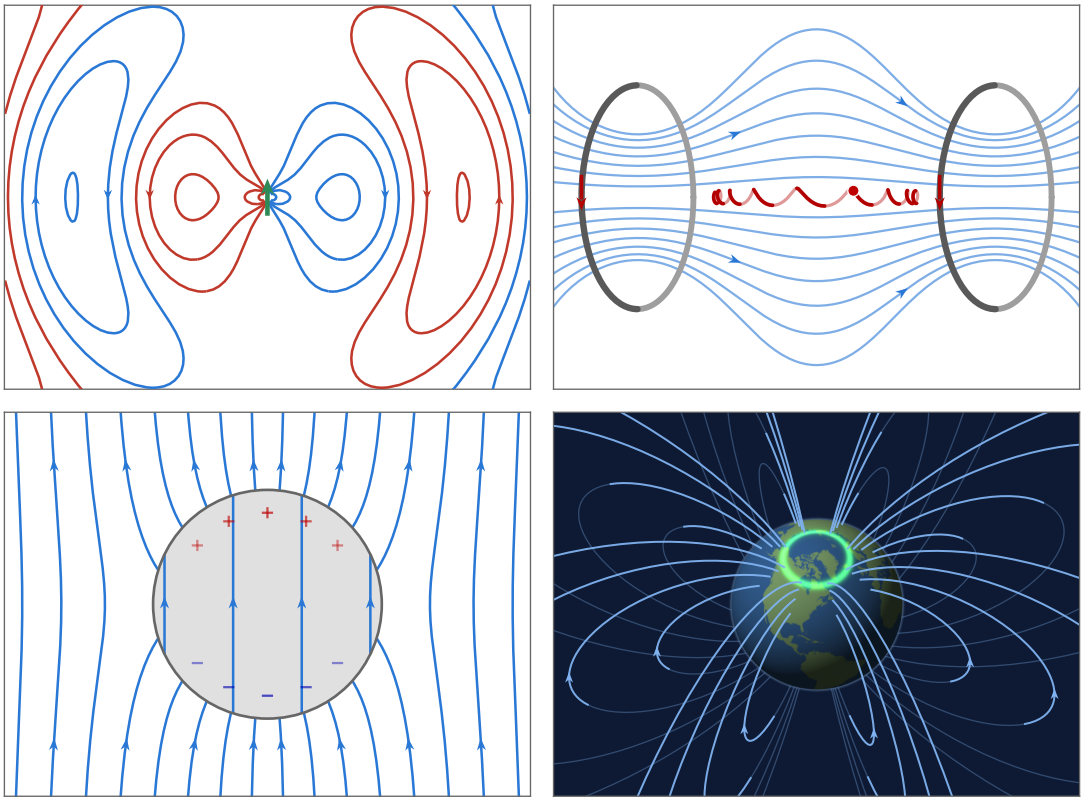
Physics of Electricity and Magnetism



John S. Colton
Brigham Young University

2026 Edition
November 14, 2026

Copyright ©2026 John S. Colton. All rights reserved.

The author retains the copyright to this book. However, the content is available free of charge at em.byu.edu. This book may be downloaded, printed, and distributed freely as long as this copyright notice is included. Any use of a portion of this book's content as part of another work requires the express written permission of the author.

ISBN xxx-x-xxx-xxxxx-x

Preface

This curriculum was originally developed for a two-semester fourth-year undergraduate electricity and magnetism sequence in the Department of Physics and Astronomy at Brigham Young University. Historically, this class has primarily used the well-known *Introduction to Electrodynamics* textbook by David J. Griffiths, and indeed this new book has received a lot of inspiration from that classic work. However, as I repeatedly taught the class I kept thinking of different ways to present some of the topics, and things that should have been included but weren't. I wrote dozens of handouts to supplement the book, and as my page count reached the several hundred level I started feeling a call to create my own textbook. In 2026 resolved to finally do so.

In addition to the new and modified material found in my handouts, I also started using tutorials and using conceptual questions in classes on exams in ways that I felt helped students' understanding, and have incorporated those into this book. I also become more aware of the physics education research and as I prepared this book I made deliberate steps to use what I have learned from that literature, see https://em.byu.edu/per. I have also highlighted and emphasized similarities and differences between the electric and magnetic field equations, and have dedicated an entire chapter for this purpose. Indeed, there are more similarities than one might assume! For more reasons to consider using this book, see https://em.byu.edu/why.

While I retain the copyright, this book is made available free of charge at the web site, https://em.byu.edu. A collection of electronic material related to the text is available, including animations to illustrate key concepts, printable tutorial sheets, and data files needed for some homework problems. This is my contribution toward a future world with free textbooks, inspired by the *Physics of Light and Optics* textbook written by BYU professors Justin Peatross and Michael Ware, see https://optics.byu.edu/textbook. This book also relied heavily on the textbook template that Peatross and Ware provide. I have also been inspired by the profile pictures and historical sketches that they used in their Optics book, and have similarly included historical sketches and pictures for dozens of the pioneers in classical electricity and magnetism to help students appreciate some of this historical context. As I compiled these it was impossible to overlook the overwhelming lack of women and non-American/European people, which is an unfortunate indictment of how these communities have not been given the same opportunities as others in the past; may we all work toward a future in which this

is different.

I may be contacted for textbook-related matters at embook@byu.edu. I enjoy hearing reports from those using the book and welcome constructive feedback, and especially while the book is young I will revise it regularly to fix errors and otherwise improve the text. The title page indicates the date of the last revision.

I would like to thank some individuals who helped teach me the physics of electricity and magnetism, whether in person or through their books, or both. These include: D.J. Griffiths, author of *Introduction to Electrodynamics*, which text has been foundational to me as mentioned above; Howard Vanfleet, who taught my senior undergraduate course in E&M at BYU; Reitz, Milford, and Christy (*Foundations of Electromagnetic Theory*), which book I used as an undergrad; J.D. Jackson, author of *Classical Electrodynamics* and teacher of my graduate course at U.C. Berkeley; and Peter Yu, co-author of *Fundamentals of Semiconductors*, my Ph.D. advisor who taught me a lot of the content of this book in the context of materials physics.

This book was written with extensive help from Claude, an artificial intelligence model made by Anthropic (Claude Opus 5.5 model, 2026). Working from my lecture notes, recorded lectures, handouts, and homework and exam problems, and following outlines and decisions that I made chapter by chapter, Claude drafted much of the text, produced or redrew essentially all of the figures, made all of the animations, helped adapt and write the exercises, researched the historical sketches and checked their sources, checked equations and numerical results by computer, and researched the physics education literature to help improve this book in many other ways. I directed the work and reviewed and revised everything that Claude produced, and take full responsibility for the content of the book.

Lastly, I would like to thank all who have supported me and helped me to improve this material, including but not limited to David Neilsen for writing several tutorials which were the inspiration for ones in this book, Carson Tenney who wrote an early tutorial draft and also helped sort through the complicated complex number conventions and misunderstandings, Michele Eggleston for being a friend and writing accountability partner during the initial stages of this book, and of course Pauline Colton for all of her support through the years, including making me feel that this particular project was a worthwhile idea.

John Colton
Provo, UT
14 Nov 2026

Suggestions for Instructors

These pages give suggestions for instructors on how to use the book's features maximize effective teaching of their electricity & magnetism courses.

Semester break. For a typical two-semester class I recommend covering Chapters 1 through 16 in the first semester and Chapters 18 through 28 in the second. Chapter 17 is an optional chapter on units that could go either semester; it is for those who feel as I do that students should be exposed to different unit systems because as they delve into the physics research literature only $\tilde{5}0\%$ of papers (in my experience) will use SI units. Topics for the first semester then include all manner of 3-dimensional scalar and vector fields: the electric field **E**, charge density ρ, electric potential V, polarization **P**, displacement field **D**, magnetic field **B**, current density **J**, magnetic vector potential **A**, magnetization **M**, "auxiliary field" **H**, and the time dependences found in Faraday's law and Maxwell's correction to Ampere's law. Dividing the book this way makes sure that students who take only the first semester will still see Maxwell's equations in their final form.

Mathematics review chapters. I believe in the ability and importance of physics topics to teach and cement mathematical understanding. For these classes I assume that students are completely familiar with basic algebra and trigonometry as well as differential and integral calculus, but I have included reviews of vector calculus, orthogonal functions, and complex numbers in Chapters 1, 6, and 18, respectively. Students with a strong mathematical background may find these chapters to be mostly review, and can perhaps skip or at least go quickly through that material when it arises. Students with less of a background may be seeing some of these topics for the first time and may require a more in-depth coverage of the mathematics chapters. It is of course left to the instructor to determine how best to use the text for their specific situation.

Sections that can be skipped. Several sections stand on their own: no later section builds on them, so an instructor can include them or skip them as time allows. Examples are the chapter on systems of units (Chapter 17); AC circuits, with complex impedance, power, and filters (Sections 19.5–19.7); forces from the Maxwell stress tensor, with its tutorial, and angular momentum in the fields (Sections 20.4 and 20.5); the Kramers–Kronig relations (Section 22.5.4); phase and group velocity, resonant cavities, and transmission lines (Sections 23.4, 23.7, and 23.8); the Doppler effect for light and the stress–energy tensor (Sections 26.8 and 26.9); and synchrotron radiation and bremsstrahlung (Sections 27.6 and 27.7).

Warm-up questions. Aside from the statics review chapter, Chapter 14, the exercises for each chapter include warm-up questions in a purple box, usually two or three per section, numbered W4.1, W4.2, ..., placed separately from the main exercises (P4.1, P4.2, ...). The questions follow Just-in-Time Teaching:[1] students answer them after reading the section and before it is discussed in class, for credit on effort rather than on correctness, and the instructor reads the answers before class to decide what to spend the class time on. Each box ends by asking students to summarize the section without looking back and to say what they found hardest to understand. Answers to these warm-up questions can be found in the Warm-up Solutions Manual, on the website (made available to students).

Reality checks and why-prompts. Chapters typically include several worked examples. The handwritten-style solutions often include a why-prompt, which asks the reason for a key step and gives the answer in a footnote. They also typically end with a reality check paragraph which tests the result in some way: its units, a limit, a known case, or the same result found another way. I recommend you encourage students to answer each why-prompt before they read its footnote, and to learn to check their own answers to homework problems the way these reality checks model.

Tutorials. The tutorials sit inside the chapters, in the sections whose ideas they practice. Each is meant for students working in small groups and takes about an hour. They can be done in class or assigned as part of homework assignments. A printable version of each tutorial, to hand out in class, is on the book's web site at https://em.byu.edu/tutorials, and the "Printable version" link in the tutorial's title bar within this textbook also opens it.

Exercises and exam reviews. The homework exercises at the end of each chapter are grouped by section. Most chapters end with true/false questions whose statements students are to assess and explain. These can for example, assist with exam reviews — I sometimes assign students to discuss conceptual questions in groups and then submit their answers as a group. Chapter 14 can also be valuable as an exam review tool. It sets the equations of electro- and magnetostatics side by side, with exercises that ask which method a problem needs, and why. In Section 14.3, Figures 14.9 and 14.10 gather the methods of electro- and magnetostatics into "Which method?" flowcharts, and Figure 16.17 does the same for time-varying fields.

Computation. Many chapters have a section on a numerical method, such as direct integration or relaxation, and exercises that need a computer because they have no analytic solution. The text gives short "Tips for Mathematica and Python" rather than full code listings, and students need access to Mathematica, Python, or a similar tool. The data files that some exercises use are at https://em.byu.edu/data. The last exercise of most chapters also asks students to use AI

[1] G. M. Novak, E. T. Patterson, A. D. Gavrin, and W. Christian, *Just-in-Time Teaching: Blending Active Learning with Web Technology* (Prentice Hall, Upper Saddle River, NJ, 1999).

resources to do something computational beyond what they could do on their own, in order to understand one of the chapter's concepts better.

Animations. Some figures have animations on the book's web site. In a figure caption, the word "Animation" is a link that opens the animation, and all of the animations are listed in book order, with their section numbers, at https://em.byu.edu/animations. They can be shown in class, and students can watch them as they read.

Solutions manuals. A solutions manual for instructors which contains written-out solutions for the exercises and tutorials can be requested by email at em-book@byu.edu, under the condition that the solutions are kept private. If you share solutions with students (e.g. after the assignments are due), make sure they also agree to not distribute them to others. As mentioned above, a warm-up solutions manual is also available on the website.

Exam corrections. One practice supported by physics education research which I suggest you consider is to let students earn back part of the credit they lost on an exam problem by turning in a corrected solution, and to release the exam's key only after the corrections are in. In a study of upper-level quantum mechanics courses, students who had this incentive did better on the final exam than students who did not, and it helped most those who had done poorly on the midterm exam.[2]

[2] "Improving performance in quantum mechanics with explicit incentives to correct mistakes," B. R. Brown, A. Mason, and C. Singh, Phys. Rev. Phys. Educ. Res. 12, 010121 (2016). https://doi.org/10.1103/PhysRevPhysEducRes.12.010121

Brief Contents

Preface . . . iii
Suggestions for Instructors . . . v

I Preliminaries **1**
1 Mathematics Review 1: Vector Calculus . . . 3
2 Maxwell's Equations and the Helmholtz Theorem . . . 41

II Electrostatics **49**
3 Coulomb's Law . . . 51
4 Gauss's Law . . . 79
5 Electric Potential and Potential Energy . . . 103
6 Mathematics Review 2: Orthogonal Functions . . . 153
7 Boundary Conditions and Boundary Value Problems . . . 171
8 Vectors, Tensors, and the Multipole Expansion . . . 209
9 Polarization and Dielectrics . . . 231

III Magnetostatics **287**
10 The Lorentz Force and the Biot–Savart Law . . . 289
11 Ampère's Law . . . 333
12 Vector Potential . . . 357
13 Magnetic Materials . . . 401
14 Review of Electro- and Magnetostatics . . . 453

IV Time-Varying Fields **467**
15 Faraday's Law . . . 469
16 The Ampère–Maxwell Law . . . 503

V Units, Circuits, and Conservation Laws **531**
17 Unit Systems . . . 533
18 Mathematics Review 3: Complex Numbers . . . 547
19 Ohm's Law, Induction, and Circuits . . . 567
20 Forces and Conservation Laws . . . 603

VI Electromagnetic Waves **633**
21 Real Waves . . . 635
22 Complex Waves . . . 665
23 Confined Waves . . . 695
24 Inhomogeneous Wave Equations . . . 729

VII Relativity **753**
25 Basic Relativity . . . 755
26 Relativity and Electromagnetic Fields . . . 807

VIII Moving Charges and Radiation **845**
27 Fields of a Point Charge . . . 847
28 Dipole and Multipole Radiation . . . 883

IX Appendices **921**
A Convolutions . . . 923
B Green's Functions? . . . 925
C The Kramers–Kronig Relations . . . 927
Portrait Credits . . . 928
Index . . . 933
Vector Derivatives . . . 940
Physical Constants . . . 941

Contents

Preface iii

Suggestions for Instructors v

I Preliminaries 1

1 Mathematics Review 1: Vector Calculus 3
- 1.1 Coordinate systems . . . 3
  - 1.1.1 Cartesian coordinates . . . 3
  - 1.1.2 Spherical coordinates . . . 4
  - 1.1.3 Cylindrical coordinates . . . 4
  - 1.1.4 Converting between coordinate systems . . . 5
  - 1.1.5 Constant and variable unit vectors . . . 7
- 1.2 Scalar and vector fields . . . 9
- 1.3 Integrating in three dimensions: Jacobians . . . 9
  - 1.3.1 Line elements . . . 9
  - 1.3.2 Volume elements . . . 10
  - 1.3.3 Area elements . . . 10
  - 1.3.4 Choosing coordinates . . . 11
  - 1.3.5 Integrals of vector fields . . . 12
- 1.4 Derivatives in vector calculus: gradient, divergence, curl, Laplacian 13
  - 1.4.1 Gradient . . . 13
  - 1.4.2 Divergence . . . 14
  - 1.4.3 Curl . . . 15
  - 1.4.4 Gauss's law and Ampère's law, conceptually . . . 16
  - 1.4.5 Second derivatives and the Laplacian . . . 17
  - 1.4.6 Cylindrical and spherical coordinates . . . 18
- 1.5 The fundamental theorems . . . 19
  - 1.5.1 Fundamental theorem of calculus . . . 19
  - 1.5.2 Gradient theorem . . . 19
  - 1.5.3 Divergence theorem . . . 21
  - 1.5.4 Curl theorem (Stokes' theorem) . . . 22
  - 1.5.5 The four theorems together . . . 24
- 1.6 The Dirac delta function . . . 24
  - 1.6.1 The delta function in one dimension . . . 24
  - 1.6.2 The delta function in three dimensions . . . 26
  - 1.6.3 Derivatives that the formulas get wrong . . . 27
- 1.7 Vector identities . . . 28

1.8 Summary . . . 30
Exercises . . . 33

2 Maxwell's Equations and the Helmholtz Theorem 41
2.1 What the equations are . . . 41
2.2 What the equations mean . . . 41
2.3 The Helmholtz theorem . . . 42
2.4 A short history of the equations . . . 43
2.5 Summary . . . 44
Exercises . . . 46

II Electrostatics 49

3 Coulomb's Law 51
3.1 Charge and Coulomb's law . . . 51
3.1.1 The fundamental problem, and how we simplify it . . . 52
3.1.2 Coulomb's law . . . 53
3.2 The electric field and superposition . . . 54
3.3 Field vectors and field lines . . . 57
3.4 Continuous charge distributions . . . 58
3.5 Solving for E when given ρ: script-r method . . . 60
3.5.1 Colton's First Commandment . . . 61
3.5.2 Worked examples . . . 62
3.5.3 The uniformly charged spherical shell . . . 66
3.6 Numerical method: direct integration . . . 68
3.7 Summary . . . 70
Exercises . . . 72

4 Gauss's Law 79
4.1 Divergence of Coulomb's law . . . 79
4.2 Statement and physical meaning . . . 82
4.2.1 Counting field lines . . . 82
4.2.2 Gauss's law directly from Coulomb's law . . . 83
4.2.3 Two forms, two uses . . . 83
4.3 Solving for E when given ρ: Gauss's law method (high symmetry) 83
4.3.1 Colton's Second Commandment . . . 84
4.3.2 The three kinds of symmetry . . . 84
4.3.3 Worked examples . . . 85
4.3.4 Superposition . . . 90
4.4 Boundary condition on $E^{\perp}$ . . . 91
4.5 Conductors and induced charges . . . 91
4.5.1 Induced charge . . . 92
4.5.2 The force on a surface charge . . . 93
4.6 Summary . . . 95
Exercises . . . 96

