 Phys Rev Lett. 2003 Jul 11;91(2):020401. doi: 10.1103/PhysRevLett.91.020401. Epub 2003 Jul 10.
- NAKANO, A, & SHIRAISHI, M. 2004. Investigation for magnetic separation of oxygen from supercritical air near the maxcondentherm point. *AIP Conf. Proc.*, **710**(1), 1923.
- NELSON, P. 2017. *From photon to neuron: Light, imaging, vision*. Princeton NJ: Princeton Univ. Press.
- NELSON, P. 2020. *Biological physics student edition: Energy, information, life*. Philadelphia: Chliagon Science.
- NEUENSCHWANDER, D E. 2015. *Tensor calculus for physics: A concise guide*. Baltimore MD: Johns Hopkins Univ. Press.

- NEUENSCHWANDER, D E. 2017. *Emmy Noether's wonderful theorem*. Rev. ed. Baltimore MD: Johns Hopkins Univ. Press.
- NIEVES, JOSÉ F, & PAL, PALASH B. 1994. Third electromagnetic constant of an isotropic medium. *Am. J. Phys.*, **62**(3), 207–216.
- NOETHER, E. 1918. Invariante variationsprobleme. *Nachr. d. König Gesellsch. d. Wiss. zu Göttingen, Math-phys. Klasse*, 235–257.
- NOVOTNY, L, & HECHT, B. 2012. *Principles of nano-optics*. 2nd ed. Cambridge UK: Cambridge Univ. Press.
- NUSSENZVEIG, H M. 1992. *Diffraction effects in semiclassical scattering*. Cambridge UK: Cambridge Univ. Press.
- NYE, J F. 1999. *Natural focusing and fine structure of light: Caustics and wave dislocations*. Bristol UK: Institute of Physics Pub.
- PAIS, A. 1982. *Subtle is the Lord: The science and the life of Albert Einstein*. Oxford, UK: Oxford University Press.
- PANTANO, C, GANAN-CALVO, A M, & BARRERO, A. 1994. Zeroth-order, electrohydrostatic solution for electro-spraying in cone-jet mode. *Journal of Aerosol Science*, **25**(6), 1065–1077.
- PARKER, A R, & TOWNLEY, H E. 2007. Biomimetics of photonic nanostructures. *Nat. Nanotech.*, **2**(6), 347–353.
- PARKER, E N. 2007. *Conversations on electric and magnetic fields in the cosmos*. Princeton NJ: Princeton Univ. Press.
- PASTEUR, L. 1848. Sur les relations qui peuvent exister entre la forme cristalline, la composition chimique et le sens de la polarisation rotatoire. *Ann. Chim. Physique 3ème Sér.*, **24**, 442–459.
- PEATROSS, J, & WARE, M. 2015. *Physics of light and optics*. 2022 revision available at <http://optics.byu.edu>.
- PERKINS, T T. 2014. Ångström-precision optical traps and applications. *Annu. Rev. Biophys.*, **43**(1), 279–302.
- PESKIN, M E, & SCHROEDER, D V. 1995. *An introduction to quantum field theory*. Addison–Wesley.
- PHILLIPS, R, KONDEV, J, THERIOT, J, & GARCIA, H. 2012. *Physical biology of the cell*. 2nd ed. New York: Garland Science.
- PIERRUS, J. 2017. *Solved problems in classical electromagnetism*. Oxford UK: Oxford Univ. Press.
- PINE, D J. 2019. *Introduction to Python for science and engineering*. Boca Raton FL: CRC Press.
- PLANCHON, THOMAS A, GAO, L, MILKIE, D E, DAVIDSON, M W, GALBRAITH, JAMES A, GALBRAITH, CATHERINE G, & BETZIG, E. 2011. Rapid three-dimensional isotropic imaging of living cells using Bessel beam plane illumination. *Nat. Methods*, **8**(5), 417–423.
- POLLACK, G L, & STUMP, D R. 2002. *Electromagnetism*. San Francisco CA: Addison-Wesley.
- RAAB, R E, & DE LANGE, O L. 2005. *Multipole theory in electromagnetism: Classical, quantum, and symmetry aspects, with applications*. Oxford Univ. Press.
- RAINVILLE, S, THOMPSON, J K, MYERS, E G, BROWN, J M, DEWEY, M S, KESSLER JR, E G, DESLATTES, R D, BORNER, H G, JENTSCH, M, MUTTI, P, & PRITCHARD, D E. 2005. A direct test of $E = mc^2$. *Science*, **438**, 1096–1097.
- RAMAN, I M, & FERSTER, D L (Eds.). 2021. *The annotated Hodgkin and Huxley*. Princeton NJ: Princeton Univ. Press.
- RICHEY, L, STEWART, B, & PEATROSS, J. 2006. Creating and analyzing a mirage. *Physics Teach.*, **44**(7), 460–464.
- ROHRLICH, F. 2001a. The correct equation of motion of a classical point charge. *Physics Letters A*, **283**(5-6), 276–278.
- ROHRLICH, F. 2001b. Why the principles of inertia and of equivalence hold despite self-interaction. *Physical Review D*, **63**(12), 621.
- ROICHMAN, Y, SUN, B, ROICHMAN, Y, AMATO-GRILL, J, & GRIER, D. 2008. Optical forces arising from phase gradients. *Phys. Rev. Lett.*, **100**(1), 013602.
- ROTHMAN, T. 2003. *Everything's relative and other fables from science and technology*. New York: Wiley.