5 Electric Potential and Potential Energy 103
5.1 Curl of Coulomb's law . . . 103
5.2 Boundary condition on $E^{\parallel}$ . . . 105
5.3 Scalar potential and its relation to electric field . . . 106
5.3.1 Conservative fields . . . 106
5.3.2 The reference point . . . 108
5.3.3 Why the potential is useful . . . 110
5.3.4 Continuity of the potential . . . 110
5.4 Potential of a point charge, relationship to potential energy . . . 112
5.4.1 Work and potential energy . . . 113
5.5 Solving for V when given ρ: script-r method . . . 115
5.5.1 Worked examples . . . 117
5.5.2 Choosing a route . . . 123
5.6 Numerical method: direct integration . . . 123
5.7 Conductors as equipotentials . . . 125
5.8 Capacitance . . . 127
5.9 Energy stored in the field . . . 130
5.9.1 The energy of a collection of point charges . . . 130
5.9.2 Continuous charge distributions . . . 130
5.9.3 Self-energy . . . 133
5.9.4 The energy of a capacitor . . . 133
5.10 Laplace's and Poisson's equations . . . 134
5.11 Summary: Triangle diagram of ρ, **E**, V . . . 135
Exercises . . . 138

6 Mathematics Review 2: Orthogonal Functions 153
6.1 Sines and cosines: Fourier series . . . 153
6.1.1 The series . . . 153
6.1.2 Orthogonality . . . 154
6.1.3 Fourier's trick . . . 155
6.1.4 Symmetry shortcuts . . . 156
6.1.5 Worked example: the square wave . . . 156
6.1.6 Other notations you will meet . . . 157
6.2 Legendre polynomials . . . 158
6.2.1 The polynomials . . . 158
6.2.2 Orthogonality . . . 159
6.2.3 Comparison with sines and cosines . . . 160
6.2.4 Colton's trick . . . 161
6.2.5 A connection to come . . . 161
6.3 Bessel functions . . . 162
6.3.1 Orthogonality, with a weight . . . 163
6.4 Summary of similarities and differences . . . 164
6.5 Summary . . . 165
Exercises . . . 166

7 Boundary Conditions and Boundary Value Problems 171
7.1 Boundary conditions and uniqueness theorems 171
7.1.1 Boundary conditions 172
7.1.2 Laplace's equation and averages 172
7.1.3 The first uniqueness theorem 174
7.1.4 The second uniqueness theorem 175
7.2 Method of images 177
7.2.1 A point charge above a grounded plane 177
7.2.2 The induced surface charge 178
7.2.3 Other image problems 179
7.3 Numerical solutions: the relaxation method 179
7.3.1 Laplace's equation on a grid 179
7.3.2 The method 180
7.3.3 An example, and a caution about the tolerance 181
7.4 Separation of variables 182
7.4.1 2d Cartesian coordinates 182
7.4.2 3d Cartesian coordinates 187
7.4.3 Spherical coordinates, no ϕ dependence 191
7.4.4 Cylindrical coordinates, no z dependence 199
7.5 Summary 201
Exercises 203

8 Vectors, Tensors, and the Multipole Expansion 209
8.1 What is a scalar? What is a vector? 209
8.1.1 The transformation rule, written out 210
8.2 Generalization—tensors 210
8.2.1 Index notation 211
8.2.2 How tensors transform 211
8.3 Application: the multipole expansion of V 212
8.3.1 What the expansion says, before any algebra 212
8.3.2 The mathematics of the expansion 214
8.3.3 What the expansion is for 216
8.3.4 The first two terms 216
8.3.5 Where θ is measured from 217
8.4 Worked example: how good is the dipole approximation? 218
8.5 The electric quadrupole moment tensor 219
8.5.1 Two of the components, written out 221
8.5.2 The pattern continues 221
8.5.3 A note on conventions 222
8.6 Summary 223
Exercises 225

9 Polarization and Dielectrics 231
9.1 Dipoles again 232
9.1.1 Torque on a dipole 232

9.1.2 Energy of a dipole in a field . . . 232
9.1.3 Force on a dipole . . . 233
9.1.4 Polar molecules . . . 234
9.1.5 Induced dipoles: polarizability . . . 235
9.2 Dielectric materials: polarization and susceptibility . . . 237
9.2.1 Polarization . . . 237
9.2.2 How a material becomes polarized . . . 238
9.2.3 Linear dielectrics and the susceptibility . . . 239
9.3 Bound charges . . . 240
9.3.1 What the bound charges are . . . 241
9.4 Solving for E when given P: bound-charge method . . . 242
9.4.1 The field inside a dielectric . . . 245
9.5 The D field and permittivity . . . 248
9.5.1 Linear dielectrics . . . 249
9.6 Gauss's law for D . . . 249
9.6.1 D is not determined by the free charge alone . . . 250
9.6.2 Bound charge in a linear dielectric . . . 252
9.6.3 What the dielectric constant measures . . . 252
9.7 Solving for E when given P: Gauss's law for D method (high symmetry) . . . 253
9.7.1 Linear dielectrics, when the free charge is given . . . 254
9.8 Boundary conditions at dielectric interfaces . . . 256
9.8.1 The perpendicular component of D . . . 257
9.8.2 The parallel components . . . 257
9.8.3 Interfaces between linear dielectrics . . . 257
9.8.4 Boundary-value problems with dielectrics . . . 259
9.9 Numerical methods: direct integration and relaxation . . . 261
9.9.1 Direct integration over the bound charges . . . 261
9.9.2 Relaxation with a dielectric . . . 262
9.10 Energy stored in the field, inside dielectrics . . . 264
9.10.1 Forces on dielectrics . . . 265
9.11 Summary . . . 266
Exercises . . . 268

III Magnetostatics 287

10 The Lorentz Force and the Biot–Savart Law 289
10.1 Forces on moving charges: the Lorentz force law . . . 289
10.1.1 What the experiments show . . . 289
10.1.2 The Lorentz force law . . . 291
10.1.3 Magnetic forces do no work . . . 294
10.1.4 Motion in a uniform field . . . 295
10.1.5 Numerical method: trajectories . . . 297
10.2 Why the force and the field need cross products . . . 298

10.3 The magnetic field and field lines . . . 299
10.4 Currents and current densities . . . 300
10.4.1 Steady currents . . . 303
10.4.2 Forces on currents . . . 303
10.5 The Biot–Savart law . . . 304
10.6 Solving for B when given J: script-r method . . . 306
10.6.1 Worked examples . . . 307
10.6.2 Tips for Mathematica and Python . . . 312
10.7 Numerical method: direct integration . . . 313
10.8 Summary . . . 314
Exercises . . . 317

11 Ampère's Law **333**
11.1 Ampère's law . . . 333
11.1.1 The divergence of B . . . 333
11.1.2 The curl of B . . . 335
11.1.3 The integral form . . . 336
11.1.4 Maxwell's equations for static fields . . . 337
11.2 Solving for B when given J: Ampère's law method (high symmetry) 338
11.2.1 Colton's Third Commandment . . . 339
11.2.2 The four geometries . . . 339
11.2.3 Worked examples . . . 340
11.2.4 Superposition . . . 346
11.3 Boundary conditions on B . . . 348
11.4 Summary . . . 349
Exercises . . . 351

12 Vector Potential **357**
12.1 The magnetic vector potential . . . 357
12.1.1 Gauge freedom and the Coulomb gauge . . . 358
12.1.2 Circulation of A and magnetic flux . . . 361
12.1.3 Continuity of the vector potential . . . 361
12.2 Poisson's equation for **A** . . . 362
12.2.1 The components of A behave like V . . . 363
12.2.2 The script-r integral for A . . . 364
12.3 Solving for A when given J: script-r method . . . 365
12.3.1 Worked examples . . . 366
12.3.2 Tips for Mathematica and Python . . . 370
12.4 Solving for A when given B: flux method (high symmetry) . . . 370
12.5 Numerical method: relaxation for A in two dimensions . . . 375
12.6 Multipole expansion of A; the magnetic dipole . . . 377
12.6.1 The monopole term is zero . . . 378
12.6.2 The dipole term . . . 378
12.6.3 The field of a magnetic dipole . . . 379
12.6.4 Near a dipole . . . 382

12.7 Inductance and the energy stored in magnetic fields 383
12.7.1 Inductance 383
12.7.2 Energy stored in the magnetic field 384
12.8 Summary: Triangle diagram of **J**, **B**, **A** 386
Exercises 389

13 Magnetic Materials **401**
13.1 Dipoles again 401
13.1.1 Torque on a dipole 402
13.1.2 Energy of a dipole in a field 402
13.1.3 Force on a dipole 403
13.1.4 Where atomic moments come from 405
13.2 Magnetic materials: magnetization and susceptibility 405
13.2.1 Magnetization 406
13.2.2 How a material becomes magnetized 406
13.3 Bound currents 408
13.3.1 What the bound currents are 410
13.3.2 The total bound current 410
13.4 Solving for B when given M: bound-current method 411
13.5 The H field and permeability 415
13.5.1 Linear magnetic materials 415
13.6 Ampère's law for H 416
13.6.1 Gauss's law for H 417
13.6.2 H is not determined by the free current alone 418
13.6.3 Bound current in a linear material 420
13.6.4 What the permeability measures 420
13.7 Solving for H when given M: Ampère's law for H method (high symmetry) 420
13.8 Ferromagnets: domains, hysteresis, and the Curie temperature . 424
13.8.1 Domains 425
13.8.2 Hysteresis 425
13.8.3 The Curie temperature 426
13.8.4 Antiferromagnets and ferrimagnets 426
13.9 Boundary conditions at magnetic interfaces 427
13.9.1 Interfaces between linear materials 427
13.9.2 Boundary-value problems with magnetic materials 429
13.10 Numerical methods: direct integration and relaxation 431
13.10.1 Direct integration over the bound currents 431
13.10.2 Relaxation with magnetic materials 431
13.11 Energy stored in the field, inside magnetic materials 433
13.11.1 Energy lost to hysteresis 435
13.11.2 A magnet lifting a nail 435
13.12 Summary 436
Exercises 438

14 Review of Electro- and Magnetostatics **453**
14.1 The two columns side by side . . . 453
14.1.1 Statics . . . 453
14.1.2 In materials . . . 455
14.1.3 Dynamics . . . 457
14.2 What the pattern is telling you . . . 457
14.2.1 The integrals are the same integral . . . 457
14.2.2 Divergence and curl trade places . . . 458
14.2.3 Boundary conditions swap . . . 459
14.2.4 The one sign that is not symmetric . . . 459
14.2.5 Time dependence: similarities and differences . . . 459
14.3 Which method? . . . 460
Exercises . . . 462

IV Time-Varying Fields **467**

15 Faraday's Law **469**
15.1 Motional EMF . . . 469
15.1.1 Electromotive force . . . 469
15.1.2 A conducting bar moving through a field . . . 470
15.1.3 The flux rule . . . 472
15.2 Faraday's law and the Faraday current . . . 474
15.2.1 Faraday's experiments . . . 474
15.2.2 Lenz's law . . . 476
15.2.3 The differential form . . . 477
15.2.4 The Faraday current . . . 478
15.2.5 Eddy currents . . . 479
15.3 Solving for E when given a changing B: Faraday-current method (high symmetry) . . . 480
15.3.1 Worked examples . . . 482
15.4 Fields from potentials: time dependence . . . 485
15.5 Numerical method: a magnet falling through a coil . . . 487
15.5.1 The model . . . 488
15.5.2 Tips for Mathematica and Python . . . 489
15.6 Summary . . . 490
Exercises . . . 493

16 The Ampère–Maxwell Law **503**
16.1 Displacement current and the Ampère–Maxwell law . . . 503
16.1.1 A contradiction: the charging capacitor . . . 503
16.1.2 Maxwell's term . . . 504
16.1.3 The displacement current . . . 505
16.1.4 In materials . . . 506

16.2 Solving for B when given a changing E: displacement-current method (high symmetry) . . . 507
16.3 How the displacement current produces magnetic fields . . . 511
16.3.1 The displacement current of a charging capacitor . . . 511
16.3.2 The real currents make the field . . . 513
16.3.3 When the displacement current does contribute . . . 514
16.4 When is fast, when is slow . . . 516
16.4.1 The size of the corrections . . . 516
16.4.2 Successive approximations . . . 517
16.5 The prediction of waves . . . 519
16.6 Summary . . . 521
Exercises . . . 523

V Units, Circuits, and Conservation Laws 531

17 Unit Systems 533
17.1 SI versus Gaussian units . . . 533
17.1.1 Where base units come from . . . 533
17.1.2 Two attempts, both of which fail . . . 534
17.1.3 The vicious cycle . . . 534
17.1.4 How SI breaks the cycle . . . 535
17.1.5 How Gaussian units break it instead . . . 535
17.1.6 What follows from the Gaussian choice . . . 536
17.1.7 Why **E** and **B** end up with the same units . . . 536
17.1.8 Three translation rules . . . 537
17.2 Maxwell's equations in Gaussian units . . . 537
17.2.1 Another example: energy stored in the fields . . . 538
17.2.2 Why anyone still uses Gaussian units . . . 538
17.3 Atomic units and other systems used in research . . . 538
17.3.1 Hartree atomic units . . . 539
17.3.2 Natural units, and others . . . 540
17.4 The 2019 change to SI units . . . 541
17.5 Summary . . . 542
Exercises . . . 543

18 Mathematics Review 3: Complex Numbers 547
18.1 Complex numbers . . . 547
18.1.1 Rectangular and polar forms . . . 547
18.1.2 Euler's formula . . . 548
18.1.3 The complex conjugate . . . 549
18.1.4 Roots . . . 549
18.2 Oscillations and waves as complex exponentials . . . 550
18.2.1 Phase and the complex amplitude . . . 550
18.2.2 Adding oscillations . . . 551

18.2.3 The driven oscillator . . . 552
18.2.4 Waves . . . 554
18.3 Complex wave numbers . . . 555
18.4 Products and time averages . . . 555
18.5 The two sign conventions . . . 557
18.5.1 Tips for Mathematica and Python . . . 558
18.6 Summary . . . 558
Exercises . . . 560

19 Ohm's Law, Induction, and Circuits 567
19.1 The Ampère–Maxwell law in matter . . . 567
19.1.1 The polarization current . . . 567
19.1.2 Maxwell's equations in matter . . . 569
19.1.3 Boundary conditions for changing fields . . . 569
19.2 Ohm's law and conductivity . . . 571
19.2.1 The field inside a wire . . . 572
19.2.2 Resistance . . . 572
19.2.3 Ohmic heating . . . 572
19.2.4 Numerical solutions: current in a conductor of any shape 573
19.3 Microscopic view of conduction . . . 574
19.4 Inductance . . . 576
19.4.1 The inductor in a circuit . . . 576
19.4.2 The energy stored in an inductor . . . 577
19.4.3 Two circuits . . . 578
19.4.4 The transformer . . . 579
19.5 Complex impedance . . . 579
19.5.1 Circuits with steady currents . . . 580
19.5.2 Complex amplitudes and impedance . . . 580
19.5.3 Solving a circuit with impedances . . . 582
19.6 Power in AC circuits . . . 585
19.7 Transfer functions and filters . . . 587
19.7.1 The voltage divider . . . 587
19.7.2 High-pass and low-pass filters . . . 587
19.7.3 The band-pass filter . . . 589
19.8 Summary . . . 589
Exercises . . . 592

20 Forces and Conservation Laws 603
20.1 Conservation of charge: the continuity equation . . . 603
20.2 Energy conservation: Poynting's theorem . . . 605
20.2.1 The work done on charges . . . 605
20.2.2 Poynting's theorem . . . 606
20.2.3 Energy flow into a wire . . . 607
20.2.4 Time averages . . . 609
20.2.5 In linear materials . . . 609

20.3 Momentum conservation: the Maxwell stress tensor 610
20.3.1 Newton's third law and moving charges 610
20.3.2 The force on charges, in terms of the fields 611
20.3.3 The Maxwell stress tensor 612
20.3.4 Momentum in the fields 614
20.4 Forces from the stress tensor 616
20.4.1 Pressure on a surface 616
20.4.2 Forces between charges 616
20.5 Angular momentum in the fields 621
20.6 Summary 623
Exercises 625

VI Electromagnetic Waves 633

21 Real Waves 635
21.1 The wave equation in vacuum and in matter 635
21.2 Sinusoidal waves 636
21.2.1 Reflection and transmission on a string 638
21.3 Polarization 640
21.4 Plane electromagnetic waves 642
21.4.1 The impedance of a material 643
21.5 Energy and momentum in waves 644
21.5.1 Momentum and radiation pressure 645
21.6 Reflection and transmission at normal incidence 647
21.6.1 Reflection from a mirror 650
21.7 Oblique incidence: the laws of optics and the Fresnel equations . 650
21.7.1 The laws of reflection and refraction 651
21.7.2 The Fresnel equations 652
21.8 Summary 655
Exercises 657

22 Complex Waves 665
22.1 Total internal reflection 665
22.1.1 The evanescent wave 666
22.1.2 The reflected wave and the energy 667
22.2 Waves in conductors 670
22.2.1 Free charge and the damped wave equation 670
22.2.2 The complex wave number and the skin depth 671
22.2.3 The fields in a conductor 673
22.3 Reflection from a conductor 674
22.4 Complex permittivity and complex conductivity 676
22.4.1 Polarization that lags the field 676
22.4.2 Conduction at high frequencies 677
22.4.3 One current, two descriptions 678

22.5 The Lorentz oscillator model . . . 679
22.5.1 Bound electrons . . . 679
22.5.2 Dispersion . . . 680
22.5.3 Free electrons and the plasma frequency . . . 681
22.5.4 Causality and the Kramers–Kronig relations . . . 684
22.6 Summary . . . 685
Exercises . . . 687