- ROTHMAN, T, & BOUGHN, S. 2009. The Lorentz force and the radiation pressure of light. *American Journal of Physics*, **77**(2), 122–127.
- ROVENCHAK, A, & KRYNYTSKYI, Y. 2018. Radiation of the electromagnetic field beyond the dipole approximation. *Am. J. Phys.*, **86**(10), 727–732.
- RUFFNER, DAVID B, & GRIER, DAVID G. 2012. Optical forces and torques in nonuniform beams of light. *Phys. Rev. Lett.*, **108**(17), 173602.
- RYBAK, A, GRZYWNA, Z J, & KASZUWARA, W. 2011. Influence of various parameters on the air separation process by magnetic membranes. *Polish Journal of Applied Chemistry*, **Vol. 55, nr 1**, 41–48.
- RYBICKI, G B, & LIGHTMAN, A P. 2004. *Radiative processes in astrophysics*. New York NY: Wiley-VCH.
- SAFRAN, S A. 2003. *Statistical thermodynamics of surfaces, interfaces, and membranes*. Boulder CO: Westview Press.
- SASLOW, W M. 2021. Voltaic cells: The good (Faraday), the bad (Volta), and the ugly (Galvani). *The Physics Teacher*, **59**(1), 22–26.
- SCHEY, H. M. 2005. *Div, grad, curl, and all that: An informal text on vector calculus*. 4th ed. W.W. Norton.
- SCHMIDT-ROHR, K. 2018. How batteries store and release energy: Explaining basic electrochemistry. *Journal of Chemical Education*, **95**(10), 1801–1810.
- SCHOBEIRI, M T. 2021. *Tensor analysis for engineers and physicists – with application to continuum mechanics, turbulence, and Einstein’s special and general theory of relativity*. Cham CH: Springer Nature Switzerland.
- SCHUTZ, B. 2022. *A first course in general relativity*. 3rd ed. Cambridge MA: Cambridge Univ. Press.
- SCHWARZ, U. 2021. *Theoretical biophysics*.
www.thphys.uni-heidelberg.de/~biophys/PDF/Skripte/TheoreticalBiophysics.pdf.
- SCHWINGER, J, DERAAD, JR., L L, MILTON, K A, & TSAI, W-Y. 1998. *Classical electrodynamics*. Reading MA: Perseus Books.
- SHANKLAND, R S. 1964. Michelson-Morley experiment. *Am. J. Phys.*, **32**, 16–35.
- SHAPIRO, I S. 1973. On the history of the discovery of the Maxwell equations. *Soviet Physics Uspekhi*, **15**(5), 651–659.
- SHARMA, V, CRNE, M, PARK, J O, & SRINIVASARAO, M. 2009. Structural origin of circularly polarized iridescence in jeweled beetles. *Science*, **325**(5939), 449–451.
- SHARMA, V, CRNE, M, PARK, J O, & SRINIVASARAO, M. 2014. Bouligand structures underlie circularly polarized iridescence of scarab beetles: A closer view. *Materials Today Proceedings*, 1–11.
- SIEG, P G, BERNER, W, HARNISH, P K, & NELSON, P C. 2019. A demonstration of the infrared activity of carbon dioxide. *Physics Teach.*, **57**(4), 246–249.
- SIEGEL, D M. 1991. *Innovation in Maxwell’s electromagnetic theory: Molecular vortices, displacement current, and light*. Cambridge UK: Cambridge Univ. Press.
- SIMON, D S. 2016. *A guided tour of light beams: From lasers to optical knots*. San Rafael CA: Morgan and Claypool.
- SINDBERT, S, KALININ, S, NGUYEN, H, KIENZLER, A, CLIMA, L, BANNWARTH, W, APPEL, B, MÜLLER, S, & SEIDEL, C A M. 2011. Accurate distance determination of nucleic acids via Förster resonance energy transfer: Implications of dye linker length and rigidity. *JACS*, **133**(8), 2463–2480.
- SMITH, A A T, BROWN, C V, & MOTTRAM, N J. 2007. Theoretical analysis of the magnetic Fréedericksz transition in the presence of flexoelectricity and ionic contamination. *Phys. Rev. E*, **75**(4), 315–8.
- SMITH, ALEXANDER M, BORKOVEC, MICHAL, & TREFALT, GREGOR. 2020. Forces between solid surfaces in aqueous electrolyte solutions. *Advances in Colloid and Interface Science*, **275**(C), 102078.
- SMITH, G S. 1997. *An introduction to classical electromagnetic radiation*. Cambridge UK: Cambridge Univ. Press.
- SMOOT, G F, GORENSTEIN, M V, & MULLER, R A. 1977. Detection of Anisotropy in the Cosmic Blackbody Radiation. *Phys. Rev. Lett.*, **39**(14), 898.

- SOHN, L L, SALEH, O A, FACER, G R, BEAVIS, A J, ALLAN, R S, & NOTTERMAN, D A. 2000. Capacitance cytometry: measuring biological cells one by one. *Proc. Natl. Acad. Sci. USA*, **97**(20), 10687–10690.
- SOMMERFELD, A. 1964a. *Electrodynamics*. New York: Academic Press.
- SOMMERFELD, A. 1964b. *Optics*. New York: Academic Press.
- SONI, V S. 1988. A note on the ray diagram of the Michelson–Morley experiment. *American Journal of Physics*, **56**(2), 178–179.
- SPIVAK, M. 1999. *A comprehensive introduction to differential geometry*. 3rd ed. Houston TX: Publish or Perish.
- SRINIVASARAO, M. 1999. Nano-optics in the biological world: Beetles, butterflies, birds, and moths. *Chem. Rev.*, **99**(7), 1935–1962.
- STONE, M, & GOLDBART, P. 2009. *Mathematics for physics: A guided tour for graduate students*. Cambridge UK: Cambridge Univ. Press.
- STOWE, S. 1980. Rapid synthesis of photoreceptor membrane and assembly of new microvilli in a crab at dusk. *Cell Tissue Res.*, **211**(3), 419–440.
- STRAUMANN, N. 2013. *General relativity*. 2d ed. New York: Springer.
- 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.
- STRUTT, J W (LORD RAYLEIGH). 1888. XXVI. On the remarkable phenomenon of crystalline reflexion described by Prof. Stokes. *Phil. Mag. Ser. 5*, **26**(160), 256–265.
- STRUTT, J W (LORD RAYLEIGH). 1890. Measurements of the amount of oil necessary in order to check the motions of camphor upon water. *Proc. R. Soc. Lond.*, **47**, 1–5.
- TANFORD, C. 1989. *Ben Franklin stilled the waters*. Durham NC: Duke Univ. Press.
- TAYLOR, G. 1964. Disintegration of water drops in an electric field. *Proc. R. Soc. Lond. A*, **280**(1382), 383–397.
- THOMPSON, A R, MORAN, J M, & SWENSON, JR., G W. 2017. *Interferometry and synthesis in radio astronomy*. 3rd ed. Berlin: Springer.
- THORNE, K S, & BLANDFORD, R D. 2017. *Modern classical physics*. Princeton NJ: Princeton Univ. Press.
- TOPRAK, E, ENDERLEIN, J, SYED, S, MCKINNEY, S A, PETSCHKE, R G, HA, T, GOLDMAN, Y E, & SELVIN, P R. 2006. Defocused orientation and position imaging (DOPI) of myosin V. *Proc. Natl. Acad. Sci. USA*, **103**(17), 6495–6499.
- TUCKER, T. 2003. *Bolt of fate: Benjamin Franklin and his electric kite hoax*. PublicAffairs.
- VAFABAKHSH, R, & HA, T. 2012. Extreme bendability of DNA less than 100 base pairs long revealed by single-molecule cyclization. *Science*, **337**(6098), 1097–1101.
- VAN DER STRATEN, P, & METCALF, H. 2016. *Atoms and molecules interacting with light: Atomic physics for the laser era*. Cambridge UK: Cambridge Univ. Press.
- VAN MAMEREN, J, WUITE, G J L, & HELLER, I. 2011. Introduction to optical tweezers: Background, system designs, and commercial solutions. *Meth. Mol. Biol.*, **783**, 1–20.
- VANDERLINDE, J. 2004. *Classical electromagnetic theory*. 2nd ed. Dordrecht NL: Kluwer Academic.
- WADHWA, N, SASSI, A, BERG, H C, & TU, Y. 2022. A multi-state dynamic process confers mechano-adaptation to a biological nanomachine. *Nat. Commun.*, **13**(1), 5327. Wadhwa, Navish Sassi, Alberto Berg, Howard C Tu, Yuhai eng K99 GM134124/GM/NIGMS NIH HHS/ R35 GM131734/GM/NIGMS NIH HHS/ Research Support, N.I.H., Extramural England 2022/09/11 Nat Commun. 2022 Sep 10;13(1):5327. doi: 10.1038/s41467-022-33075-5.
- WALD, R. 2022. *Advanced classical electromagnetism*. Princeton NJ: Princeton Univ. Press.
- WALMSLEY, I. 2015. *Light: A very short introduction*. Oxford UK: Oxford Univ. Press.
- WEINBERG, S. 1972. *Gravitation and cosmology*. New York: Wiley.
- WEINBERG, S. 2005a. *The quantum theory of fields*. Vol. 3. Cambridge UK: Cambridge Univ. Press.

- WEINBERG, S. 2005b. *The quantum theory of fields*. Vol. 1. Cambridge UK: Cambridge Univ. Press.
- WESS, J, & BAGGER, J. 1992. *Supersymmetry and supergravity*. 2nd ed. Princeton NJ: Princeton University Press.
- WHITTAKER, E T. 1951. *A history of the theories of aether and electricity*. Vol. 1-2. New York: Harper.
- WILL, C M. 2006a. Special relativity: A centenary perspective. *Pages 33–58 of: DAMOUR, T, DARRIGOL, O, DUPLANTIER, B, & RIVASSEAU, V (Eds.), Einstein, 1905–2005: Poincaré Seminar 2005*, Vol. 47. Basel CH: Birkhäuser Basel. arxiv.org/abs/gr-qc/0504085.
- WILL, C M. 2006b. Was Einstein right? *Annalen der Physik*, **15**(1-2), 19–33.
- WOLF, E. 2007. *Introduction to the theory of coherence and polarization of light*. Cambridge UK: Cambridge Univ. Press.
- YABLONOVITCH, E. 2001. Photonic crystals: Semiconductors of light. *Sci. Am.*, **285**(6), 47–55.
- ZAHN, C T. 1926. The electric moment of CO₂, NH₃, and SO₂. *Phys. Rev.*, **27**(4), 455–459.
- ZANGWILL, A. 2013. *Modern electrodynamics*. Cambridge UK: Cambridge Univ. Press.
- ZHANG, Y-C. 1997. *Special relativity and its experimental foundations*. Singapore: World Scientific.

Index

Bold references indicate the main or defining instance of a key term. Symbol names and mathematical notation are defined in Appendix B.

- aberration
 - spherical, 298, 305, 309
 - starlight, 394–397
- absolute temperature scale, 678
- absolute value of, **313**
- acceptor, **57**, 58, 61
- achiral object, 612
- acoustic mode in plasma, 644
- action, **525**
- action functional, **525**
- action potential, **167**, 168, 170, 171, 174, 182
 - mechanical analog, 179
 - speed, 181
- æther, 12, 17, 78, 261, 273–275, 361, 364, **365**, 365–367, 369, 370, 372, 373, 376, 381–383, 390, 395, 396, 467, 670, 691
- afterhyperpolarization, 175, 186
- ampere (unit), **679**
- Ampère's law, 663
- amplitude
 - complex, *see* Jones vector
- angle, 681
 - azimuthal, **10**
 - polar, **10**
- Ångstrom (unit), 680
- angular area, **10**
- angular frequency, **263**, **312**
- angular momentum flux tensor, **487**, 534
- angular velocity, **190**
- anion, **132**
- antenna
 - dipole, electric, 566
- antenna theory, **337**
- arbitrary units, 681
- arc length, 309, 310
- arcminute (unit), 681
- arcsecond (unit), 681
- area, angular, 682
- ATP
 - hydrolysis of, 137
- Avogadro number, 694
- axon, **118**, **167**
 - giant, 175, 183
- axoplasm, **91**
- battery, **150**, *see* voltaic cell
- beam, 512
 - gaussian, **515**
 - laguerre-gaussian, **515**
 - waist, **515**
- beauty, 450
- benzene, 51
- Berg, Howard, 89
- Bessel
 - beam, **516**
 - function, **516**
- bialy, 508
- biaxial symmetry, **48**, 52, 108
- Big Bang, *see* early Universe
- bilayer membrane, **86**, 126, 127, 138
- Biot–Savart formula, 222
- birefringence
 - circular, 625, *see* circular birefringence
 - ordinary, 193, 609, 611, **625**
- Birkhoff's theorem, 41
- Bjerrum length, **155**
- Boltzmann constant, 694
- Boltzmann distribution, 86, 93
- boost
 - combination, 381
 - galilean, **348**, 349, 359, 361, 365, 375, 378, 390, 397, 410
 - Lorentz, 353, **390**, 390, 391, 393, 395, 396, 401, 402, 404, 406, 411, 412, 416, 421, 430, 431, 435, 438, 446, 448, 502, 504
 - combination, 402
 - provisional, 378, 383, 387–389
 - rapidity, 404
- Born self-energy, **84**, 86, 137
- Born, Max, 84
- Bradley, James, 387, 394
- bremsstrahlung, 477, 634
- de Broglie relation, 661
- de Broglie, Louis, 306, 436
- cable equation
 - linear, 166
 - nonlinear, **181**, 182
- capacitance, **74**, 143
- capacitor, 143
- cation, **132**
- Cavendish, Henry, 21, 29, 115
- CD, *see* circular dichroism
- Celsius temperature scale, 678
- Čerenkov radiation, 549, **633**
- Čerenkov, Pavel, 633
- charge
 - density, **3**
 - bound, 82, **83**, 605
 - surface, 138, 139
 - surface, bound, **77**
 - surface, free, **77**
- density, surface
 - bound, **82**
- flux, **3**
 - 1D, **112**
 - 2D, **113**, 590, 591
 - 3D, **114**
 - 4-vector, **462**
- bound, 605, 606
 - wire, 223
- charge renormalization, **150**
- chicken-egg problem, 127, 139, 140, 421
- chimera, **57**
- chiral medium, *see* medium
- chirality, **612**, 612
- chirp, 646
- chromatic aberration, **609**
- chromophore, 668
- chutzpah, 12
- circular birefringence, **611**, 615–616, 648
- circular dichroism, **617**
- circular frequency, **263**
- classical electron radius, **596**
- Clausius–Mossotti formula, 82
- Clausius–Mossotti relation, 80
- clumping catastrophe, 136
- CMBR, *see* cosmic microwave background radiation
- Cockcroft, John, 414
- coherent states, **662**
- coion, 139, 148
- colloidal suspension, 136
- complex conjugate, **313**
- complex numbers, **313**
- compliance tensor, **193**
- components, 191, 422–424, 429, 430, 440, 456, 471
 - 3D, **189**, **191**
 - 4D, 472
- components of a tensor, **201**
- Compton scattering, **415**, 599
- Compton wavelength, **599**
- conductance, **115**
 - resting, 179
- conductivity, **115**
 - electrical, 164
 - tensor, **193**
- cone
 - cell, 668
- constitutive relation, **78**
 - electric, **608**
 - magnetic, **608**
- continuity equation, 113, **114**, 221–223, 261, 330, 362, **462**, 462, 464, 479–483, 492, 543, 575, 605
 - charge, 528
 - energy of a spring, 362
 - momentum of a spring, 362
 - SO(2) charge, 531
- contraction, 426, 432
 - 3D, **426**, **428**
 - 4D, 458, *see* invariant inner product
- coordinate system
 - cartesian, **202**