23 Confined Waves **695**
23.1 Boundary conditions at a perfect conductor . . . 695
23.2 Waves in a hollow guide . . . 696
23.2.1 The longitudinal-to-transverse equations . . . 697
23.2.2 TE, TM, and TEM modes . . . 699
23.3 Rectangular waveguides . . . 700
23.3.1 The TE modes and their cutoff frequencies . . . 702
23.3.2 The TE_{10} mode . . . 703
23.3.3 Below cutoff . . . 704
23.4 Phase velocity and group velocity . . . 706
23.4.1 Beats and the group velocity . . . 706
23.4.2 Group velocity in a material . . . 707
23.5 Cylindrical waveguides . . . 708
23.5.1 TM modes . . . 709
23.6 Numerical methods: modes of a guide of any shape . . . 710
23.7 Resonant cavities . . . 712
23.7.1 The resonant frequencies . . . 713
23.7.2 Losses and the quality factor . . . 715
23.8 Transmission lines . . . 715
23.8.1 The TEM mode . . . 715
23.8.2 The telegrapher's equations . . . 716
23.8.3 Pulses and reflections . . . 717
23.9 Summary . . . 718
Exercises . . . 721

24 Inhomogeneous Wave Equations **729**
24.1 The Lorenz gauge . . . 729
24.1.1 Maxwell's equations for the potentials . . . 730
24.1.2 The Coulomb gauge again . . . 730
24.1.3 The Lorenz condition . . . 731
24.2 The wave equations for V and A . . . 732
24.3 Retarded time . . . 732
24.3.1 Apparent shapes . . . 735
24.4 Retarded potentials . . . 735
24.5 Jefimenko's equations . . . 738
24.6 Numerical methods: retarded times and integrals . . . 740
24.6.1 Retarded times . . . 741

24.6.2 Retarded integrals . . . 741
24.6.3 Tips for Mathematica and Python . . . 741
24.7 Summary: triangle diagram of sources, fields, and potentials . . . 743
Exercises . . . 746

VII Relativity 753

25 Basic Relativity 755
25.1 The postulates . . . 756
25.1.1 Galilean relativity . . . 756
25.1.2 The speed of light . . . 756
25.1.3 Einstein's postulates . . . 757
25.2 Events, world lines, and spacetime diagrams . . . 759
25.3 Simultaneity, time dilation, and length contraction . . . 760
25.3.1 Simultaneity . . . 760
25.3.2 Time dilation . . . 761
25.3.3 Length contraction . . . 763
25.4 The Lorentz transformations . . . 764
25.5 The structure of spacetime . . . 768
25.5.1 The invariant interval . . . 768
25.5.2 Cause and effect . . . 769
25.5.3 Reading a spacetime diagram . . . 769
25.5.4 The twin paradox . . . 770
25.6 Velocity addition . . . 772
25.7 Four-vectors . . . 774
25.7.1 The scalar product . . . 775
25.7.2 The summation convention . . . 775
25.7.3 Four-tensors . . . 776
25.8 Proper time and four-velocity . . . 776
25.9 Four-acceleration and the instantaneous rest frame . . . 777
25.10 Energy and momentum . . . 780
25.11 Collisions . . . 782
25.12 Relativistic dynamics . . . 786
25.13 Numerical method: relativistic trajectories . . . 788
25.13.1Tips for Mathematica and Python . . . 789
25.14 Summary . . . 791
Exercises . . . 794

26 Relativity and Electromagnetic Fields 807
26.1 Magnetism from relativity . . . 807
26.2 The transformation of E and B . . . 809
26.2.1 Moving capacitor plates: the electric field . . . 809
26.2.2 Moving plates and a moving solenoid: the magnetic field 810
26.2.3 The transformation rules . . . 811

26.2.4 The magnet and the conductor . . . 813
26.3 The field tensor . . . 814
26.4 The four-current and the four-gradient . . . 816
26.5 Maxwell's equations in tensor form . . . 817
26.6 The four-potential . . . 818
26.6.1 Proof that the retarded potentials form a four-vector . . . 819
26.7 The fields of a uniformly moving charge . . . 823
26.8 The Doppler effect for light . . . 826
26.9 The stress–energy tensor . . . 828
26.10 Numerical method: motion in crossed fields . . . 829
26.11 Summary . . . 831
Exercises . . . 834

VIII Moving Charges and Radiation 845

27 Fields of a Point Charge 847
27.1 The Liénard–Wiechert potentials . . . 847
27.1.1 Which velocity? . . . 849
27.2 The fields of a point charge . . . 851
27.2.1 The velocity field . . . 852
27.2.2 The acceleration field and Thomson's kink . . . 853
27.3 What is radiation? . . . 855
27.4 Larmor's formula . . . 856
27.5 Liénard's generalization . . . 859
27.6 Synchrotron radiation . . . 861
27.7 Bremsstrahlung . . . 863
27.7.1 The spectrum . . . 864
27.8 Numerical method: fields of an arbitrarily moving charge . . . 865
27.8.1 Tips for Mathematica and Python . . . 867
27.9 Summary . . . 868
Exercises . . . 871

28 Dipole and Multipole Radiation 883
28.1 Radiation from small sources: power and angular distribution . . . 883
28.2 Electric dipole radiation . . . 885
28.2.1 The potentials . . . 885
28.2.2 The fields and the power . . . 887
28.2.3 Radiation resistance . . . 890
28.2.4 Why the sky is blue . . . 892
28.3 Magnetic dipole radiation . . . 893
28.4 Radiation from an arbitrary small source: the multipole expansion 895
28.4.1 The fields far from a small source . . . 896
28.4.2 The next order . . . 898
28.5 Numerical methods: antennas of any length . . . 905

28.5.1 The far field of a given current 905
28.5.2 Arrays of antennas 907
28.5.3 Tips for Mathematica and Python 907
28.6 Summary 908
Exercises 911

IX Appendices **921**

A Convolutions **923**

B Green's Functions? **925**

C The Kramers–Kronig Relations **927**

Portrait Credits **928**

Index **933**

Vector Derivatives **940**

Physical Constants **941**

Part I

Preliminaries

Mathematics is an experimental science, and definitions do not come first, but later on. They make themselves, when the nature of the subject has developed itself.

— Oliver Heaviside,
"On Operations in Physical Mathematics, Part II" (1893)*

*"On operations in physical mathematics. Part II," O. Heaviside, Proc. R. Soc. Lond. 54, 105–143 (1893–94), p. 121. https://doi.org/10.1098/rspl.1893.0059.

Chapter 1

Mathematics Review 1: Vector Calculus

Vector calculus is in large part the study of functions of three spatial variables, which we will call x, y, and z and typically refer to as the vector $\mathbf{r} = (x, y, z) = x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}}$. Here we have used the traditional "hat" notation to refer to unit vectors: e.g. $\hat{\mathbf{x}} = (1, 0, 0)$, a dimensionless[1] vector in the direction of the x axis.

This chapter encompasses the vector calculus that the rest of the book uses: coordinate systems; line, surface, and volume integrals; the derivatives called the gradient, divergence, curl, and Laplacian, and what each one means physically; the theorems that connect derivatives to integrals; the Dirac delta function; and a set of identities. Chapter 2 then presents Maxwell's equations, which are written in the language of vector calculus and are the primary subject of the rest of this book.

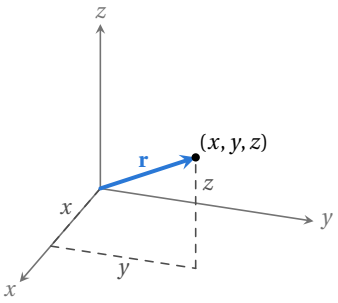


Figure 1.1 The point $\mathbf{r}$, depicted using Cartesian coordinates.

1.1 Coordinate systems

There are many ways of referring to a point in 3-dimensional space, $\mathbf{r}$. The symmetry of the situation usually determines which coordinate system to use. The three common coordinate systems we will use in this book are Cartesian, spherical, and cylindrical, but others exist. In the three-dimensional figures of this book the z axis points up, the y axis to the right, and the x axis out of the page, toward the lower left.

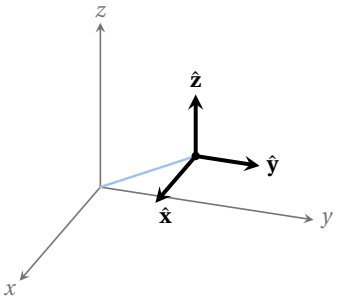


Figure 1.2 The unit vectors of Cartesian coordinates, depicted at the location of $\mathbf{r}$.

1.1.1 Cartesian coordinates

The coordinate system used above, namely

$$\mathbf{r} = (x, y, z) = x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}}, \qquad \text{(Cartesian coordinates)} \quad (1.1)$$

is called "Cartesian" or "rectangular" coordinates (Figure 1.1). Note that the unit vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ point in the directions of increasing x, y, and z (Figure 1.2).

[1] Somewhat ironically, the *unit* vectors have no *units*.

This is true of all coordinate systems. The equation simply says, "To reach the point $\mathbf{r}$ from the origin, you move a distance x in the $\hat{\mathbf{x}}$ direction, a distance y in the $\hat{\mathbf{y}}$ direction, and a distance z in the $\hat{\mathbf{z}}$ direction." When listed in this order, $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ form a right-handed coordinate system, which means that $\hat{\mathbf{x}} \times \hat{\mathbf{y}} = \hat{\mathbf{z}}$, $\hat{\mathbf{y}} \times \hat{\mathbf{z}} = \hat{\mathbf{x}}$, and $\hat{\mathbf{z}} \times \hat{\mathbf{x}} = \hat{\mathbf{y}}$.

1.1.2 Spherical coordinates

In **spherical coordinates**, the location is specified by the symbols r (distance from the origin), θ (polar angle, i.e. angle from the positive z axis), and ϕ (azimuthal angle, i.e. angle of the projection of $\mathbf{r}$ in the xy plane from the positive x axis), as shown in Figure 1.3.[2] The angle θ runs from 0 (the positive z axis) to π (the negative z axis), and ϕ runs from 0 to 2π. The projection of $\mathbf{r}$ onto the xy plane has length $r\sin\theta$, so

$$x = r\sin\theta\cos\phi, \qquad y = r\sin\theta\sin\phi, \qquad z = r\cos\theta. \tag{1.2}$$

Just as x, y, and z each have a unit vector associated with them, so do r, θ, and ϕ: $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\phi}}$ (Figure 1.4). They point in the directions of increasing r, θ, and ϕ, and also form a right-handed coordinate system when listed in that order: $\hat{\mathbf{r}} \times \hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\phi}}$, $\hat{\boldsymbol{\theta}} \times \hat{\boldsymbol{\phi}} = \hat{\mathbf{r}}$, and $\hat{\boldsymbol{\phi}} \times \hat{\mathbf{r}} = \hat{\boldsymbol{\theta}}$. To find the direction of one of them, imagine increasing that coordinate alone and watch which way the point moves. Increasing r moves the point directly away from the origin, so $\hat{\mathbf{r}}$ points outward. Increasing θ swings the point down, away from the positive z axis, along a circle through the two poles, so $\hat{\boldsymbol{\theta}}$ is tangent to that circle and points away from the positive z axis. And increasing ϕ moves the point horizontally around the z axis, so $\hat{\boldsymbol{\phi}}$ is tangent to a horizontal circle.

However, whereas the point $\mathbf{r}$ is equal to $x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}}$, it is *not* equal to $r\hat{\mathbf{r}} + \theta\hat{\boldsymbol{\theta}} + \phi\hat{\boldsymbol{\phi}}$. Indeed, that last expression does not make sense because r has units of distance whereas θ and ϕ are dimensionless. Instead we more simply have

$$\mathbf{r} = r\hat{\mathbf{r}}. \qquad \text{(Spherical coordinates)} \tag{1.3}$$

This is commonly misunderstood by students, so make sure you can visualize and understand why going a distance r in the $\hat{\mathbf{r}}$ direction takes you from the origin to the point $\mathbf{r}$. No $\hat{\boldsymbol{\theta}}$ or $\hat{\boldsymbol{\phi}}$ components are needed.

1.1.3 Cylindrical coordinates

In **cylindrical coordinates**, the location is specified by the symbols s (distance from the z axis), ϕ (azimuthal angle, as in spherical coordinates), and z (the usual

☞ *Caution*: Physics and mathematics use the symbols θ and ϕ differently. Physicists, and the international standard ISO 80000-2, use θ for the polar angle, measured from the z axis, and ϕ for the azimuthal angle. Many mathematics texts, especially American calculus books, swap the two, so that θ stays the angle in the xy plane, as it is in two-dimensional polar coordinates. This being a physics textbook, I shall use the physics convention.

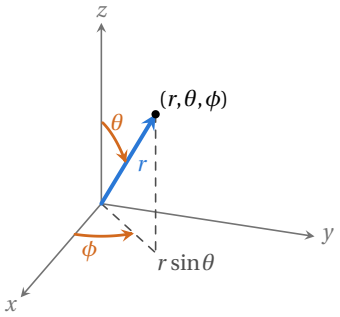


Figure 1.3 The point $\mathbf{r}$, depicted using spherical coordinates.

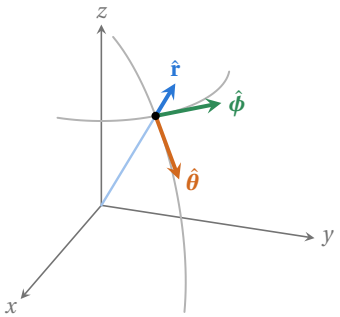


Figure 1.4 The unit vectors of spherical coordinates, depicted at the location of $\mathbf{r}$. Gray: the circles of constant ϕ and of constant θ through the point.

[2] On where the two conventions for θ and ϕ came from: the physics usage is the older one. Laplace's *Mécanique céleste* (1799) already used θ for the polar angle, in his equation for the gravitational potential written in spherical coordinates, and Maxwell's *Treatise on Electricity and Magnetism* (1873) uses θ and ϕ just as we do. The current physics convention very likely grew out of the early work on differential equations and celestial mechanics. The current mathematics usage keeps θ as the angle in the xy plane, as in two-dimensional polar coordinates.

z-coordinate), as shown in Figure 1.5. In the xy plane, s and ϕ are just the usual polar coordinates:

$$x = s\cos\phi, \qquad y = s\sin\phi, \qquad z = z. \tag{1.4}$$

The coordinates each have unit vectors associated with them, namely $\hat{\mathbf{s}}$, $\hat{\boldsymbol{\phi}}$, and $\hat{\mathbf{z}}$ (Figure 1.6). They point in the directions of increasing s, ϕ, and z, and also form a right-handed coordinate system, $\hat{\mathbf{s}} \times \hat{\boldsymbol{\phi}} = \hat{\mathbf{z}}$. The unit vector $\hat{\mathbf{s}}$ points directly away from the z axis, and $\hat{\boldsymbol{\phi}}$ is the same as in spherical coordinates.

Again, note that the point $\mathbf{r}$ is *not* equal to $s\hat{\mathbf{s}} + \phi\hat{\boldsymbol{\phi}} + z\hat{\mathbf{z}}$, among other reasons because s and z have units of distance whereas ϕ is dimensionless. Instead we have

$$\mathbf{r} = s\hat{\mathbf{s}} + z\hat{\mathbf{z}}. \qquad \text{(Cylindrical coordinates)} \tag{1.5}$$

Again, make sure you can visualize and understand why going a distance s in the $\hat{\mathbf{s}}$ direction along with a distance z in the $\hat{\mathbf{z}}$ direction takes you from the origin to the point $\mathbf{r}$. No $\hat{\boldsymbol{\phi}}$ component is needed. The angle ϕ has not disappeared, though: it is hidden in $\hat{\mathbf{s}}$, whose direction depends on ϕ.

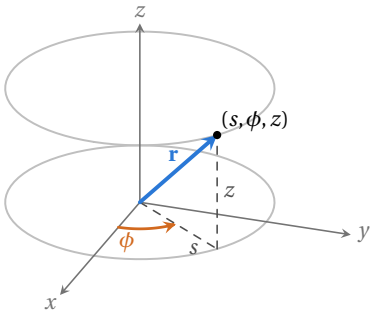


Figure 1.5 The point $\mathbf{r}$, depicted using cylindrical coordinates. The gray circles have radius s.

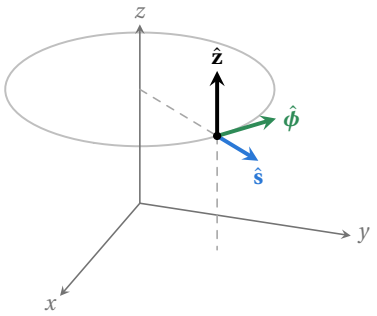


Figure 1.6 The unit vectors of cylindrical coordinates, depicted at the location of $\mathbf{r}$.

1.1.4 Converting between coordinate systems

Equations (1.2) and (1.4) give x, y, and z in terms of the spherical and cylindrical coordinates. Going the other way,

$$r = \sqrt{x^2+y^2+z^2}, \qquad s = \sqrt{x^2+y^2} = r\sin\theta, \qquad \theta = \cos^{-1}\left(\frac{z}{r}\right), \tag{1.6}$$

and ϕ is the angle whose cosine is x/s and whose sine is y/s. Two cautions are in order. First, θ runs from 0 to π, and so does the inverse cosine, so $\cos^{-1}(z/r)$ always gives the right angle. Formulas such as $\theta = \tan^{-1}(s/z)$, which you will find in some tables, give a negative angle when $z < 0$, and you then have to add π. Second, ϕ runs from 0 to 2π, but $\tan^{-1}(y/x)$ returns an angle between $-\pi/2$ and $\pi/2$: for points with $x < 0$ you have to add π, and for points with $x > 0$ and $y < 0$ you have to add 2π. Most programming languages provide a two-argument inverse tangent that gets the quadrant right, such as `atan2(y, x)` in Python or `ArcTan[x, y]` in Wolfram Mathematica (note the order of the arguments). These return an angle between $-\pi$ and π, so add 2π to a negative result if you want the equivalent ϕ between 0 and 2π.