- curvilinear, 4, 5, 33, **66**, 65–71, 202, 209, 221, 264, 269, 422, 427, 454, 684
inertial, *see* inertial coordinate system
on spacetime, 348
inertial, *see* inertial coordinate system
separable, *see* separable coordinate system
- cosmic microwave background radiation (CMBR), 354, 400, 406, 504, 597
dipole anisotropy, 400, 406
- Coulomb
attraction, 664
gauge, 656, 657, 663
coulomb (unit), 678, **679**
- Coulomb gauge condition, *see* gauge
- counterion, 135–147, 158
cloud, 139
release, 146
- counterion release, 147
- covector, 199
- covector, 4D, **454**
- creation and destruction operators, **661**, 690, 693
- cross product, **196**
- cross-section
scattering
differential, **596**
Thomson, **596**
- cross-susceptibility, 648
- curl, **7**
-free vector field, **9**
- current
density, *see* charge flux
surface, *see* charge flux, 2D, *see* flux, charge, 2D
density (term not used in this book), *see* flux
displacement, *see* flux, charge, displacement
- eddy, 241
- electric, 679
through ion channel, 169
- curvature, **301**
curve in plane, **98**
surface
Gauss, **97**, **101**, 102, 106
mean, **97**, **101**, 102–104, 106, 110
principal, **101**
- cutoff, **279**
- cyclotron motion, **445**, 647
- d'Alembert equation, **330**, 497, 537, 539–541, 543, 551
- d'Alembert operator, *see* wave operator
- dalembertian, *see* wave operator
- de Sitter, Willem, 352
- debye (unit), **235**
- Debye screening length, **148**, 149, 150, 152
- defocused orientation imaging, 655, 666
- degree (angular unit), 681
- degree (temperature unit), 678
- degree (unit), 10
- delta function, **10**, 17, *see* Dirac delta function
dimensions of, 10
- dendrites, **118**
- density matrix, 199
- depletion layer, **31**
- depolarization, **120**
- depth
optical, 615
- deRham cohomology, **227**
- dialectical materialism, 74
- diamagnetism, **249**
- dielectric, **76**
- dielectric constant, **78**
- dielectric susceptibility, *see* susceptibility
- differential forms, **227**
- diffraction pattern, **508**
- diffuse charge layer, **136**, 143
- diffusion
constant, 166
equation, 166
- diffusion constant, 116
- diffusion equation, 181
- fundamental pulse solution, 166
- dilation, **347**, 347, 348, 351, 388–391, 402
- dimensional analysis, **678**, 679–682
- dimensionless quantities, **678**, 681
- dimensions, **679**
- dipole
anisotropy of CMBR, *see* cosmic microwave background radiation
approximation
electric, 664
doughnut, **508**, 509, 665
electric, 69
moment, 693
moment, *see* multipole moments
point, *see* dipole, electric, pure
radiation pattern, 655, 666, *see* dipole doughnut
transition, **665**, 665, 667, 668, 693
- Dirac equations, **475**
- dispersion, **609**, 694
- dispersion interaction, **46**, 60
- dispersion relation, **268**, **270**, **279**
- displacement, **78**
- displacement, electric, **605**
- dissipation, 27, 115, 116, 168, 350, 592, 604, 608, 617, 619, 627
- dissociation, 158
- divergence operator, **5**
- divergence theorem, **9**, 18, 67, 527
- DNA, 61
denaturation, 146
dissociation, 135
FRET measurements on, 58
Z form, 617
- donor, **57**, 58, 61
- Doppler shift
longitudinal, **396**
transverse, **396**
- dot product, **4**, **205**
- drag force, **192**
- drift velocity, 116
- duality, electric–magnetic, 471
- du Châtelet, Émilie, 410
- dyad product, **191**, 191, 426, 458, 683
- e-folding time, 277
- early Universe, 400, 501, 594, 597
- EEG, *see* electroencephalography
- eikonal, 300
equation
in medium, **302**
vacuum, **301**
function, **300**
- eikonal trial solution, **300**
- Einstein relation
diffusion versus drag, 116, 133
frequency versus energy, **436**, 661
- Einstein ring, **311**
- Einstein thinking, **411**, 414, 436, 441, 442, 458, 468, 475, 478, 479, 483, 527, 528, 537, 539, 541, 621, 674
- Einstein, Albert, 2, 11, 20, 215, 289, 305, 364, 365, 412, 415, 420, 436, 467
chutzpah, 12
- EKG, *see* electrocardiogram
- electret, *see* ferroelectric
- electric
charge, 679
field, 656
- electric double layer, **31**, **136**
- electro-optical effect, 628, 642
- electrocardiogram, **123**
- electrodynamics
classical, 666
quantum, 658–669
- electroencephalography, **122**
- electrolyte, **143**
- electromotive force (deprecated term), *see* EMF
- electron, 663, 664
charge, 694
mass, 689, 694
spin, 669
volt (unit), 680
- electrophoretic flux, **133**
- electrospinning, **105**
- electrospray, **105**
- electrostatic potential, 140
- electrostatics, 663
- electrostriction, **89**
- electroweak theory, 403
- EMF (deprecated term), 259
- emission spectrum, **57**
- Empire State Building, 347
- enantiomers, **612**, 617, 620
- endoscopy, **299**
- energy
alternative units, 680
dimensions, 679
free, 84, 102, 119, 158, 171, 172, 178, 212
counterions, 147, 155–157
particle
newtonian, **410**
relativistic, **413**
photon, 356, 660
potential
of capacitor, 143
electrostatic, 75, 156
solar, 304
thermal, room temperature, 694
- energy–momentum flux 4-tensor, **479**, 531
- entropic force, 155
- entropy
counterions, 135, 139, 143, 145, 147
- equilibrium
electrochemical, 152, 153
- erg (unit), 680
- esu (unit), **238**
- euclidean group, *see* group
- Euler–Lagrange equations
field, **526**
mechanics, **525**
- event, **344**
- excitable medium, **178**, 183
- excitation
transfer, 61
- excitation spectrum, **57**
- extensive quantity, **257**
- exterior derivative, **227**
- eye
bee, 668