The unit vectors of spherical and cylindrical coordinates can be written in terms of $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$:

$$\begin{aligned}
\hat{\mathbf{r}} &= \sin\theta\cos\phi\,\hat{\mathbf{x}} + \sin\theta\sin\phi\,\hat{\mathbf{y}} + \cos\theta\,\hat{\mathbf{z}},\\
\hat{\boldsymbol{\theta}} &= \cos\theta\cos\phi\,\hat{\mathbf{x}} + \cos\theta\sin\phi\,\hat{\mathbf{y}} - \sin\theta\,\hat{\mathbf{z}},\\
\hat{\boldsymbol{\phi}} &= -\sin\phi\,\hat{\mathbf{x}} + \cos\phi\,\hat{\mathbf{y}},\\
\hat{\mathbf{s}} &= \cos\phi\,\hat{\mathbf{x}} + \sin\phi\,\hat{\mathbf{y}}.
\end{aligned} \tag{1.7}$$

These follow directly from the definition of each unit vector as the direction in which the point moves when one coordinate increases, as the derivation below shows. The inverse relations are

$$\begin{aligned}\hat{\mathbf{x}} &= \sin\theta\cos\phi\,\hat{\mathbf{r}} + \cos\theta\cos\phi\,\hat{\boldsymbol{\theta}} - \sin\phi\,\hat{\boldsymbol{\phi}} = \cos\phi\,\hat{\mathbf{s}} - \sin\phi\,\hat{\boldsymbol{\phi}},\\ \hat{\mathbf{y}} &= \sin\theta\sin\phi\,\hat{\mathbf{r}} + \cos\theta\sin\phi\,\hat{\boldsymbol{\theta}} + \cos\phi\,\hat{\boldsymbol{\phi}} = \sin\phi\,\hat{\mathbf{s}} + \cos\phi\,\hat{\boldsymbol{\phi}},\\ \hat{\mathbf{z}} &= \cos\theta\,\hat{\mathbf{r}} - \sin\theta\,\hat{\boldsymbol{\theta}}.\end{aligned} \tag{1.8}$$

Exercise P1.3 shows how to obtain them. All of these relations between coordinates and unit vectors are compiled in the chapter summary, Section 1.8.

☞ *Caution*: The most common symbols used for the cylindrical distance are r and ρ. However both of these are problematic — r, because this is a *different* distance than r used above (from the origin to the point $\mathbf{r}$), and ρ, because ρ is universally used to represent density. While some duplication of symbols is inherently needed in physics, this one seems especially egregious as we will soon be integrating charge density in order to determine the amount of charge, which in cylindrical coordinates would lead to the ridiculous $\int \rho(\rho,\phi,z)\,\rho\,d\rho\,d\phi\,dz$. Therefore we will use the "s" notation, popularized by David Griffiths in *Introduction to Electrodynamics*.

The unit vectors and their scale factors

Write the position in Cartesian components, with the spherical coordinates as parameters:

$$\mathbf{r} = r\sin\theta\cos\phi\,\hat{\mathbf{x}} + r\sin\theta\sin\phi\,\hat{\mathbf{y}} + r\cos\theta\,\hat{\mathbf{z}}.$$

The direction in which the point moves when r alone increases is $\partial\mathbf{r}/\partial r$; when θ alone increases, $\partial\mathbf{r}/\partial\theta$; and when ϕ alone increases, $\partial\mathbf{r}/\partial\phi$:

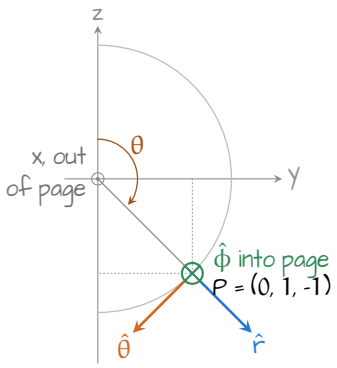
$$\begin{aligned}\frac{\partial\mathbf{r}}{\partial r} &= \sin\theta\cos\phi\,\hat{\mathbf{x}} + \sin\theta\sin\phi\,\hat{\mathbf{y}} + \cos\theta\,\hat{\mathbf{z}},\\ \frac{\partial\mathbf{r}}{\partial\theta} &= r\left(\cos\theta\cos\phi\,\hat{\mathbf{x}} + \cos\theta\sin\phi\,\hat{\mathbf{y}} - \sin\theta\,\hat{\mathbf{z}}\right),\\ \frac{\partial\mathbf{r}}{\partial\phi} &= r\sin\theta\left(-\sin\phi\,\hat{\mathbf{x}} + \cos\phi\,\hat{\mathbf{y}}\right).\end{aligned}$$

These three vectors have lengths 1, r, and $r\sin\theta$. Dividing each one by its length gives $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\phi}}$, as in (1.7). The lengths themselves are called *scale factors*. They say how far the point moves for a small change in each coordinate: a change dr moves it a distance dr, a change $d\theta$ a distance $r\,d\theta$, and a change $d\phi$ a distance $r\sin\theta\,d\phi$. They will appear again in the line and volume elements (Section 1.3) and in the derivatives (Section 1.4.6). The same steps in cylindrical coordinates give $\hat{\mathbf{s}}$ and $\hat{\boldsymbol{\phi}}$, with scale factors 1 for s, s for ϕ, and 1 for z.

Example 1.1

Find the spherical and cylindrical coordinates of the point $(x,y,z) = (0,1,-1)$, and express $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\phi}}$, and $\hat{\mathbf{s}}$ at that point in terms of $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$.

Solution:



To start out, we plot the point. (In almost every problem, the first thing to do is to visualize what is given and what is being asked by making a sketch, and nearly every example in this book begins with one.) Here, the point is in the yz plane, one unit to the right of the z axis and one unit below the xy plane. Then $\hat{\mathbf{r}}$ points away from the origin, $\hat{\boldsymbol{\theta}}$ points in the direction of increasing θ (tangent to the circle in the yz plane shown in the sketch), and $\hat{\boldsymbol{\phi}}$ points in the direction of increasing ϕ (the way the point moves if ϕ increases, i.e. around the z axis counterclockwise as seen from above).

Its distance from the origin is $r = \sqrt{0+1+1} = \sqrt{2}$, and its distance from the z axis is $s = 1$. The point is $45°$ below the xy plane, so θ is $90° + 45° = 135°$, or $3\pi/4$ radians if you prefer. The projection of the point onto the xy plane lies on the positive y axis, so $\phi = 90° = \pi/2$.

From the sketch, $\hat{\mathbf{r}}$ points in the $\mathbf{r}$ direction, down and to the right, and is $\mathbf{r}/r = \frac{\hat{\mathbf{y}}-\hat{\mathbf{z}}}{\sqrt{2}}$. $\hat{\boldsymbol{\theta}}$ points down and to the left, which is $\hat{\boldsymbol{\theta}} = -\frac{\hat{\mathbf{y}}+\hat{\mathbf{z}}}{\sqrt{2}}$; and $\hat{\boldsymbol{\phi}}$ points into the page, which is $\hat{\boldsymbol{\phi}} = -\hat{\mathbf{x}}$. Meanwhile, $\hat{\mathbf{s}}$ (not shown) points in the direction of increasing s, directly away from the z axis, which here is $\hat{\mathbf{y}}$.

In summary, we therefore have

$$\hat{\mathbf{r}} = \frac{\hat{\mathbf{y}}-\hat{\mathbf{z}}}{\sqrt{2}}, \qquad \hat{\boldsymbol{\theta}} = -\frac{\hat{\mathbf{y}}+\hat{\mathbf{z}}}{\sqrt{2}}, \qquad \hat{\boldsymbol{\phi}} = -\hat{\mathbf{x}}, \qquad \hat{\mathbf{s}} = \hat{\mathbf{y}}.$$

You can check each one against (1.7) if you'd like.

Reality check: Each answer should be a unit vector. Notice the factors of $1/\sqrt{2}$: an answer like $\hat{\boldsymbol{\theta}} = -\hat{\mathbf{y}} - \hat{\mathbf{z}}$ points in the right direction but is not a unit vector. Also, $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\phi}}$ should be perpendicular to each other, and they are: $\hat{\mathbf{r}}\cdot\hat{\boldsymbol{\theta}} = -(1-1)/2 = 0$, and $\hat{\boldsymbol{\phi}} = -\hat{\mathbf{x}}$ has no $\hat{\mathbf{y}}$ or $\hat{\mathbf{z}}$ component.



Alexis Claude Clairaut (1713–1765, French) Clairaut was admitted to the Paris Academy of Sciences in 1731, at 18, below the legal age for membership. In 1736–37 he took part in Maupertuis's expedition to Lapland to measure the shape of the Earth, and his *Théorie de la figure de la Terre* (1743) related the strength of gravity at the Earth's surface to its flattening at the poles. He and Euler both published (flawed) proofs in 1740 that mixed partial derivatives are equal, $\partial^2 f/\partial x\partial y = \partial^2 f/\partial y\partial x$, a result often called Clairaut's theorem, and he helped to predict the date of the 1759 return of Halley's comet. He is credited with being the first to think of polar coordinates in three dimensions; Euler was the first to develop them. (Wikipedia)

1.1.5 Constant and variable unit vectors

The unit vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ point in the same directions everywhere: they are constants. The unit vectors of spherical and cylindrical coordinates are not. Each one always has length one, but its direction depends on the location of the point $\mathbf{r}$. As (1.7) shows, $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\phi}}$, and $\hat{\mathbf{s}}$ all depend on θ, on ϕ, or on both (Figure 1.7). This has two consequences that come up again and again.

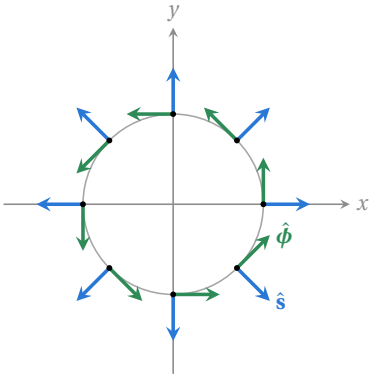


Figure 1.7 The unit vectors $\hat{\mathbf{s}}$ (blue) and $\hat{\boldsymbol{\phi}}$ (green) at eight points on a circle around the z axis, seen from above. Each has length one, but its direction depends on ϕ.

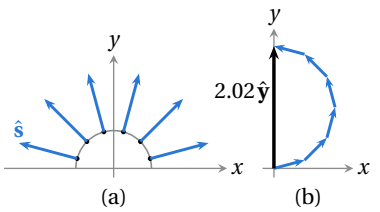


Figure 1.8 Adding up $\hat{\mathbf{s}}$ around half of a circle. (a) The unit vector $\hat{\mathbf{s}}$ at six points, $\phi = 15°, 45°, \ldots, 165°$. (b) The same six vectors, each multiplied by $\Delta\phi = 30° = \pi/6$ and placed head to tail. The x components cancel, and the sum, $2.02\hat{\mathbf{y}}$, approaches $2\hat{\mathbf{y}}$ as $\Delta\phi \to 0$.



Leonhard Euler (1707–1783, Swiss) Euler worked at the academies of St. Petersburg (1727–1741 and from 1766) and Berlin (1741–1766). He became almost blind in his right eye in 1738 and nearly blind in both in his last years, yet his output increased: with the help of scribes, in 1775 he produced on average one paper a week. Much of our notation is his, including $f(x)$ for a function, e, i, Σ for a sum, and the modern notation for the trigonometric functions. In 1769 he gave the rule for changing the variables in a double integral, an early form of the Jacobian; Lagrange extended it to triple integrals in 1773. Laplace is quoted as saying, "Read Euler, read Euler, he is the master of us all." (Wikipedia)

First, their derivatives are not zero. Differentiating (1.7) gives, for example,

$$\frac{\partial \hat{\mathbf{s}}}{\partial \phi} = \hat{\boldsymbol{\phi}}, \qquad \frac{\partial \hat{\boldsymbol{\phi}}}{\partial \phi} = -\hat{\mathbf{s}}, \qquad \frac{\partial \hat{\mathbf{r}}}{\partial \theta} = \hat{\boldsymbol{\theta}}, \qquad \frac{\partial \hat{\mathbf{r}}}{\partial \phi} = \sin\theta\, \hat{\boldsymbol{\phi}}. \tag{1.9}$$

Derivatives like these are part of the reason that the formulas for the divergence and curl in cylindrical and spherical coordinates (Section 1.4.6) contain extra terms.

Second, a variable unit vector cannot be taken outside an integral the way a constant can. For example, adding up the unit vectors $\hat{\mathbf{s}}$ around half of a circle (Figure 1.8),

$$\begin{aligned}\int_0^\pi \hat{\mathbf{s}}\, d\phi &= \int_0^\pi \left(\cos\phi\,\hat{\mathbf{x}} + \sin\phi\,\hat{\mathbf{y}}\right) d\phi \\ &= \hat{\mathbf{x}} \int_0^\pi \cos\phi\, d\phi + \hat{\mathbf{y}} \int_0^\pi \sin\phi\, d\phi \\ &= 2\hat{\mathbf{y}}.\end{aligned}$$

Taking $\hat{\mathbf{s}}$ outside the integral would give "$\pi\hat{\mathbf{s}}$", which does not even say which direction is meant, since $\hat{\mathbf{s}}$ points a different way at every ϕ. The right way, as here, is to write the variable unit vectors in terms of $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ first, and only then take those constants outside. Section 1.3.5 has another example, and Chapter 3 makes this a rule, Colton's First Commandment (Section 3.5.1).

One more point about θ and $\hat{\boldsymbol{\theta}}$. The angle θ is measured from a *polar axis*, and so far the polar axis has always been the z axis. It need not be. In Section 8.3.5, for example, the natural polar axis is the direction of an electric dipole moment, which can point any way. In general θ is the angle between $\hat{\mathbf{r}}$ and a unit vector $\hat{\mathbf{p}}$ along the polar axis, so that $\cos\theta = \hat{\mathbf{r}} \cdot \hat{\mathbf{p}}$, and $\hat{\boldsymbol{\theta}}$ is still the direction in which the point moves as θ increases: perpendicular to $\hat{\mathbf{r}}$, in the plane that contains $\hat{\mathbf{r}}$ and $\hat{\mathbf{p}}$, pointing away from the direction of $\hat{\mathbf{p}}$. (On the polar axis itself, $\hat{\boldsymbol{\theta}}$ is undefined, just as it is on the z axis.) The same point therefore has different values of θ and $\hat{\boldsymbol{\theta}}$ for different polar axes, as Figure 1.9 shows. Exercise P1.4 has more examples.

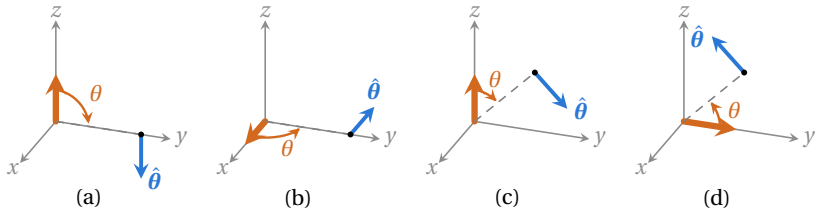


Figure 1.9 The direction of $\hat{\boldsymbol{\theta}}$ depends on the polar axis (thick orange arrow). (a) Polar axis $\hat{\mathbf{z}}$, point on the y axis: $\theta = 90°$ and $\hat{\boldsymbol{\theta}} = -\hat{\mathbf{z}}$. (b) Polar axis $\hat{\mathbf{x}}$, same point: $\theta = 90°$ and $\hat{\boldsymbol{\theta}} = -\hat{\mathbf{x}}$. (c) Polar axis $\hat{\mathbf{z}}$, point in the yz plane, on the $+y$ side, 45° from the z axis: $\theta = 45°$ and $\hat{\boldsymbol{\theta}} = (\hat{\mathbf{y}} - \hat{\mathbf{z}})/\sqrt{2}$. (d) Polar axis $\hat{\mathbf{y}}$, same point: $\theta = 45°$ and $\hat{\boldsymbol{\theta}} = (-\hat{\mathbf{y}} + \hat{\mathbf{z}})/\sqrt{2}$.

1.2 Scalar and vector fields

In physics, a *field* is simply a function which is defined at every point in space, in other words, a function of $\mathbf{r}$. Functions of $\mathbf{r}$ generally come in two flavors: ones which return a scalar value, such as the temperature, pressure, or density, and ones which return a vector value, such as heat flow, wind velocity, or gravitational field. We will refer to these as scalar and vector functions, or equivalently, scalar and vector fields. There are higher-order generalizations of scalar and vector fields called *tensor fields*, or often just "tensors", which we will encounter in a few places in this book, but we will leave a discussion of those for later chapters, starting with Chapter 8.

A vector field is really three scalar fields at once, one for each component:

$$\mathbf{A}(\mathbf{r}) = A_x(x,y,z)\,\hat{\mathbf{x}} + A_y(x,y,z)\,\hat{\mathbf{y}} + A_z(x,y,z)\,\hat{\mathbf{z}}. \tag{1.10}$$

It can equally well be written in terms of spherical or cylindrical unit vectors, as $\mathbf{A} = A_r\,\hat{\mathbf{r}} + A_\theta\,\hat{\boldsymbol{\theta}} + A_\phi\,\hat{\boldsymbol{\phi}}$, where $A_r = \mathbf{A}\cdot\hat{\mathbf{r}}$ and so on, as long as you remember that these unit vectors themselves change from point to point.

Two fields make good analogies for the fields of electromagnetism, and this chapter uses them often: the temperature in a room, which is a scalar field, and heat flow, which is a vector field. Sometimes fluid flow is also used as an example vector field. Heat flows from hot to cold, and in the same way the electric field points from high electric potential toward low; Chapter 5 shows that the electric field is related to the potential in the same way that the flow of heat is related to the temperature. The fields in this chapter depend only on position. Fields that also change with time appear from Chapter 15 on.

1.3 Integrating in three dimensions: Jacobians

Integrals in this book are taken along curves (line integrals), over surfaces (surface integrals), and through volumes (volume integrals). Each needs the size of a small piece of the region: a length, an area, or a volume. In Cartesian coordinates these are easy. In spherical and cylindrical coordinates they take a little more thought, and the thought is always the same: how far does the point move when each coordinate changes a little?



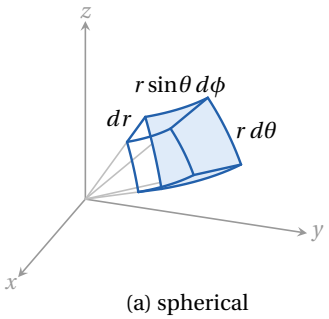
Carl Gustav Jacob Jacobi (1804–1851, German) Jacobi was a professor at Königsberg (from 1827) when his treatise on elliptic functions, *Fundamenta nova theoriae functionum ellipticarum*, appeared in 1829. He made fundamental contributions to classical mechanics, notably the Hamilton–Jacobi theory, and to the theory of determinants; in an 1841 paper he studied the determinant of the n^2 partial derivatives of n functions of n variables, now called the Jacobian. The Jacobi method of Section 7.3.2, for solving Laplace's equation numerically, is also named after him. He is said to have advised his students, "man muss immer umkehren" ("one must always invert!"). He died of smallpox at 46. (Wikipedia)

1.3.1 Line elements

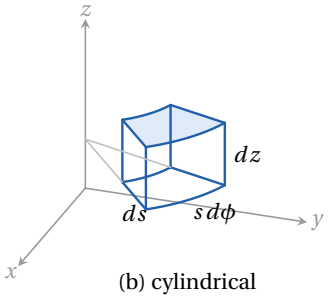
A small displacement of the point $\mathbf{r}$ is

$$d\boldsymbol{\ell} = dx\,\hat{\mathbf{x}} + dy\,\hat{\mathbf{y}} + dz\,\hat{\mathbf{z}}. \tag{1.11}$$

In spherical coordinates, changing r by dr moves the point a distance dr in the $\hat{\mathbf{r}}$ direction. Changing θ by $d\theta$ moves it along a circle of radius r centered on the origin, a distance $r\,d\theta$ in the $\hat{\boldsymbol{\theta}}$ direction. And changing ϕ by $d\phi$ moves it along a circle of radius $r\sin\theta$ around the z axis, a distance $r\sin\theta\,d\phi$ in the $\hat{\boldsymbol{\phi}}$



(a) spherical



(b) cylindrical

Figure 1.10 Volume elements. (a) In spherical coordinates, a small box with edges dr, $r\,d\theta$, and $r\sin\theta\,d\phi$. (b) In cylindrical coordinates, a small wedge with edges ds, $s\,d\phi$, and dz. Both are drawn much larger than they should be.

direction. These are the scale factors of Section 1.1.4. The same reasoning works in cylindrical coordinates, so

$$\begin{aligned} d\boldsymbol{\ell} &= dr\,\hat{\mathbf{r}} + r\,d\theta\,\hat{\boldsymbol{\theta}} + r\sin\theta\,d\phi\,\hat{\boldsymbol{\phi}}, && \text{(spherical)} \\ d\boldsymbol{\ell} &= ds\,\hat{\mathbf{s}} + s\,d\phi\,\hat{\boldsymbol{\phi}} + dz\,\hat{\mathbf{z}}. && \text{(cylindrical)} \end{aligned} \tag{1.12}$$

1.3.2 Volume elements

A small box whose edges are the three displacements has a volume equal to the product of the three edge lengths (Figure 1.10):

$$\begin{aligned} d\tau &= dx\,dy\,dz, && \text{(Cartesian)} \\ d\tau &= r^2\sin\theta\,dr\,d\theta\,d\phi, && \text{(spherical)} \\ d\tau &= s\,ds\,d\phi\,dz. && \text{(cylindrical)} \end{aligned} \tag{1.13}$$

In the spherical case the box is curved, but when dr, $d\theta$, and $d\phi$ are small enough it is a rectangular box with edges dr, $r\,d\theta$, and $r\sin\theta\,d\phi$. We write the volume element as $d\tau$, as Griffiths does, rather than dV, because V will be the electric potential.

It is tempting to write the spherical volume element as $dr\,d\theta\,d\phi$. A quick check of units shows that something is missing: θ and ϕ are dimensionless, so $dr\,d\theta\,d\phi$ has units of length, not volume. The missing factor, $r^2\sin\theta$ (or s in cylindrical coordinates), is called the **Jacobian** of the change of coordinates. Formally it is the determinant of the matrix of partial derivatives of x, y, and z with respect to the new coordinates,

$$\frac{\partial(x,y,z)}{\partial(r,\theta,\phi)} = \begin{vmatrix} \partial x/\partial r & \partial x/\partial\theta & \partial x/\partial\phi \\ \partial y/\partial r & \partial y/\partial\theta & \partial y/\partial\phi \\ \partial z/\partial r & \partial z/\partial\theta & \partial z/\partial\phi \end{vmatrix} = r^2\sin\theta, \tag{1.14}$$

which is the volume of the box spanned by $\partial\mathbf{r}/\partial r$, $\partial\mathbf{r}/\partial\theta$, and $\partial\mathbf{r}/\partial\phi$. When those three vectors are perpendicular to each other, as they are in spherical and cylindrical coordinates, the determinant is just the product of their lengths, the three scale factors. That is why multiplying the three edge lengths works (Exercise P1.5).

1.3.3 Area elements

For an integral over a surface, the area element is the product of the two edge lengths that lie in the surface (Figure 1.11). On the surface of a sphere of radius R, where r is fixed at R and there is no change in the r direction (i.e. no dr),

$$da = (R\,d\theta)(R\sin\theta\,d\phi) = R^2\sin\theta\,d\theta\,d\phi. \tag{1.15}$$

On the curved side of a cylinder of radius R, where s is fixed, $da = (R\,d\phi)(dz) = R\,d\phi\,dz$, and on a flat end of the cylinder, where z is fixed, $da = (ds)(s\,d\phi) =$

$s\,ds\,d\phi$, the same as in plane polar coordinates. A capital R is a reminder that the radius is fixed on the surface, unlike the variable r or s of a volume integral.

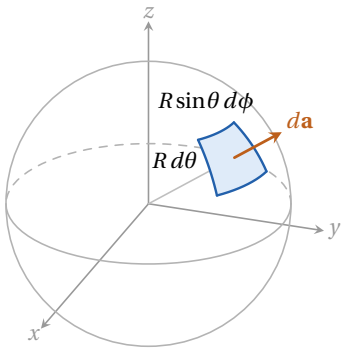
For flux integrals (Section 1.5) we need the **vector area element** $d\mathbf{a}$, whose magnitude is da and whose direction is perpendicular to the surface. On a closed surface, $d\mathbf{a}$ points outward:

$$\begin{aligned} d\mathbf{a} &= R^2 \sin\theta\, d\theta\, d\phi\, \hat{\mathbf{r}}, && \text{(sphere)} \\ d\mathbf{a} &= R\, d\phi\, dz\, \hat{\mathbf{s}}, && \text{(curved side of a cylinder)} \\ d\mathbf{a} &= \pm\, s\, ds\, d\phi\, \hat{\mathbf{z}}. && \text{(top and bottom of a cylinder)} \end{aligned} \tag{1.16}$$

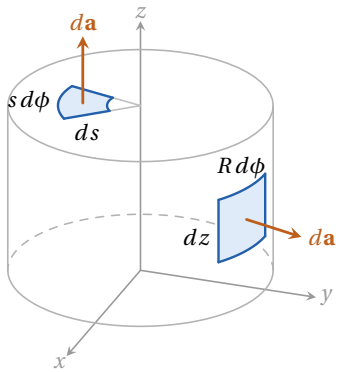
1.3.4 Choosing coordinates

Choose the coordinates for an integral to make the *limits* of integration simple, even if that makes the integrand more complicated. The shape of the region is what matters: a ball calls for spherical coordinates and a cylinder for cylindrical ones, whatever the form of the integrand. Nearly all of the integrals in this book are over regions of a few simple shapes — line segments and circles, rectangles and disks, boxes, balls, and cylinders.

One of the types of scalar functions we will need to integrate from time to time is the **charge density** ρ. It relates to charge in the same way that regular density relates to mass: it is $dq/d\tau$.



(a) sphere



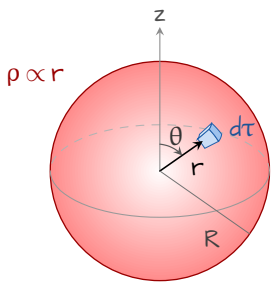
(b) cylinder

Figure 1.11 Area elements, drawn much larger than they should be. (a) On a sphere of radius R, a patch with edges $R\,d\theta$ and $R\sin\theta\,d\phi$. (b) On a cylinder of radius R, a patch on the curved side, with edges $R\,d\phi$ and dz, and a patch on the top, with edges ds and $s\,d\phi$. The arrows are the vector area elements $d\mathbf{a}$ of (1.16).

Example 1.2

A ball of radius R, centered on the origin, has the charge density $\rho = \rho_0\, r/R$, where ρ_0 is a constant. Use spherical coordinates to find the total charge of the ball.

Solution:



We start with a picture of the ball, shaded darker where the density is larger, with a volume element $d\tau$ at an arbitrary point (r,θ,ϕ) inside it. The density grows from zero at the center to ρ_0 at the surface. To cover the whole ball, r runs from 0 to R, θ from 0 to π, and ϕ from 0 to 2π.

The charge in the volume element is $dq = \rho\, d\tau$. Using $d\tau$ from (1.13), the total charge is

$$\begin{aligned} Q &= \int_0^{2\pi}\int_0^{\pi}\int_0^{R} \frac{\rho_0\, r}{R}\, r^2 \sin\theta\, dr\, d\theta\, d\phi \\ &= \frac{\rho_0}{R}\left(\int_0^R r^3\, dr\right)\left(\int_0^\pi \sin\theta\, d\theta\right)\left(\int_0^{2\pi} d\phi\right) \\ &= \frac{\rho_0}{R}\cdot\frac{R^4}{4}\cdot 2\cdot 2\pi = \pi\rho_0 R^3. \end{aligned}$$

The integral splits into three one-dimensional integrals. (Why can the triple integral be split this way? Answer in footnote.[3]) The θ and ϕ integrals together give 4π, a combination you will see often: whenever an integrand does not depend on θ or ϕ, the angular integrals give 4π.

Reality check: The factor r/R is a pure number, so ρ_0 has the units of a charge density, $\mathrm{C/m^3}$, and $\pi\rho_0 R^3$ is in $(\mathrm{C/m^3})(\mathrm{m^3}) = \mathrm{C}$, a charge, as it should be.

1.3.5 Integrals of vector fields

An integral of a vector field is a vector, and it is done one Cartesian component at a time. The constant unit vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ can be taken outside the integral; the variable ones cannot, as Section 1.1.5 explained.

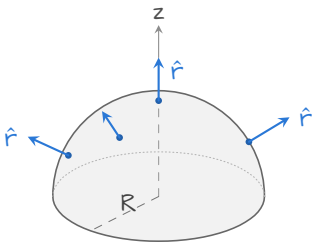


René Descartes (1596–1650, French) Descartes was educated at the Jesuit college of La Flèche and took a degree in law at Poitiers in 1616. He then served as a volunteer in the army of Maurice of Nassau and later in the Bavarian army, and from about 1628 he lived in the Netherlands for some twenty years. His *Discourse on the Method* (1637), the source of "I think, therefore I am," was published with three essays. One of them, *La Géométrie*, described curves by algebraic equations relating the distances of their points from fixed lines: the beginning of analytic geometry, which Pierre de Fermat developed independently. "Cartesian" comes from Cartesius, the Latin form of his name. Another essay, *La Dioptrique*, stated the law of refraction, which Willebrord Snell had found earlier but not published. In 1649 Descartes accepted an invitation from Queen Christina of Sweden, and he died in Stockholm the following February. (Wikipedia)

Example 1.3

Evaluate $\int \hat{\mathbf{r}}\, da$ over the upper half of a sphere of radius R centered on the origin (the hemisphere with $z \geq 0$).

Solution:



Before integrating, we sketch the hemisphere and draw $\hat{\mathbf{r}}$ at a few points on it. The tempting answer is $\hat{\mathbf{r}} \int da = 2\pi R^2\, \hat{\mathbf{r}}$, with $\hat{\mathbf{r}}$ taken outside the integral. But, as the sketch shows, $\hat{\mathbf{r}}$ points in a different direction at each point of the hemisphere, so "$2\pi R^2\, \hat{\mathbf{r}}$" does not even say which direction is meant.

Instead, write $\hat{\mathbf{r}}$ in Cartesian components with (1.7), use (1.15), and take the constants $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ outside:

$$\int \hat{\mathbf{r}}\, da = \int_0^{2\pi} \int_0^{\pi/2} \left(\sin\theta\cos\phi\,\hat{\mathbf{x}} + \sin\theta\sin\phi\,\hat{\mathbf{y}} + \cos\theta\,\hat{\mathbf{z}}\right) R^2 \sin\theta\, d\theta\, d\phi.$$

The $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ terms contain $\int_0^{2\pi} \cos\phi\, d\phi = 0$ and $\int_0^{2\pi} \sin\phi\, d\phi = 0$. (Why should the $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ terms be zero? Answer in footnote.[4]) What is left is

$$\begin{aligned} \int \hat{\mathbf{r}}\, da &= \hat{\mathbf{z}} R^2 \left(\int_0^{2\pi} d\phi\right)\left(\int_0^{\pi/2} \cos\theta \sin\theta\, d\theta\right) \\ &= \hat{\mathbf{z}} R^2\, (2\pi)\left(\frac{1}{2}\right) \\ &= \pi R^2\, \hat{\mathbf{z}}. \end{aligned}$$

[3] The limits are constants, and the integrand is a product of a function of r alone, a function of θ alone, and a function of ϕ alone.

[4] For every point of the hemisphere there is another one on the opposite side of the z axis whose $\hat{\mathbf{r}}$ has the opposite horizontal component.

Reality check: The answer has a nice interpretation. The quantity $\hat{\mathbf{z}}\cdot\hat{\mathbf{r}}\,da = \cos\theta\,da$ is the area of the patch da projected onto the xy plane, so the z component of the integral is the total area of the hemisphere projected onto the xy plane: a disk of area πR^2.

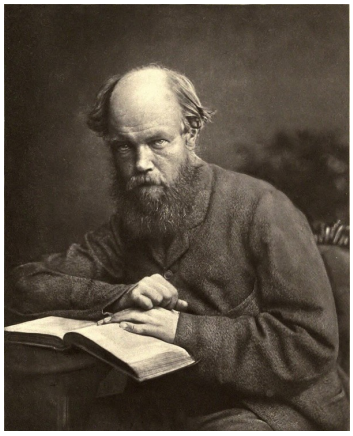
Exercise P1.7 gives more practice with variable unit vectors.

1.4 Derivatives in vector calculus: gradient, divergence, curl, Laplacian

A function of one variable has one derivative. A field depends on three variables, so it has three partial derivatives (and a vector field has three components, each with three partial derivatives), and there are several useful ways of combining them. All of them are built from the operator called "del",

$$\nabla = \hat{\mathbf{x}}\frac{\partial}{\partial x} + \hat{\mathbf{y}}\frac{\partial}{\partial y} + \hat{\mathbf{z}}\frac{\partial}{\partial z}, \tag{1.17}$$

which behaves like a vector whose components are derivatives. Applied to a scalar field, it gives the *gradient*; dotted into a vector field, it gives the *divergence*; crossed with a vector field, it gives the *curl*. Each of the three has a physical as well as a mathematical meaning, explored below. This section gives the formulas in Cartesian coordinates first; the cylindrical and spherical versions are in Section 1.4.6, and also are summarized at the very end of the textbook.



Peter Guthrie Tait (1831–1901, Scottish) Tait and James Clerk Maxwell were schoolmates at the Edinburgh Academy and lifelong friends; from 1860 Tait was professor of natural philosophy at Edinburgh. He was the leading champion of William Rowan Hamilton's quaternions and gave Hamilton's operator ∇ its full development. The name *nabla*, which is what the ∇ symbol is called in markup languages such as LaTeX, was suggested to him in an 1870 letter by William Robertson Smith due to its resemblance to a harp of that shape. Maxwell and Tait used the term, mostly in fun, in their letters, and Maxwell dedicated a humorous ode to the "Chief Musician upon Nabla", meaning Tait. With William Thomson (Lord Kelvin) he wrote the *Treatise on Natural Philosophy* (1867), known to its authors as "T and T′". (Wikipedia)

1.4.1 Gradient

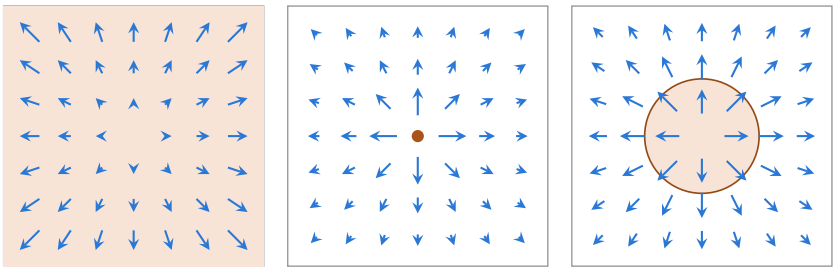
The gradient of a scalar field $f(x,y,z)$ is the vector field

$$\nabla f = \frac{\partial f}{\partial x}\hat{\mathbf{x}} + \frac{\partial f}{\partial y}\hat{\mathbf{y}} + \frac{\partial f}{\partial z}\hat{\mathbf{z}}. \tag{1.18}$$

At each point, ∇f points in the direction in which f increases fastest, and its magnitude is the rate of increase (the slope) in that direction. If f were the height of a hill, ∇f would point straight uphill, and its length would be how steep the hill is there. Individual components of ∇f can of course be negative: $-3\hat{\mathbf{x}}$ just means that f decreases as x increases. Heat flows down the temperature gradient, along $-\nabla T$, and a conservative force is $\mathbf{F} = -\nabla U$, where $U(x,y,z)$ is the potential energy as a function of position. In Chapter 5 the electric field will turn out to be $\mathbf{E} = -\nabla V$. In a small displacement $d\boldsymbol{\ell}$ the change in f is

$$df = \nabla f \cdot d\boldsymbol{\ell}, \tag{1.19}$$

the component of the slope along the direction you move, times the distance moved.



(a) $\mathbf{r}$ (b) $\hat{\mathbf{r}}/r^2$ (c) uniform ball

Figure 1.12 Three vector fields, drawn in the xy plane (x to the right, y up, z out of the page). Shading, or a dot, marks where the divergence is nonzero: (a) $\mathbf{r}$, with $\nabla\cdot\mathbf{r}=3$ everywhere; (b) $\hat{\mathbf{r}}/r^2$, whose divergence is zero everywhere except at the origin; (c) the field of a uniformly charged ball, whose divergence is nonzero only inside the ball. In (b) and (c) the arrow lengths are compressed so that the weak arrows stay visible.

1.4.2 Divergence

The divergence of a vector field $\mathbf{A}=A_x\hat{\mathbf{x}}+A_y\hat{\mathbf{y}}+A_z\hat{\mathbf{z}}$ is the scalar field

$$\nabla\cdot\mathbf{A}=\frac{\partial A_x}{\partial x}+\frac{\partial A_y}{\partial y}+\frac{\partial A_z}{\partial z}. \tag{1.20}$$

It measures how much the field flows *out of* a point. Imagine that $\mathbf{A}$ is the velocity of a fluid, and draw a tiny closed surface around the point. If more fluid leaves the surface than enters it, something inside must be producing fluid, like a faucet, and the divergence is positive there. If more enters than leaves, there is a drain, and the divergence is negative. If what enters equals what leaves, the divergence is zero. Stated precisely, the divergence of a vector function at a point in space is the net outward flow of the function through a small closed surface (the *flux*), divided by the volume it encloses, in the limit that the volume shrinks to a point:

$$\nabla\cdot\mathbf{A}=\lim_{\Delta\tau\to 0}\frac{1}{\Delta\tau}\oint\mathbf{A}\cdot d\mathbf{a}. \tag{1.21}$$

In terms of field lines, *the divergence is nonzero exactly where field lines begin or end*: positive where they begin, negative where they end.