- crustacean, 668
- fish, 309
- insect, 668
- Fahrenheit temperature scale, 678
- far field
 - dynamic, 507, **560**
 - in solution, 146
 - static, **36**, 36, 39, 243, 245, 560
 - synchrotron, 583, 585, 586
- far field (dynamic), 334, 335, 338–340, 507, 508, 558, 560, 561, 564, 565, 567, 568, 571, 573–576, 578, 579, 581, 586, 597
- far field (static), 334
- farad (unit), **235**
- Faraday effect, *see* magneto-optical effects
- Faraday tensor, *see* tensor, 4D
 - plane wave, 499
- Faraday, Michael, 132, 153, 162, 257, 261, 491–495, 642
- feedback, 177, 179, 180
- Fenn, John, 105
- Fermi, Enrico, 437
- ferroelectric, **89**, 620
- ferromagnet, 620
- ferromagnetism, **249**, 621
- Fick's law, 133
- field
 - electric, **3**
 - intensity, 3
 - magnetic, **3**
 - tensor, **198**
- field line, **8**
- field lines, **492**
- field point, **28**
- field quantization, 656–669
- first fundamental form, *see* tensor, 2D
 - metric
- fission, nuclear, 137
- FitzGerald, George, 333
- FitzHugh–Nagumo model, **186**
- fixed point, 181
- Fizeau experiment, 276, 366, 381–383, 398, 421, 623
- flagellum
 - bacterial, 193
- fluorescence
 - resonance energy transfer (FRET), 57, 58, 63
 - efficiency, 58, 61, 62
- fluorescence microscopy, **57**
- fluorescence resonance energy transfer, **57**
- fluorophore, **57**, 61
 - Cy2, Cy3, Cy5, Cy5.5, 58
 - fluorescein, 58
 - Texas red, 58
- flux, **114**
 - 1D, **112**
 - charge, **113**, 193, 663, 664
 - 2D, 224
 - bound, electric dipole, **605**
 - dielectric displacement, 115
 - 4D, free, 620
 - surface (bound), **605**
 - vacuum displacement, 275, 276
 - electrophoretic, **133**
 - magnetic (not used in this book), 115, 257
- focal length, 516
- focus, **297**, 298, 305
- forbidden transition, 668
- form, differential, **471**
- forms, *p*, *see* differential forms
- Förster radius, 61, 693
- Förster radius, **61**
- Foucault pendulum, 405
- frame of reference, *see* coordinate system
 - on spacetime
- Franklin, Ben, 25, 69
- Franklin, Benjamin, 65, 103, 126, 130, 356
- free energy
 - counterions, 143
- Fresnel, Augustin-Jean, 616
- FRET efficiency, **61**
- Fricke, Hugo, 126, 127
- friction
 - coefficient, viscous, 132
- von Frisch, Karl Ritter, 666
- fundamental form
 - first, 108
 - second, 108
- fussiness, 226
- galilean boost, *see* boost
- galilean invariance, **344**
- galvanic cell, *see* voltaic cell
- gauge choice
 - Coulomb, **219**, 221, 222, 269–271, 330, 331, 497, 509, 542, 656
 - Lorenz, *see* Lorenz gauge
- gauge fixing, **219**
- gauge invariance, **219**, **463**
- gauge transformation, **219**
- gauss (unit), **237**
- Gauss Law
 - at a surface, 138
- Gauss law, 138, 140, 141
 - electric, 78, 138, 663
 - in bulk, 138
- gaussian units, 235–238
- general relativity, 68, 106, 372, 427, 447, 450, 454, 467, 478, 675
- generator of rotation, **45**
- Gibbs, J. Willard, 672
- Golden Rule, 664, 668
- Gouy–Chapman layer, **142**, 144
- gradient-index
 - (GRIN) lens, 309
- Grand Unification
 - Newton's, 23
- grand unification, 226, 240
 - Newton's, 19
- gravitation, 27
 - force due to, 132
 - potential, *see* potential
- gravitomagnetic effect, 257
- Green function, **19**
 - causal, **542**
 - Laplace operator, **30**, 331
 - retarded, **542**
- greenhouse gas, **567**
- group
 - covering, 472
 - euclidean, 393, 416
 - galilean, **349**, **350**
 - Lorentz, 381, **393**
 - proper, **431**
 - Lorentz (O(3,1)), **430**
 - orthochronous Lorentz (O⁺(3,1)), **431**
 - orthogonal (O(3)), 423, 431, 620
 - proper Lorentz (SO(3,1)), **431**
 - proper orthochronous (restricted) Lorentz (SO⁺(3,1)), **431**, 472–474
- representation
 - fundamental, **471**, **472**
 - rotation (SO(3)), 423, 431, 471–473, 620
 - SL(2,C), 473, 474
 - Spin(3), 472
 - Spin(3,1), 473
 - SU(2), 450, 472–474
 - SU(3), 450
 - SU(5), 675
 - \mathbb{Z}_2 , 620
- gyromagnetic ratio, 277
- hairsplitting, 46
- Hamiltonian operator, 688, 693
- hamiltonian operator, 658, 659, 664
- harmonic oscillator, 657–660
- heat equation, 166
- Heaviside layer, *see* ionosphere
- Heaviside, Oliver, 11, 260, 274, 279, 303, 447, 483, 633, 672
- helicity
 - negative, **272**
 - positive, **272**
- helicity basis, **272**, 280, 615
- Helmholtz
 - coils, 229
 - equation, **512**, 516
 - free energy, 155
- Helmholtz, Hermann von, 167
- henry (unit), **235**
- Henry, Joseph, 275, 672
- Herschel, Sir John, 642
- hertz (unit), 680
- Hertz, Heinrich, 275, 672
- Hodge dual, 471
 - 4D, **471**
- Hodgkin, Alan, 177
- homonuclear molecule, **567**
- honeybee, 666
- Hooke, Robert, 24, 41
- Huggins, Margaret, 399
- Huggins, William, 399
- Huxley, Andrew, 177
- Huygens, Christiaan, 364
- hydrophobic effect, **86**
- i (square root of -1), 687
- ice, 84
- identity tensor, **191**, **685**
- improper Lorentz transformations, **431**
- index
 - dummy, **8**
 - loose, **8**
- index lowering, **455**
- index raising, **455**
- indistinguishable particles, 661
- induced dipole, **46**
- induction
 - magnetic, 3
- inductor, **278**
- inertial coordinate system
 - Einstein (E-), **390**, 390, 393, 395, 399, 400, 403, 407, 410, 416, 429, 432, 538, 688, 690–692
 - galilean (G-), **348**, 348, 369, 370, 390, 538, 690
- inertial frame
 - Einstein (E-), 438, 457, 460, 461
- instanton, 671
- integrability lemma, **29**
- integrable, **533**
- integrable model, **533**