That last statement deserves care, because it is not the same as "the divergence is nonzero where the arrows spread apart." Figure 1.12 shows three fields that all point away from the origin.

- Panel (a) is $\mathbf{A}=\mathbf{r}=x\hat{\mathbf{x}}+y\hat{\mathbf{y}}+z\hat{\mathbf{z}}$. From (1.20), $\nabla\cdot\mathbf{r}=1+1+1=3$, the same at every point. New field lines begin everywhere in space, as if there were a small faucet at every point, and that is why the arrows grow longer as you move out.

- Panel (b) is $\mathbf{A}=\hat{\mathbf{r}}/r^2=\mathbf{r}/r^3$, the shape of the electric field of a point charge. The arrows spread apart, which might suggest that this field has divergence

everywhere, but it does not. The x term of (1.20) is

$$\frac{\partial}{\partial x}\left(\frac{x}{r^3}\right) = \frac{1}{r^3} - \frac{3x^2}{r^5},$$

since $\partial r/\partial x = x/r$ (see footnote[5]), and adding the three terms together gives

$$\nabla\cdot\left(\frac{\hat{\mathbf{r}}}{r^2}\right) = \frac{3}{r^3} - \frac{3(x^2+y^2+z^2)}{r^5} = 0 \qquad (r \neq 0).$$

The lines spread out, but the field also weakens as $1/r^2$, exactly enough that as much flows out of any small volume as flows in (Figure 1.13).[6] No field lines begin or end anywhere except at the origin, where all of them begin. At the origin itself the field is infinite and (1.20) cannot be used; all of the divergence is concentrated there, in a way that Section 1.6.3 makes precise.

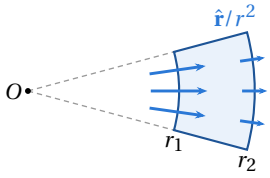
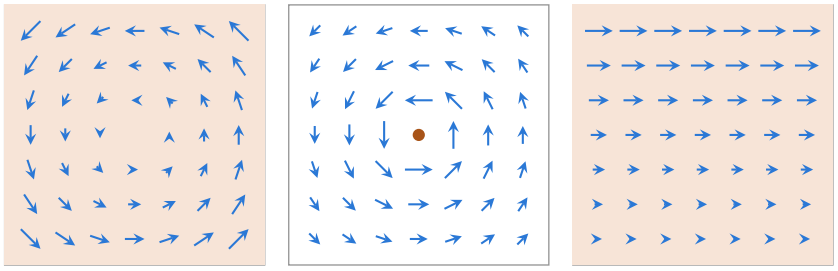


Figure 1.13 A thin cone between the spheres $r = r_1$ and $r = r_2$, in cross section, in the field $\hat{\mathbf{r}}/r^2$. The arrows are drawn with lengths proportional to the field, which is weaker at the outer face by the factor $(r_1/r_2)^2$, while the outer face is larger by $(r_2/r_1)^2$. As much flows out as flows in. Animation of the limit (1.21) in four fields.

- Panel (c) is the electric field of a uniformly charged ball (Gauss's law gives it easily; see Exercise P4.5): proportional to $\mathbf{r}$ inside the ball and to $\hat{\mathbf{r}}/r^2$ outside. Its divergence is a nonzero constant inside and zero outside. New lines begin throughout the ball, and none begin outside it.



(a) $s\hat{\boldsymbol{\phi}}$ (b) $\hat{\boldsymbol{\phi}}/s$ (c) $(c+y)\hat{\mathbf{x}}$

Figure 1.14 Three vector fields, drawn in the xy plane (x to the right, y up, z out of the page). Shading, or a dot, marks where the curl is nonzero: (a) $s\hat{\boldsymbol{\phi}}$, with curl $2\hat{\mathbf{z}}$ everywhere; (b) $\hat{\boldsymbol{\phi}}/s$, whose curl is zero everywhere except on the z axis; (c) $(c+y)\hat{\mathbf{x}}$, which has straight, parallel lines but curl $-\hat{\mathbf{z}}$ everywhere. In (b) the arrow lengths are compressed so that the weak arrows stay visible. Animation of paddle wheels in (a) and (b).

1.4.3 Curl

The curl of a vector field is itself a vector field:

$$\nabla\times\mathbf{A} = \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ A_x & A_y & A_z \end{vmatrix} = \left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}\right)\hat{\mathbf{x}} + \left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}\right)\hat{\mathbf{y}} + \left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right)\hat{\mathbf{z}}. \tag{1.22}$$

[5] With $r = (x^2+y^2+z^2)^{1/2}$, the chain rule gives $\partial r/\partial x = \frac{1}{2}(x^2+y^2+z^2)^{-1/2}(2x) = x/r$.

[6] The inverse square is special in this respect. For a radial field of the form $\hat{\mathbf{r}}/r^n$, the divergence away from the origin is $(2-n)/r^{n+1}$ (use the spherical formula (1.30)), which vanishes only for $n = 2$.

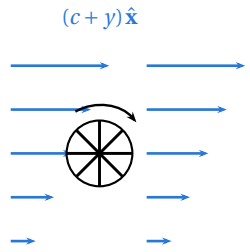


Figure 1.15 A small paddle wheel in the flow $(c+y)\,\hat{\mathbf{x}}$ of Figure 1.14(c). The flow is faster across the top of the wheel than across its bottom, so the wheel turns clockwise, and the curl points into the page, along $-\hat{\mathbf{z}}$. Animation of paddle wheels in the flows of Figure 1.14(a) and (b).

It measures how much the field circulates around a point — its "swirliness", as I like to call it. A good test is to imagine $\mathbf{A}$ as the velocity of a flowing fluid and to put a tiny paddle wheel into it. If the wheel turns, the curl is nonzero there (Figure 1.15). The curl points along the wheel's axis, by the right-hand rule (fingers in the direction of rotation, thumb along the curl), and its magnitude is proportional to how fast the wheel turns when its axis is oriented to make it turn fastest. Stated precisely, the component of the curl along a unit vector $\hat{\mathbf{n}}$ is the "circulation" $\oint \mathbf{A}\cdot d\boldsymbol{\ell}$ around a small loop perpendicular to $\hat{\mathbf{n}}$, divided by the loop's area, in the limit that the loop shrinks to the point. The loop is traversed counterclockwise as seen from the tip of $\hat{\mathbf{n}}$ (right-hand rule, thumb along $\hat{\mathbf{n}}$):

$$(\nabla\times\mathbf{A})\cdot\hat{\mathbf{n}} = \lim_{\Delta a\to 0}\frac{1}{\Delta a}\oint \mathbf{A}\cdot d\boldsymbol{\ell}. \tag{1.23}$$

The curl is nonzero where a small loop has a net circulation, which is not always where the field lines curve. Figure 1.14 shows three examples.

- Panel (a) is $\mathbf{A} = s\hat{\boldsymbol{\phi}} = -y\hat{\mathbf{x}} + x\hat{\mathbf{y}}$, the velocity field of a rigidly rotating turntable. From (1.22), $\nabla\times\mathbf{A} = 2\hat{\mathbf{z}}$ everywhere. A paddle wheel placed anywhere, not just at the center, turns counterclockwise.

- Panel (b) is $\mathbf{A} = \hat{\boldsymbol{\phi}}/s = (-y\hat{\mathbf{x}} + x\hat{\mathbf{y}})/(x^2+y^2)$, the shape of the magnetic field of a long straight wire (Example 10.5). The field lines are circles around the z axis, which might suggest that this field has curl everywhere, but it does not. The z component of (1.22) is

  $$\frac{\partial}{\partial x}\left(\frac{x}{s^2}\right) - \frac{\partial}{\partial y}\left(\frac{-y}{s^2}\right) = \left(\frac{1}{s^2} - \frac{2x^2}{s^4}\right) + \left(\frac{1}{s^2} - \frac{2y^2}{s^4}\right) = 0 \qquad (s\neq 0),$$

  and the other two components are zero because nothing depends on z. To see why a paddle wheel away from the axis does not turn, consider a small loop made of two arcs of a circle centered on the axis, at radii s and $s+\Delta s$, joined by two radial segments (Figure 1.16). The radial segments contribute nothing, because the field is perpendicular to them. The field is stronger on the inner arc, but the inner arc is shorter, in exactly the same proportion: each arc contributes $(1/s)(s\,\Delta\phi) = \Delta\phi$, with opposite signs, so the circulation is zero. A loop around the axis, on the other hand, has circulation 2π, however small it is. All of the curl is on the axis.

- Panel (c) is $\mathbf{A} = (c+y)\,\hat{\mathbf{x}}$, with c a positive constant large enough that $c+y>0$ throughout the panel. Every field line is straight, and they all point the same way, yet $\nabla\times\mathbf{A} = -\hat{\mathbf{z}}$ everywhere. A paddle wheel turns because the flow is faster across its top than across its bottom, so it turns clockwise (Figure 1.15).

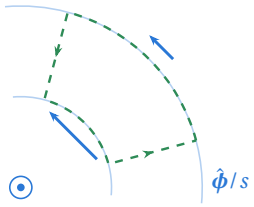


Figure 1.16 The small loop of the text in the field $\hat{\boldsymbol{\phi}}/s$, seen from $+z$ (the z axis is the dot), drawn with $\Delta s = s$. The field (blue) is twice as strong on the inner arc, which is half as long, so the two arcs contribute equal and opposite amounts, and the radial segments contribute nothing.

1.4.4 Gauss's law and Ampère's law, conceptually

Two of Maxwell's equations (Section 2.1) answer, for electrostatics and magnetostatics, exactly the question this section has been asking: where does the field have divergence, and where does it have curl?

Gauss's law says that the divergence of the electric field is proportional to the charge density,

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0}.$$

With words, it says that electric field lines begin and end only where there is charge: they begin on positive charge and end on negative charge. (The exceptions are lines that run off to infinity, and lines that run into a point where the field is zero; see Section 3.3.) The field of a point charge (Figure 1.12b) has divergence only at the charge; the field of a uniformly charged ball (Figure 1.12c) has divergence only inside the ball. Chapter 4 derives Gauss's law from Coulomb's law.

Ampère's law says that, for steady currents, the curl of the magnetic field is proportional to the current density,

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J}.$$

With words, magnetostatic field lines circle only around the places where current flows, and they have curl only at the location of the current. The field outside a thin wire (Figure 1.14b) has curl only on the wire, even though its field lines circle the wire everywhere. Inside a thick wire carrying a uniform current, the field is proportional to $s\hat{\boldsymbol{\phi}}$ (Figure 1.14a), and the curl is uniform throughout the wire. Section 11.1 derives Ampère's law from the Biot–Savart law.

By contrast, the curl of the electric field and the divergence of the magnetic field are both everywhere zero in statics. A static electric field never has curl (Section 5.1): its field lines can curve, but no small loop in it has any net circulation. The magnetic field never has divergence, so magnetic field lines never begin or end. Each law also has an integral form, which involves the flux of **E** through a closed surface or the circulation of **B** around a closed loop; the integral forms are connected to the differential forms shown above by the divergence and curl theorems of Section 1.5.



Pierre-Simon Laplace (1749–1827, French) Laplace's five-volume *Traité de mécanique céleste* (1799–1825) gave Newton's theory of the Solar System a systematic mathematical form. In the 1780s, studying the gravitational attraction of spheroids, he showed that the gravitational potential outside the attracting body satisfies the equation now named after him, $\nabla^2 V = 0$, and he introduced the functions now called spherical harmonics. His *Théorie analytique des probabilités* (1812) laid down many of the foundations of probability and statistics, he developed the nebular hypothesis for the formation of the Solar System, and he pioneered the transform that bears his name. In 1799 he served for six weeks as Napoleon's Minister of the Interior. (Wikipedia)

1.4.5 Second derivatives and the Laplacian

Derivatives can be applied twice. The most important combination is the divergence of a gradient, called the **Laplacian**:

$$\nabla^2 f = \nabla \cdot (\nabla f) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}. \tag{1.24}$$

It compares the value of f at a point with the average of f in a small region around it: $\nabla^2 f$ is positive where f is smaller than the average of its surroundings (the bottom of a bowl) and negative where it is larger (the top of a hill). Chapter 7 makes heavy use of this idea. The Laplacian of a vector field, in Cartesian coordinates only, is the Laplacian of each component, $\nabla^2 \mathbf{A} = (\nabla^2 A_x)\,\hat{\mathbf{x}} + (\nabla^2 A_y)\,\hat{\mathbf{y}} + (\nabla^2 A_z)\,\hat{\mathbf{z}}$.

Two other combinations of derivatives are always zero: the curl of a gradient field, $\nabla \times (\nabla f) = 0$, and the divergence of a curl, $\nabla \cdot (\nabla \times \mathbf{A}) = 0$.

1.4.6 Cylindrical and spherical coordinates

In cylindrical and spherical coordinates the formulas are more complicated, because the coordinates do not all measure distance. A small step $d\phi$ in cylindrical coordinates moves you a distance $s\,d\phi$, not $d\phi$, so the rate of change of f in the $\hat{\boldsymbol{\phi}}$ direction is $(1/s)\,\partial f/\partial\phi$. Similar factors appear throughout.

Cylindrical (s,ϕ,z):

$$\nabla f = \frac{\partial f}{\partial s}\hat{\mathbf{s}} + \frac{1}{s}\frac{\partial f}{\partial \phi}\hat{\boldsymbol{\phi}} + \frac{\partial f}{\partial z}\hat{\mathbf{z}}, \tag{1.25}$$

$$\nabla\cdot\mathbf{A} = \frac{1}{s}\frac{\partial(sA_s)}{\partial s} + \frac{1}{s}\frac{\partial A_\phi}{\partial\phi} + \frac{\partial A_z}{\partial z}, \tag{1.26}$$

$$\nabla\times\mathbf{A} = \left(\frac{1}{s}\frac{\partial A_z}{\partial\phi} - \frac{\partial A_\phi}{\partial z}\right)\hat{\mathbf{s}} + \left(\frac{\partial A_s}{\partial z} - \frac{\partial A_z}{\partial s}\right)\hat{\boldsymbol{\phi}} + \frac{1}{s}\left(\frac{\partial(sA_\phi)}{\partial s} - \frac{\partial A_s}{\partial\phi}\right)\hat{\mathbf{z}}, \tag{1.27}$$

$$\nabla^2 f = \frac{1}{s}\frac{\partial}{\partial s}\left(s\frac{\partial f}{\partial s}\right) + \frac{1}{s^2}\frac{\partial^2 f}{\partial\phi^2} + \frac{\partial^2 f}{\partial z^2}. \tag{1.28}$$

Spherical (r,θ,ϕ):

$$\nabla f = \frac{\partial f}{\partial r}\hat{\mathbf{r}} + \frac{1}{r}\frac{\partial f}{\partial\theta}\hat{\boldsymbol{\theta}} + \frac{1}{r\sin\theta}\frac{\partial f}{\partial\phi}\hat{\boldsymbol{\phi}}, \tag{1.29}$$

$$\nabla\cdot\mathbf{A} = \frac{1}{r^2}\frac{\partial(r^2A_r)}{\partial r} + \frac{1}{r\sin\theta}\frac{\partial(\sin\theta\,A_\theta)}{\partial\theta} + \frac{1}{r\sin\theta}\frac{\partial A_\phi}{\partial\phi}, \tag{1.30}$$

$$\begin{aligned}\nabla\times\mathbf{A} = {}& \frac{1}{r\sin\theta}\left(\frac{\partial(\sin\theta\,A_\phi)}{\partial\theta} - \frac{\partial A_\theta}{\partial\phi}\right)\hat{\mathbf{r}} + \frac{1}{r}\left(\frac{1}{\sin\theta}\frac{\partial A_r}{\partial\phi} - \frac{\partial(rA_\phi)}{\partial r}\right)\hat{\boldsymbol{\theta}} \\ & + \frac{1}{r}\left(\frac{\partial(rA_\theta)}{\partial r} - \frac{\partial A_r}{\partial\theta}\right)\hat{\boldsymbol{\phi}},\end{aligned} \tag{1.31}$$

$$\nabla^2 f = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial f}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial f}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 f}{\partial\phi^2}. \tag{1.32}$$

These formulas, like (1.20) and (1.22), assume that the field is finite and differentiable at the point in question. Where it is not — at the location of a point charge, for example — they can give the wrong answer, as Section 1.6.3 shows. The same formulas are printed on page 940, near the end of the book, facing the table of physical constants.

Tips for Mathematica and Python. Symbolic programs know these formulas.

- Mathematica's `Grad`, `Div`, `Curl`, and `Laplacian` accept a coordinate system as a third argument. For example, `Div[{r, 0, 0}, {r, th, ph}, "Spherical"]` returns 3, the divergence of $r\,\hat{\mathbf{r}}$ (with `th` and `ph` standing for θ and ϕ). Its `"Spherical"` system uses the physics convention for the two angles.

- In Python, the `sympy.vector` module does the same. Create the coordinate system with `CoordSys3D('S', transformation='spherical')`

(or `'cylindrical'`), and pass fields built from its coordinates and unit vectors to the module's `gradient`, `divergence`, and `curl` functions. For example, `divergence(S.r*S.i)` returns 3. Beware that sympy calls the cylindrical coordinates `r`, `theta`, and `z`, where this book uses s, ϕ, and z.

- Neither program knows about the delta functions of Section 1.6.3: both report that the divergence of $\hat{\mathbf{r}}/r^2$ is zero, just as the formulas above do.

1.5 The fundamental theorems

Section 1.3 was about integrating fields and Section 1.4 about differentiating them. This section asks what you get when you integrate a derivative. In one dimension the answer is the fundamental theorem of calculus, and each of the three derivatives of vector calculus has a theorem of its own. All four say the same thing: *the integral of a derivative over a region is determined by the values of the function on the boundary of the region.*

1.5.1 Fundamental theorem of calculus

The fundamental theorem of calculus is

$$\int_a^b \frac{df}{dx}\,dx = f(b) - f(a). \tag{1.33}$$

With words, it says, "the integral of a derivative is equal to the difference of the function evaluated at the endpoints." There is a nice geometrical interpretation if you think of the function f as being like a height. An integral is just a summation. What you are summing up is the slope at each point (df/dx), multiplied by the horizontal distance at each point (dx), which gives you the rise at that point (Figure 1.17). You add up all of the rises and you get the *total* rise, which is just the height difference between b and a.