- integral curve, *see* streamline
interference, **317**
invariance, **345**
 manifest, 11, 12, 226, 229, **428**, 428, 435, 441, 445, 449, 451, 456, 463, 469, 562
invariant inner product, **432**
invariant interval, **391**, 402, 412, 416, 419, 432, 433, 467, 540, 692
 (1+1)D, **391**
 (3+1)D, **393**
ion, 132
 channel, **86**, 119, 121, 169, 172, 186
 voltage-gated, 178
ionosphere, 299, 303, 304, 643, 645
isosurface, **54**
isotropic medium, *see* medium
- Jones tensor, **592**
Jones vector, **266**, **327**, 592
joule (unit), **679**
Joule heating, *see* ohmic heating
jumpstart box, 155
- Keesom interaction, **45**, 60
Kelvin, *see* Thomson, William
kelvin (unit), 680
Kelvin temperature scale, 678
Kerr effects, *see* electro-optical effect *and* magneto-optical effects
Kerr, John, 628, 642, 648
kink, 179
Klein-Gordon equation, **475**
Kronecker symbol, **7**, **195**
- lagrangian density, **525**, 526–528, 530–532
Laplace equation, **28**, 39, 69, 76, 104, 120, 128
Laplace operator, *see* laplacian
laplacian, **5**, 30, 52, 66, 68, 70–72, 124, 125, 221, 391, 683
laser, 663, **669**
law of refraction
 generalized, 303
leak conductance, 164
Legendre
 equation, **69**, 104
 function, 104, 110
 function, fractional order, 104
 polynomial, 104
lens, 516
Lenz's law, **258**, 259, 614
level set, **54**
Levi-Civita
 tensor, *see also* tensor
Levi-Civita symbol
 3D, **6**
Levi-Civita tensor
 4D, **469**, 621, 691
Liénard-Weichert
 fields, **582**
 potentials, **544**, 545, 549, 581, 632
light
 speed of, 356, 694
 unpolarized, **326**, 326, 327
light cone, **433**, 540, 542, 544, 557, 635
 coordinates, **402**
lightlike separation, **432**, 433, 538, 540
linear cable equation, **165**
liquid crystal, 194
liter (unit), 680
local action functional, **526**
 localization microscopy, 669
 Loligo forbesi, 118
 London force, **46**
 Lorentz boost, **390**
 Lorentz force law, **3**, 12, 28, 234, 237, 247, 389, 420, 421, 442, 444–447, 451, 482, 486, 489, 495, 529, 530
 4-vector, 442
 Lorentz group, *see* group
 Lorentz transformation, **430**
 proper and orthochronous, **473**
 provisional, 378
 Lorentz, Hendrik, 215, 261, 275, 282, 497
 Lorenz gauge, **497**, 497, 498, 502–504, 507, 509, 512, 523, 539, 542, 543, 562
 Lorenz, Ludvig V., 542
 Lorenz, Ludvig Valentin, 497
 lowering operator, **659**, 660
 luminiferous æther, *see* æther
- macroion, 135–136
magnetar, **282**
magnetic dipole moment, *see* multipole moments
magnetic field, 656
 intensity (H), **606**
 tensor, 214, 425
magnetic susceptibility, *see* susceptibility
magnetic vector potential, **218**
magneto-optical effects, 642, 646–649
magneto-electrophoresis, 228
magnetostatics, **219**
mantis shrimp, 667
Marangoni effect, 102
Marconi, Guglielmo, 303
mass
 relativistic (deprecated term), 412
 rest (deprecated term), 412
mass defect, **414**
matrix
 orthogonal, **202**, **423**
 rotation, 101
Maxwell equations, 656, 657, 663
Maxwell, James Clerk, 2, 235, 274–275, 369, 373, 642, 672
mean
 rate, 665
mean-field approximation, **139**, 140
medium
 chiral, **612**
 isotropic, **608**
Mensing, Lucie, 60
metamaterial, 609
metric, **430**
mho (unit, not used in this book), *see* siemens
Michelson
 interferometer, 500
Michelson-Morley, 366, 367, 376, 381
 improved version of Fizeau experiment, 381, 383
Michelson, Albert, 261, 381
microvillus, 667, **668**
Minkowski, Hermann, 190, 215, 419, 447, 478
mirage, 304
molar (unit), 680
mole (unit), 680
moment of inertia, 476
moments, *see also* multipole moments
 first, **37**
 of inertia, 40, **190**, 424
 second, **37**, 37, 245, 575
 zeroth, **37**
 zeroth and first, **37**
momentum
 4D, **435**
 angular, **190**
 electromagnetic, 657, 659, 690, 693
 particle
 newtonian, **409**
 relativistic, **412**
 momentum flux 3-tensor, **194**, 476, 484, 492, 494, 495, 500, 501
 momentum flux tensor, 480
 monochromatic light, **324**
monopole
 magnetic, 244
monovalent ion, **139**
Mössbauer effect, 397, **417**, 417
multipole
 potentials, **37**
multipole approximation, **561**
 newtonian gravitation, 41
 radiation, **563**
 ED (electric dipole), **563**, 565, 568–570, 577
 EQ (electric quadrupole), 576
 MD (magnetic dipole), 575
multipole moments
 electric dipole, **37**, 665
 fundamental particles, 48
 electric dipole, induced, **193**
 electric monopole, **37**
 electric quadrupole, **37**, 191, 424
 magnetic dipole
 fundamental particles, 49
 magnetic dipole (tensor form), **244**, 425
 magnetic dipole (vector form), **245**
 magnetic quadrupole, **246**
 octupole, 42
murder mystery, 617
- Nernst-Planck formula, **133**, 133, 135, 152, 153
Nernst potential, 133, **134**, 168, 169
neuron, **117**
Newton
 constant, 694
Newton, Isaac, 356, 364
newtonian potential, **16**
nightmare, 30, 40
Noether theorem, 411, 524, **531**, 530–532
nonchiral object, *see* achiral
nonlinear optics, 620
nonrelativistic limit, **380**
normal coordinates, 97, **101**, 109
normal vector, **9**
normalization
 of a vector, **5**
null
 experiment, 381, 382, 387
 separation, **432**
 wavevector, 498
- O(3), *see* group
O(3,1), *see* group
O(4), *see* group
observer, *see* coordinate system on spacetime
occupation numbers, **661**
oersted (unit not used in this book), **237**
ohm (unit), 115, **235**, 680
Ohm's law, *see* ohmic material