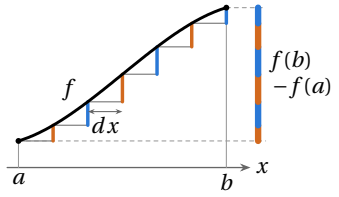


Figure 1.17 The fundamental theorem of calculus, with f drawn as a height. Over each small step dx the curve rises by $(df/dx)\,dx$ (colored). Stacked end to end (right), the rises add up to the total rise, $f(b) - f(a)$.

1.5.2 Gradient theorem

The **gradient theorem** is

$$\int_{\mathbf{a}}^{\mathbf{b}} \nabla f \cdot d\boldsymbol{\ell} = f(\mathbf{b}) - f(\mathbf{a}), \tag{1.34}$$

where the integral is taken along any path from the point **a** to the point **b**. It has the same geometrical interpretation as the fundamental theorem if f is considered to be a "height". The gradient gives you the slope in the "uphill" direction, so by (1.19) the gradient dotted into $d\boldsymbol{\ell}$ gives you the rise at each point along the path, and adding up the rises gives the total rise from **a** to **b**.

Two consequences follow at once. First, the integral is the same for every path from **a** to **b**, since the right side depends only on the endpoints (Figure 1.18).

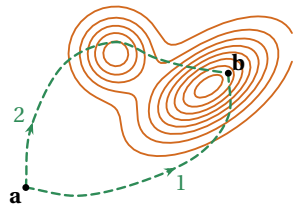


Figure 1.18 Two paths from **a** to **b** across a contour map of f (orange), a landscape of two hills. Along either one, $\int \nabla f \cdot d\boldsymbol{\ell}$ adds up the rises in f (a fall counts as a negative rise), so both give $f(\mathbf{b}) - f(\mathbf{a})$. Going out along one path and back along the other makes a closed loop, around which the integral is zero.

Second, the integral around any closed path is zero,

$$\oint \nabla f \cdot d\boldsymbol{\ell} = 0, \tag{1.35}$$

since a closed path ends where it starts. A force with this property, $\mathbf{F} = -\nabla U$, is called *conservative*: the work it does, $\int \mathbf{F} \cdot d\boldsymbol{\ell}$, depends only on the endpoints, and the work it does around a closed path is zero. (Friction is the standard example of a force that is not conservative.) In Chapter 5 the same two facts will make the electric potential possible.

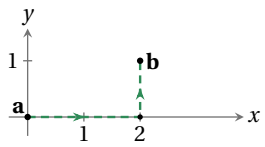
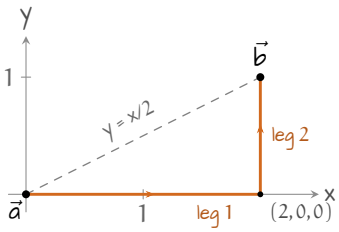


Figure 1.19 Example 1.4: the points **a** and **b** and the path between them, along the x axis and then parallel to the y axis.

Example 1.4

Check the gradient theorem for $T = xy^2$ between $\mathbf{a} = (0,0,0)$ and $\mathbf{b} = (2,1,0)$, along the path that goes first along the x axis to $(2,0,0)$ and then parallel to the y axis to $(2,1,0)$.

Solution:



First we draw the path in the xy plane: leg 1 along the x axis from **a** to $(2,0,0)$, then leg 2 parallel to the y axis up to **b**. (The dashed line is a second path, used at the end.) The gradient is $\nabla T = y^2\hat{\mathbf{x}} + 2xy\hat{\mathbf{y}}$. On the first leg $y = 0$ and $d\boldsymbol{\ell} = dx\hat{\mathbf{x}}$, so $\nabla T \cdot d\boldsymbol{\ell} = y^2\,dx = 0$: the first leg contributes nothing.

On the second leg $x = 2$ and $d\boldsymbol{\ell} = dy\hat{\mathbf{y}}$, so $\nabla T \cdot d\boldsymbol{\ell} = 2xy\,dy = 4y\,dy$, and

$$\int_0^1 4y\,dy = 2.$$

The left side of (1.34) is therefore $0 + 2 = 2$. The right side is $T(2,1,0) - T(0,0,0) = 2 \cdot 1^2 - 0 = 2$, so the two sides agree.

Reality check: A different path gives the same answer, as it must. On the straight line from **a** to **b**, $y = x/2$, so with x as the variable of integration $d\boldsymbol{\ell} = dx\hat{\mathbf{x}} + dy\hat{\mathbf{y}} = dx\hat{\mathbf{x}} + \frac{1}{2}dx\hat{\mathbf{y}}$ and

$$\begin{aligned}\int \nabla T \cdot d\boldsymbol{\ell} &= \int_0^2 \left[\frac{x^2}{4} + 2x\left(\frac{x}{2}\right)\frac{1}{2}\right] dx \\ &= \int_0^2 \frac{3x^2}{4}\,dx \\ &= 2.\end{aligned}$$

1.5.3 Divergence theorem

The **divergence theorem** is

$$\int_{\mathcal{V}} (\nabla \cdot \mathbf{A})\, d\tau = \oint_{\mathcal{S}} \mathbf{A} \cdot d\mathbf{a}. \tag{1.36}$$

Here $\mathcal{V}$ (a calligraphic capital letter V) is a volume of space and $\mathcal{S}$ (a calligraphic capital letter S) is the closed surface that bounds it. The circle on the integral sign is a reminder that the surface is closed, and the area element $d\mathbf{a}$ points outward, perpendicular to $\mathcal{S}$. Just like the fundamental theorem and the gradient theorem, the divergence theorem says that if you integrate the derivative of a function (in this case, the divergence of a vector function) over a region of space (a volume), the result involves evaluating the function on the boundary of the region (the surface).

Geometrically, this theorem involves *flux*. If you have sources of field lines, then there are two ways of finding how much flux they produce: you can add up the *sources* of flux throughout the volume (the left-hand side), or you can add up the flux itself passing out through the surface (the right-hand side). Think of $\mathbf{A}$ as the flow of water, with faucets running here and there inside the volume. You can find how much water is being produced either by adding up the output of every faucet or by measuring how much water flows out across the boundary. Flow that merely passes through the volume, entering on one side and leaving on the other, adds nothing to either side: there is no divergence where it passes, and its inward and outward flux cancel.

The theorem also follows from the meaning of the divergence, (1.21). Divide the volume into many small cubes. The integral of $\nabla \cdot \mathbf{A}$ over one small cube is the flux out of that cube. Wherever two cubes share a face, the flux out of one through that face is the flux into the other, so in the sum over all of the cubes those contributions cancel, and only the flux through the outer boundary is left (Figure 1.20).

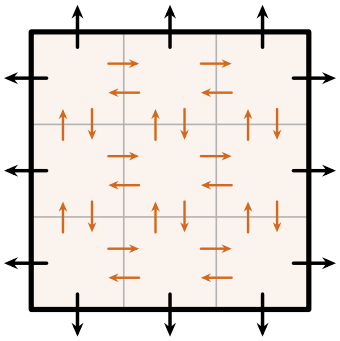


Figure 1.20 Adding up the fluxes out of many small cubes, seen in cross section. Through every inside face, the outward directions of the two neighboring cubes (orange arrows) are opposite, so their fluxes through it cancel, leaving only the flux out through the outer surface (thick).

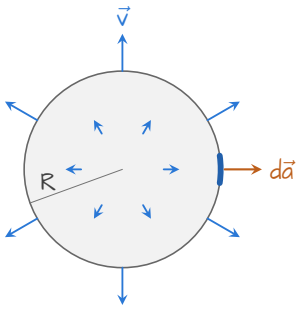
Example 1.5

Check the divergence theorem for $\mathbf{v} = r\,\hat{\mathbf{r}}$ over a ball of radius R centered on the origin.



George Gabriel Stokes (1819–1903, Irish) Born in County Sligo, Ireland, Stokes graduated from Cambridge as senior wrangler in 1841 and became Lucasian Professor of Mathematics in 1849, holding the chair for 54 years, longer than anyone else. His work on viscous fluids produced the Navier–Stokes equations and Stokes' law for the drag on a slowly moving sphere, and in 1852 he named and explained fluorescence. He was president of the Royal Society from 1885 to 1890 and a Member of Parliament for Cambridge University from 1887 to 1892. The curl theorem carries his name because he set it as a question on a Cambridge examination in 1854, but the first known statement of the curl theorem is in a letter from William Thomson (Lord Kelvin) to Stokes, which is why it is sometimes called the Kelvin–Stokes theorem. (Wikipedia)

Solution:



As usual, a sketch of what is given comes first: the ball, and the field, which points radially outward and grows in proportion to r. On the surface, the vector area $d\mathbf{a}$ of a patch points outward too, along $\hat{\mathbf{r}}$. For the left side, use the spherical divergence (1.30). Only its first term survives, since $\mathbf{v}$ has only an $\hat{\mathbf{r}}$ component:

$$\begin{aligned}\nabla\cdot\mathbf{v} &= \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\cdot r\right) \\ &\quad + \cancel{\frac{1}{r\sin\theta}\frac{\partial(\sin\theta\, v_\theta)}{\partial\theta}} + \cancel{\frac{1}{r\sin\theta}\frac{\partial v_\phi}{\partial\phi}} \\ &= \frac{3r^2}{r^2} = 3.\end{aligned}$$

Reality check: This is the field $\mathbf{r}$ of Figure 1.12(a), whose divergence we found in Cartesian coordinates to be $1+1+1=3$ as well. (Why, then, is the divergence not simply $\partial v_r/\partial r = 1$? Answer in footnote.[7])

The left side is therefore 3 times the volume of the ball,

$$\int 3\,d\tau = 3\cdot\frac{4}{3}\pi R^3 = 4\pi R^3.$$

For the right side, on the surface $r=R$ the field is $\mathbf{v}=R\hat{\mathbf{r}}$, and by (1.16) $d\mathbf{a} = R^2\sin\theta\,d\theta\,d\phi\,\hat{\mathbf{r}}$. Since $\hat{\mathbf{r}}\cdot\hat{\mathbf{r}}=1$,

$$\oint \mathbf{v}\cdot d\mathbf{a} = \int_0^{2\pi}\int_0^{\pi} R\cdot R^2\sin\theta\,d\theta\,d\phi = 4\pi R^3.$$

The two sides agree. The divergence of this field is the same at every point, so new field lines begin everywhere inside the ball, at the same rate. This is what the electric field looks like inside a ball of uniform charge density (Chapter 4).

1.5.4 Curl theorem (Stokes' theorem)

The **curl theorem** is

$$\int_{\mathcal{S}} (\nabla\times\mathbf{A})\cdot d\mathbf{a} = \oint_{\mathcal{P}} \mathbf{A}\cdot d\boldsymbol{\ell}. \tag{1.37}$$

Here $\mathcal{S}$ is a surface that is *not* closed — think of a soap film on a wire loop — and $\mathcal{P}$ (a calligraphic capital letter P) is the closed curve, the perimeter, that bounds it. It is also called **Stokes' theorem**, but I typically use the term "curl theorem", because the name Stokes' theorem is also used for a more general result

[7] Because $\hat{\mathbf{r}}$ is not a constant unit vector (Section 1.1.5): its direction changes from point to point, which adds a term to $\partial v_r/\partial r$. The factors $1/r^2$ and r^2 in (1.30) produce that term, which here is 2.

in mathematics. Just like the other three theorems, the curl theorem says that if you integrate the derivative of the function (in this case, the curl of a vector function) over a region of space (a surface), then the result involves evaluating the function on the boundary of the region of space (the perimeter).

Geometrically, this theorem involves *circulation*, or as I called it in Section 1.4.3, "swirliness". There are two ways of finding the total amount of swirl: you can add up the curl over the surface (the left-hand side of the theorem), or you can go around the edge and add up how much the field runs along the boundary (the right-hand side of the theorem). The curl theorem says the two must agree. Divide the surface into many small patches (Figure 1.21). By (1.23), the curl integrated over one patch is the circulation around its edge. Where two patches share an edge, they run along it in opposite directions and their contributions cancel, so in the sum only the edges on the outer boundary survive.

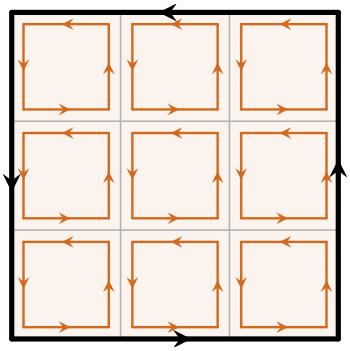


Figure 1.21 Adding up the circulations around many small squares. Along every inside edge, two neighboring squares run in opposite directions and cancel, leaving only the circulation around the outer boundary (thick).

Two details matter in practice. First, the direction of $d\mathbf{a}$ and the direction around $\mathscr{P}$ are linked by a right-hand rule: if the fingers of your right hand curl in the direction you go around $\mathscr{P}$, your thumb points along $d\mathbf{a}$ (Figure 1.22). Second, the surface is not unique. Any surface bounded by $\mathscr{P}$ gives the same left side; a flat disk spanning a circular loop and a hemispherical bowl on the same loop give the same answer. In this book we will nearly always use flat surfaces.

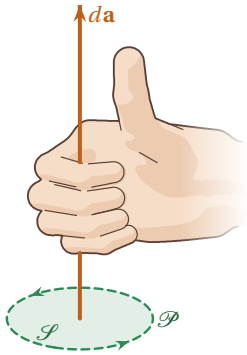
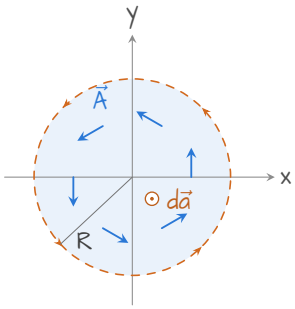


Figure 1.22 The right-hand rule of the curl theorem. When the fingers of your right hand curl in the direction you go around the boundary $\mathscr{P}$ of the surface $\mathscr{S}$, your thumb points along $d\mathbf{a}$.

Example 1.6

Check the curl theorem for $\mathbf{A} = -y\hat{\mathbf{x}} + x\hat{\mathbf{y}}$, the field of Figure 1.14(a), on a disk of radius R in the xy plane, centered on the origin, with $d\mathbf{a}$ in the $+z$ direction.

Solution:



We begin by drawing the disk as seen from above, so that $d\mathbf{a}$, along $+\hat{\mathbf{z}}$, points out of the page. The field circles the origin counterclockwise and grows with distance from it, and the dashed circle is the disk's boundary. In Section 1.4.3 we found $\nabla \times \mathbf{A} = 2\hat{\mathbf{z}}$. With $d\mathbf{a} = s\,ds\,d\phi\,\hat{\mathbf{z}}$, the left side is

$$\int (2\hat{\mathbf{z}}) \cdot \hat{\mathbf{z}}\, da = 2 \times (\text{area of the disk}) = 2\pi R^2.$$

For the right side, we go around the circle counterclockwise as seen from above. (Why counterclockwise, and not clockwise? Answer in footnote.[8]) On the circle $s = R$, with $x = R\cos\phi$ and $y = R\sin\phi$, the field is $\mathbf{A} = R(-\sin\phi\,\hat{\mathbf{x}} + \cos\phi\,\hat{\mathbf{y}}) = R\hat{\boldsymbol{\phi}}$ by (1.7), and by (1.12) $d\boldsymbol{\ell} = R\,d\phi\,\hat{\boldsymbol{\phi}}$. So

$$\oint \mathbf{A} \cdot d\boldsymbol{\ell} = \int_0^{2\pi} R \cdot R\,d\phi = 2\pi R^2,$$

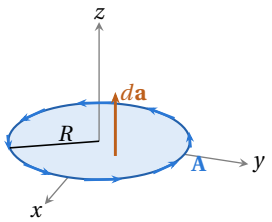


Figure 1.23 Example 1.6: a disk of radius R in the xy plane, with $d\mathbf{a}$ along $+\hat{\mathbf{z}}$, in the field $\mathbf{A} = -y\hat{\mathbf{x}} + x\hat{\mathbf{y}}$, drawn on the rim.

[8] By the right-hand rule: with the thumb of your right hand along $d\mathbf{a}$, that is, along $+\hat{\mathbf{z}}$, your fingers curl counterclockwise as seen from above.

and the two sides agree.

Reality check: Both sides are positive, as they should be. The field circles the origin counterclockwise (see the sketch), the same way the path goes, so $\mathbf{A}\cdot d\boldsymbol{\ell}$ is positive all the way around.

1.5.5 The four theorems together

Table 1.1 lists the four theorems side by side. In each one, the integral of a derivative over a region equals an integral of the function itself over the boundary of that region.

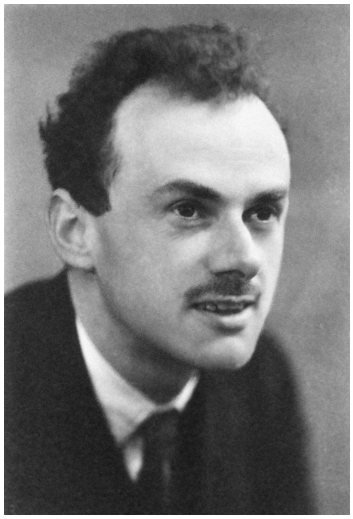
Theorem	Derivative	Region	Boundary
Fundamental	df/dx	interval from a to b	endpoints a and b
Gradient	∇f	path from **a** to **b**	endpoints **a** and **b**
Divergence	$\nabla\cdot\mathbf{A}$	volume $\mathcal{V}$	closed surface $\mathcal{S}$
Curl	$\nabla\times\mathbf{A}$	surface $\mathcal{S}$	closed curve $\mathcal{P}$

Table 1.1 The four fundamental theorems side by side.

We will use the divergence theorem and the curl theorem often in this book, starting with Gauss's law in Chapter 4. The gradient theorem appears mainly in Chapter 5, where it connects the electric field to the potential.

1.6 The Dirac delta function

How would you write the density of a point mass, or the charge density of a point charge? It must be zero everywhere except at one point, and yet its integral over any volume containing that point must be the total mass or charge. No ordinary function behaves like that. The object that does is the Dirac delta function, or as I shall more simply call it, the delta function.