- Ohm, Georg, 115
ohmic
 heating, **116**, 277
 hypothesis (neuron), **169**, **174**, 174
 material, **115**, 193, 590, 644
ommatidium, 667, **668**
one-shot action potential model, 178–183
optical activity, *see* circular birefringence
optical rotation, 611
optical rotatory dispersion, **617**
optical rotatory power, *see* circular birefringence
optical torque wrench, 515
optical vortex, **515**
ORD, *see* optical rotatory dispersion
orientation
 curve (1D), 9, 256, 258
 disambiguation, 698
 space (3D), 204, 209
 spacetime (4D), 469–472
 time, 470
orthogonal matrix, 687
osmotic pressure, 144
outer product, *see* dyad product

paramagnetism, **249**
paraxial
 conditions, 513
 equation, 513
paraxial equation, 522
Parker, Edwin, 226
passive-spread solution, **166**
Pasteur, Louis, 616
patch-clamp, **127**
Pauli matrices, **473**, 475
Pauli, Wolfgang, 473
Peebles, P. James E., 400
period, **263**
permeability
 magnetic, **608**
 vacuum, **3**, **234**
permeability of vacuum, 694
permittivity, **78**, 92, 193, **608**, 635
 vacuum, **3**, **234**
permittivity of vacuum, 694
phosphorescence, **669**
photobleaching, 58
photoisomerization, 667
photon
 ground state, **660**
 indistinguishability of, 661
 momentum, 660
photonic bandgap materials, **638**
photoreceptor, 167, 668
 rhabdomic, **668**
piezoelectricity, **89**, 620
Planck
 constant, 693, 694
Planck units, **240**
plane wave, **263**
plasma
 cold, **643**
 frequency, **645**
Plus Ultra, xl
Poincaré group, **393**, 398, 416, *see* group
Poincaré lemma, xxxvi, **218**, 227, 463, 470
Poincaré sphere, **325**, 327
Poincaré, Henri, 343, 373
point dipole, **42**
Poisson equation, **28**, 140
Poisson–Boltzmann equation, **141**, 141,
 144, 148, 155, 156, 159
 boundary condition, 141
 linearized, 149, 159
polarimeter, 611
polarizability
 anisotropic, 609
 bulk, **77**
 molecular, **46**
polarization
 circular, **271**
 elliptical, **271**
 of light, 657, 661, 662, 664–666, 668
 basis vectors, **656**, 691, 693
 linear, **271**
 magnetic, 249
 partial, degree, **327**
polarization vector, **270**
polaroid filter, 592
positron emission tomography, 31, 437
potential
 electric, 656, 663, 664, 680
 electrostatic, **28**
 multipole, *see* multipole potentials
 newtonian (gravitational), **16**
 vector, 656
 4D, **463**
power, 679
Poynting
 theorem, 483
 vector, 484, 575, 636, 652
Poynting theorem, **485**
Poynting vector, 586, 658
Poynting, John, 288, 290, 483
pressure, **195**
 radiation, 288, 292, 484
 isotropic, 501
Principle of Relativity, **343**, 343, 344, 351,
 353, 357, 364, 366, 375, 376, 379,
 380, 387, 389, 397
probability
 density function, 655, 665, 690
probability distribution
 Boltzmann
 and Nernst relation, 134
 and Poisson equation, 140, 148
product
 dyad, **191**, *see* dyad product
 tensor, *see* tensor product
propagator, *see* Green function
proper time, **432**
pseudotensor, concept not used in this
 book, 205, 226
pulsar, 579
pure dipole, **42**

quadrupole, *see also* multipole moments
 magnetic, **245**
quarterwave plate, 630
quasi-static approximation, 117, 120, 125,
 127, 163, 165, 168, 673, 693

racemic mixture, 617
radian (unit), 10, 681
radiation
 anapole, **577**
rainbow, 323
raising operator, **660**
rank
 4-tensor, 440
 alternate definition not used in this
 book, 199
spin
 3D, 473
 4D, 474
rapidity parameter, **391**
ray
 equation, **302**, 309, 310
 of light, 298, 323
ray of light, 296, **299**
ray optics, **300**
Rayleigh (J. Strutt), 126
Rayleigh cross-section, **598**
Rayleigh range, **513**
real part, **313**
redshift
 gravitational, 418
refraction
 law of, 298, 309
refractive index, **295**, 299, 694
 effective gravitational, 305, 309, 310
 graded, 305
Relativity Strategy, **351**, 382, 393, 395,
 398, 406, 407, 438, 446, 447
representation
 group, **471**
residue, 136
resistance, **115**, 135
resistivity, **115**
response function, **77**, 607, 608
rest frame, **348**, 392, 393, 399, 404–406,
 432, 446, 453, 476, 502, 632, *see also*
 coordinate system on spacetime
RET, *see* FRET
retinal (cofactor), 667
rhabdom, 667
rhabdomere, **668**
rhodopsin, 667
Ricci-Curbastro, Gregorio, 189
rod cell, 668
ross-susceptibilities, **613**
rotation
 proper and improper, terms not used in
 this book, 202
rotation group, *see* group
rotation matrix
 infinitesimal, **44**
rotation measure, **650**
the Rules
 3D, 427, 428
 4D, 421, 441, 457, 459, 460, 463, 476,
 483, 486, 526
 spinor
 3D, 473
 4D, 474

scalar
 3D, **189**, **424**
 4D (4-scalar), **432**, 432, 441
scalar product
 3D, **4**
Schrödinger, Erwin, 306
Schwarzschild, Karl, 524
second fundamental form, *see* tensor, 2D
 curvature
 selection rule, **669**
self-inductance, **257**
semiconductor, 147
separable coordinate system, **66**, 66
short circuit, 116
siemens (unit), 115, **235**
simultaneity, relativity of, **538**
skip, 303
skip (skywave transmission), 303, 645
SL(2, \mathbb{C}), *see* group
SL(2, \mathbb{R}), *see* group

- sludge, 136
 small source approximation, **561**
 small source limit, **334**, 334, 561, 562, 564, 574
 Smoot, George, 400
 SO(3), *see* group
 SO(3,1), *see* group
 SO⁺(3,1), *see* group
 solenoid, **255**
 solid angle, **10**
 soliton, **186**, 298
 solubility, 85, 86, 95
 source point, **30**
 space constant of axon, **165**
 spacelike separation, **433**, 538, 540
 spacetime, **344**
 specific absorbed rate (SAR), 277
 spectrum
 absorption or excitation, 58
 emission, 58
 spherical aberration, **298**
 spherical harmonics, 48
 spherical wave, **506**
 Spin(3), *see* group
 Spin(3,1), *see* group
 spin
 1/2, 468
 spinor, 199, 428, 441, 447
 3D, **472**
 Dirac, **475**
 Weyl, **474**
 spring constant tensor, **193**
 standard deviation, also called
 root-mean-square deviation (RMSD)
 relative (RSD), 662
 static, **26**
 static system, **219**
 stationary charges, currents, fields, **26**, 115, 216, **219**, 220, 221, 243, 244, 250, 255, 331, 622
 steady state, 179
 steradian (unit), 10, 682
 stereospecific binding, **137**
 Stokes parameters, **325**, 474
 Stokes shift, **57**
 Stokes Theorem, **9**
 strain rate tensor, **427**
 streamline, *iv*, **7**, 53, 230, 299–301, 309, 335, 339, 491–493, 571
 stress
 -energy tensor, term not used in this book, *see* energy–momentum flux 4-tensor
 shear, **195**
 stress tensor, Maxwell, term not used in this book, 484
 stress 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), *see* group
 SU(3), *see* group
 SU(5), *see* group
 subgroup, **388**, **423**
 summation convention, **8**
 supersymmetry, **534**
 surface attraction, electrostatic, 146
 surface repulsion, electrostatic, 144
 susceptibility
 4-tensor operator, **621**
 dielectric, **77**, 193
 susceptibility, dielectric, **607**
 susceptibility, magnetic, **608**
 symmetries, **345**
 synchrotron radiation, **581**
 Taylor cone, **104**, 105
 Taylor's theorem, 680
 tension
 interfacial or surface, 96, 97, **99**, 99, 100, 102, 104, 110
 line, **99**, 99, 100
 tensor
 antisymmetric, 196, 425, 443
 part, 217, 425
 field, *see also* field
 4D, 416, **440**, 441, **456**
 energy–momentum flux, **479**, 531
 Faraday, **442**, 475
 from Heaven, 201, 203, 205, 207, **209**, 212, 430, 457, 501
 spin, 472–474
 magnetic dipole moment, *see* multipole moments
 magnetic field, *see* magnetic field
 moment of inertia, *see* multipole moments
 operator, 199
 product, 191, **425**, 473, 474, 683
 group representation, 471–473
 Riemann, 106
 strain rate, **427**
 symmetric, **191**, 195–197, 425
 part, 425
 3D, **189**, **423**
 compliance, **193**
 Levi-Civita, **196**, 258
 metric, **195**
 mobility, **194**, *see* mobility tensor
 moment of inertia, *see* moment of inertia tensor
 momentum flux, *see* momentum flux 3-tensor
 polarizability, **193**, *see* polarizability tensor
 rank 2, **190**, **424**
 rank 3, **196**
 spring constant, **193**, *see* spring constant tensor
 totally antisymmetric, 197
 viscous drag, **192**, *see* viscous drag tensor
 trace, 195
 traceless, **195**
 2D, **189**
 curvature, 106
 metric, 108
 Tensor Principle
 3D, **428**
 4D, **441**, 441, 450, 462, 463
 tensor product
 3D, **428**
 4D, **458**
 tesla (unit), **235**
 test body, **3**
 Thomson, J. J., 553
 Thomson, William, 161–163, 166, 255, 260
 threshold stimulus, 181, 182
 time
 proper, **412**
 retarded
 distributed source, **332**
 point source, 544
 time constant of axon, **165**
 time-reversal invariance, 117
 timelike separation, **433**, 538
 TIR, *see* total internal reflection
 torque
 on fixed magnetic dipole, 248
 total confusion, 457
 total internal reflection, 295, 297, **298**
 trace, **191**, 426, *see also* contraction
 trajectory, **344**
 transfer matrix, **640**
 transformation
 active, **344**
 galilean, **350**
 Lorentz, *see* Lorentz transformation
 passive, **345**
 traveling wave, 120, 166–168, 174, 176, 178, 181, 182, 185, 186, 260, **263**, 270, 286, 298, 312, 358, 359, 361, 378
 attenuated, 268
 muscle, 122
 trigger waves, **178**
 tweezers
 magnetic, 250
 optical, **46**, 289, 296, 297
 twinlead cable, 76
 twisted tensor, concept not used in this book, 210
 uncertainty relation, 668
 uniaxial symmetry, **48**, 52, 108
 uniform transparent medium, 309
 unit tensor, *see* identity tensor
 unitary matrix, **472**, 687
 units, 231–238, 678–682
 base, **232**, 678
 dimensionless, 681
 Système Internationale (SI), **678**
 vacuum state, **661**
 van der Waals, **46**
 van der Waals force, 136
 Vavilov, Sergey, 633
 vector
 3D, **4**, **423**
 4D, **413**, **430**
 null, *see* null
 contravariant (term not used in this book), 454
 covariant (term not used in this book), 454
 4D (4-vector), **430**, 441
 vector potential, 668
 velocity
 4D, **433**
 velocity addition
 galilean, **349**
 Verdet constant, **649**
 vesicle, 143
 Victoria, Queen, 162
 viscous drag tensor, **192**
 volt (unit), **235**
 voltage-gating hypothesis, **172**, 174, 180, 182
 prompt, 178
 voltaic cell, 150–155
 Walton, Ernest, 414
 watt (unit), **679**
 wave equation, **358**
 inhomogeneous, *see* d'Alembert equation

- wave operator, **332**, 390, 391, 393, 402, **456**, 456, 550, 683, 688
- wavelength, **263**
- wavenumber, **263**, 312
 - spectroscopic (alternate definition not used in this book), 263
- wavevector, 656
 - 4D, **434**, 498
- weak interaction, 403
- weber (unit), **235**
- Weyl equation, **475**
- Wilkinson, David, 400
- winding number, 515
- world-line, 344
- Yang–Mills theory, 447
- Young–Laplace formula, **103**, 106, 110
- Yukawa equation, **527**
- z-stack, 54
- Zeeman
 - effect, 281, 579
 - experiment on light velocity, 381
- Zeeman effect, **281**
- Zeeman, Pieter, 282