Paul Dirac (1902–1984, British) Paul Adrien Maurice Dirac took a degree in electrical engineering at Bristol in 1921 before turning to mathematics. His 1928 equation for the electron, which combined quantum mechanics with special relativity, predicted the positron, the electron's antiparticle, and he shared the 1933 Nobel Prize in Physics with Erwin Schrödinger. He introduced the delta function in a 1927 paper and made it widely known through his book *The Principles of Quantum Mechanics* (1930), although versions of it had been used by Fourier, Cauchy, Poisson, Kirchhoff, and Heaviside; Laurent Schwartz put it on a rigorous footing in the 1940s. In 1931 Dirac showed that if even one magnetic monopole exists, electric charge must be quantized. (Wikipedia)

1.6.1 The delta function in one dimension

The one-dimensional delta function $\delta(x)$ is defined by two properties:

$$\delta(x)=0 \quad \text{for } x\neq 0, \qquad \int_{-a}^{b}\delta(x)\,dx=1 \quad (a,b>0). \tag{1.38}$$

It is an infinitely tall, infinitely narrow spike at $x=0$ with an area of exactly one. The limits of the integral need not be infinite; they only need to include the spike. Strictly speaking, $\delta(x)$ is not a function: it has no well-defined value at $x=0$, only a well-defined integral. (Mathematicians call it a *distribution*.) Physicists nevertheless treat it as a function, and that is safe as long as it ends up inside an integral.

There are at least two good ways to picture it. The first is as a limit. The normalized Gaussian

$$g_\sigma(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-x^2/2\sigma^2} \tag{1.39}$$

has an area of one for every value of σ. As σ shrinks it becomes narrower and taller, keeping its area fixed (Figure 1.24), and in the limit $\sigma \to 0$ it becomes $\delta(x)$. A rectangle of width ε and height $1/\varepsilon$, with $\varepsilon \to 0$, works equally well.

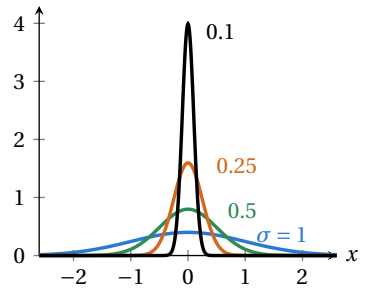


Figure 1.24 Normalized Gaussians, (1.39), for four values of σ. Each has area one; as $\sigma \to 0$ they approach $\delta(x)$.

A second way to picture the delta function is as the derivative of the step function, sometimes called the Heaviside step function,

$$\Theta(x) = \begin{cases} 0, & x < 0, \\ 1, & x > 0. \end{cases}$$

The derivative of the step function must be zero everywhere except at the step, where the slope is infinite — a spike at the origin! And the step of 1 is the right size for the spike to be exactly the delta function, since the integral of the derivative across the step gives $\int_{-a}^{b} (d\Theta/dx)\, dx = \Theta(b) - \Theta(-a) = 1$. Thus $\delta(x) = d\Theta/dx$.

A spike at x_0 instead of at the origin is $\delta(x - x_0)$, in the same way that $(x-5)^2$ is the parabola x^2 moved over to $x = 5$. This leads to the extremely important **sifting property** of the delta function:

$$\int f(x)\, \delta(x - x_0)\, dx = f(x_0), \tag{1.40}$$

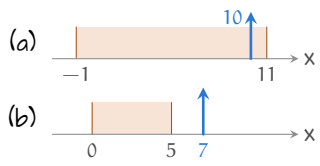
as long as the range of integration includes x_0 (otherwise the integral is zero). The reason is that the product $f(x)\,\delta(x - x_0)$ is zero everywhere except at x_0, so only the value $f(x_0)$ matters, and it multiplies a spike of area one.

One more fact follows from (1.38). Since $\int \delta(x)\, dx$ is the pure number 1 and dx has the units of x, $\delta(x)$ *has the units of* $1/x$: if x is a length, $\delta(x)$ is measured in m^{-1}.

Example 1.7

Evaluate (a) $\int_{-1}^{11} \sin x\, \delta(x - 10)\, dx$ and (b) $\int_0^5 x^2\, \delta(x - 7)\, dx$.

Solution:



Even for integrals like these, a sketch helps: for each one we draw the spike of the delta function and the range of integration. (a) The range includes $x = 10$, so by (1.40) the integral is $\sin 10 \approx -0.544$ (with 10 in radians). (b) The spike at $x = 7$ is outside the range, so the integral is zero.

Reality check: The answer to (a) should be negative: 10 radians is about 3.18π, a little more than 3π, and just past 3π the sine is negative. A calculator set to degrees would give a positive number instead.

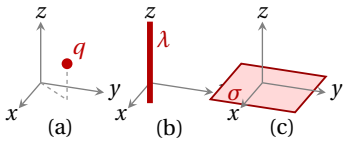
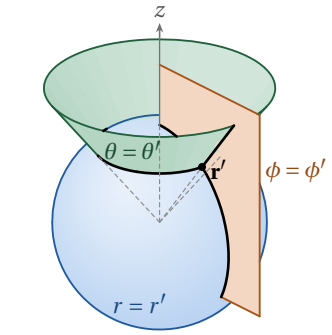
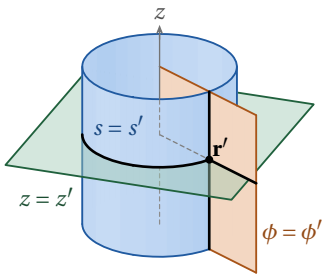


Figure 1.25 Charge densities made of delta functions: (a) a point charge q at $\mathbf{r}'$, $\rho = q\,\delta^3(\mathbf{r}-\mathbf{r}')$; (b) a line of charge λ along the z axis, $\rho = \lambda\,\delta(x)\,\delta(y)$; (c) a sheet of charge σ in the plane $z = 0$, $\rho = \sigma\,\delta(z)$.



(a) spherical



(b) cylindrical

Figure 1.26 The surfaces on which the one-dimensional delta functions of (1.43) are not zero, and the point $\mathbf{r}'$ where they meet. (a) The sphere $r = r'$, the cone $\theta = \theta'$, and the half-plane $\phi = \phi'$. (b) The cylinder $s = s'$, the half-plane $\phi = \phi'$, and the plane $z = z'$. The black curves are where two of the surfaces meet.

1.6.2 The delta function in three dimensions

The three-dimensional delta function is a product of three one-dimensional ones:

$$\delta^3(\mathbf{r}) = \delta(x)\,\delta(y)\,\delta(z). \tag{1.41}$$

The superscript 3 is not a power; it is a reminder that this is three delta functions multiplied together. $\delta^3(\mathbf{r})$ is zero everywhere except at the origin, its integral over any volume containing the origin is one, and it has units of m^{-3}. Shifted to the point $\mathbf{a}$, it is $\delta^3(\mathbf{r}-\mathbf{a}) = \delta(x-a_x)\,\delta(y-a_y)\,\delta(z-a_z)$, and the sifting property becomes

$$\int f(\mathbf{r})\,\delta^3(\mathbf{r}-\mathbf{a})\,d\tau = f(\mathbf{a}), \tag{1.42}$$

for any volume that contains $\mathbf{a}$.

This answers the question at the start of the section. A point charge q at $\mathbf{r}'$ has the charge density $\rho(\mathbf{r}) = q\,\delta^3(\mathbf{r}-\mathbf{r}')$ (Chapter 3 uses this as (3.10)). Linear charge densities like a very thin string with a static charge and surface charge densities like electrons on the surface of a conductor can be described using fewer one-dimensional delta functions (Figure 1.25). A line of charge λ along the z axis has $\rho = \lambda\,\delta(x)\,\delta(y)$: integrating over a slab of thickness h gives λh, as it should. A sheet of charge σ in the plane $z = 0$ has $\rho = \sigma\,\delta(z)$.

With the point $\mathbf{r}'$ at (x', y', z'), (r', θ', ϕ'), or (s', ϕ', z'), the delta function in each of the three coordinate systems is

$$\begin{aligned}
\delta^3(\mathbf{r}-\mathbf{r}') &= \delta(x-x')\,\delta(y-y')\,\delta(z-z'), && \text{(Cartesian)}\\
\delta^3(\mathbf{r}-\mathbf{r}') &= \frac{1}{r^2\sin\theta}\,\delta(r-r')\,\delta(\theta-\theta')\,\delta(\phi-\phi'), && \text{(spherical)}\\
\delta^3(\mathbf{r}-\mathbf{r}') &= \frac{1}{s}\,\delta(s-s')\,\delta(\phi-\phi')\,\delta(z-z'). && \text{(cylindrical)}
\end{aligned} \tag{1.43}$$

Each of the one-dimensional delta functions in these products is zero except on a surface on which its own coordinate is constant. In spherical coordinates, $\delta(r-r')$ is zero except on the sphere $r = r'$, $\delta(\theta-\theta')$ except on the cone $\theta = \theta'$, and $\delta(\phi-\phi')$ except on the half-plane $\phi = \phi'$, so their product is zero except where all three surfaces meet: at the single point $\mathbf{r}'$ (Figure 1.26). In cylindrical coordinates the three surfaces are a cylinder, a half-plane, and a plane. (In three dimensions, then, a single delta function is not a point: $\delta(x-2)$ is zero except on the whole plane $x = 2$.) The factors $1/(r^2\sin\theta)$ and $1/s$ cancel the Jacobians in the volume elements (1.13), so that the integral over any volume containing $\mathbf{r}'$ is still one. They also supply the missing units, since $\delta(\theta-\theta')$ and $\delta(\phi-\phi')$, like the angles themselves, have none.[9]

[9] These two forms fail on the z axis, where $s' = 0$ (or $\sin\theta' = 0$) and the angle ϕ' has no definite value, and at the origin, where θ' has none either. For a point there, use the Cartesian form.

1.6.3 Derivatives that the formulas get wrong

The field $\hat{\mathbf{r}}/r^2$ of Figure 1.12(b) has zero divergence everywhere except at the origin. However, the spherical coordinates divergence formula (1.30) would predict

$$\nabla\cdot\left(\frac{\hat{\mathbf{r}}}{r^2}\right) = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\cdot\frac{1}{r^2}\right) + \cancel{\frac{1}{r\sin\theta}\frac{\partial(\sin\theta\cdot 0)}{\partial\theta}} + \cancel{\frac{1}{r\sin\theta}\frac{\partial(0)}{\partial\phi}} = 0.$$

If you took this formula at face value, you would conclude that the divergence is zero everywhere. But the flux of $\hat{\mathbf{r}}/r^2$ out of a sphere of radius R centered on the origin, where $d\mathbf{a} = R^2\sin\theta\,d\theta\,d\phi\,\hat{\mathbf{r}}$, is

$$\oint\frac{\hat{\mathbf{r}}}{R^2}\cdot d\mathbf{a} = \int_0^{2\pi}\int_0^{\pi}\frac{1}{R^2}R^2\sin\theta\,d\theta\,d\phi = 4\pi,$$

whatever the radius (Figure 1.27). By the divergence theorem (Section 1.5.3), the integral of $\nabla\cdot(\hat{\mathbf{r}}/r^2)$ over the ball inside that sphere is therefore 4π, however small the ball. A function that is zero everywhere except at the origin, and whose integral over any volume containing the origin is 4π, is $4\pi\,\delta^3(\mathbf{r})$:

$$\nabla\cdot\left(\frac{\hat{\mathbf{r}}}{r^2}\right) = 4\pi\,\delta^3(\mathbf{r}). \tag{1.44}$$

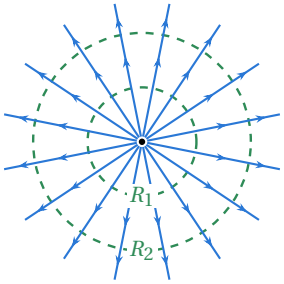


Figure 1.27 Field lines of $\hat{\mathbf{r}}/r^2$ in a plane through the origin, and two spheres centered there (dashed). Every line begins at the origin and crosses both spheres, so the flux through each is the same, 4π.

All of the field lines begin at the origin, and this is where the divergence is. A similar thing is true if the (point) source of divergence is moved to $\mathbf{r}'$. With the separation vector $\boldsymbol{\imath} = \mathbf{r} - \mathbf{r}'$ of Chapter 3, and derivatives taken with respect to the unprimed coordinates,

$$\nabla\cdot\left(\frac{\hat{\boldsymbol{\imath}}}{\imath^2}\right) = 4\pi\,\delta^3(\boldsymbol{\imath}) = 4\pi\,\delta^3(\mathbf{r}-\mathbf{r}'). \tag{1.45}$$

Section 4.1 uses this to derive Gauss's law from Coulomb's law. Also, since $\nabla(1/r) = -\hat{\mathbf{r}}/r^2$, the same result can be written as a Laplacian,

$$\nabla^2\left(\frac{1}{r}\right) = -4\pi\,\delta^3(\mathbf{r}), \qquad \nabla^2\left(\frac{1}{\imath}\right) = -4\pi\,\delta^3(\mathbf{r}-\mathbf{r}'), \tag{1.46}$$

which Section 5.10 needs.

In general, the standard formulas for the divergence, curl, and Laplacian assume a field that is finite and differentiable, and where a field blows up, they can report zero when the answer is a delta function. The curl of $\hat{\boldsymbol{\phi}}/s$ in Figure 1.14(b) is the same story in two dimensions. The formula gives zero for $s \neq 0$, but the circulation around any circle centered on the axis is $\oint(\hat{\boldsymbol{\phi}}/s)\cdot(s\,d\phi\,\hat{\boldsymbol{\phi}}) = 2\pi$, so

$$\nabla\times\left(\frac{\hat{\boldsymbol{\phi}}}{s}\right) = 2\pi\,\delta(x)\,\delta(y)\,\hat{\mathbf{z}}. \tag{1.47}$$

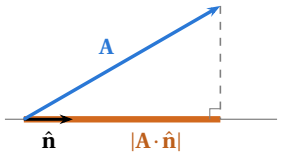


Figure 1.28 The projection (orange) of a vector **A** onto the direction of a unit vector $\hat{\mathbf{n}}$ has length $|\mathbf{A}\cdot\hat{\mathbf{n}}|$.

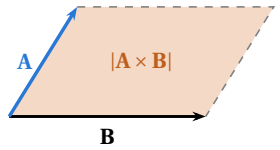


Figure 1.29 The parallelogram (orange) formed by two vectors **A** and **B** has area $|\mathbf{A}\times\mathbf{B}|$.

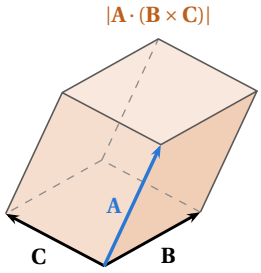


Figure 1.30 The parallelepiped (orange) formed by three vectors **A**, **B**, and **C** has volume $|\mathbf{A}\cdot(\mathbf{B}\times\mathbf{C})|$.

1.7 Vector identities

Vector calculus has a number of identities, most of which we won't prove but which are worth having compiled in one place. Example 1.8 and Exercises P1.22 and P1.23 prove a few of them. In what follows, f and g are scalar fields and **A**, **B**, and **C** are vector fields (or, in the triple products, just vectors).

Triple products. The scalar triple product can be written three ways, by cycling the vectors:

$$\mathbf{A}\cdot(\mathbf{B}\times\mathbf{C}) = \mathbf{B}\cdot(\mathbf{C}\times\mathbf{A}) = \mathbf{C}\cdot(\mathbf{A}\times\mathbf{B}). \tag{1.48}$$

Just as the length of the projection of a vector **A** onto the direction of a unit vector $\hat{\mathbf{n}}$ is given by $|\mathbf{A}\cdot\hat{\mathbf{n}}|$ (Figure 1.28), and the area of the parallelogram formed by two vectors is given by $|\mathbf{A}\times\mathbf{B}|$ (Figure 1.29), the absolute value of the scalar triple product, $|\mathbf{A}\cdot(\mathbf{B}\times\mathbf{C})|$, gives the volume of the parallelepiped formed by three vectors (Figure 1.30). The vector triple product is

$$\mathbf{A}\times(\mathbf{B}\times\mathbf{C}) = \mathbf{B}(\mathbf{A}\cdot\mathbf{C}) - \mathbf{C}(\mathbf{A}\cdot\mathbf{B}), \tag{1.49}$$

which is easiest to remember as the "BAC-CAB rule". Note that $\mathbf{A}\times(\mathbf{B}\times\mathbf{C})$ and $(\mathbf{A}\times\mathbf{B})\times\mathbf{C}$ are not the same: the cross product is not associative.

Product rules. Derivatives of the product of two functions follow patterns somewhat reminiscent of the ordinary product rule, $(fg)' = f'g + fg'$. For the gradient,

$$\nabla(fg) = f\nabla g + g\nabla f, \tag{1.50}$$

$$\nabla(\mathbf{A}\cdot\mathbf{B}) = \mathbf{A}\times(\nabla\times\mathbf{B}) + \mathbf{B}\times(\nabla\times\mathbf{A}) + (\mathbf{A}\cdot\nabla)\mathbf{B} + (\mathbf{B}\cdot\nabla)\mathbf{A}; \tag{1.51}$$

for the divergence,

$$\nabla\cdot(f\mathbf{A}) = f(\nabla\cdot\mathbf{A}) + \mathbf{A}\cdot(\nabla f), \tag{1.52}$$

$$\nabla\cdot(\mathbf{A}\times\mathbf{B}) = \mathbf{B}\cdot(\nabla\times\mathbf{A}) - \mathbf{A}\cdot(\nabla\times\mathbf{B}); \tag{1.53}$$

and for the curl,

$$\nabla\times(f\mathbf{A}) = f(\nabla\times\mathbf{A}) - \mathbf{A}\times(\nabla f), \tag{1.54}$$

$$\nabla\times(\mathbf{A}\times\mathbf{B}) = (\mathbf{B}\cdot\nabla)\mathbf{A} - (\mathbf{A}\cdot\nabla)\mathbf{B} + \mathbf{A}(\nabla\cdot\mathbf{B}) - \mathbf{B}(\nabla\cdot\mathbf{A}). \tag{1.55}$$

Two of these use the operator, $(\mathbf{A}\cdot\nabla)$, which translates to

$$\mathbf{A}\cdot\nabla = A_x\frac{\partial}{\partial x} + A_y\frac{\partial}{\partial y} + A_z\frac{\partial}{\partial z},$$

and acts on each Cartesian component of the vector that follows it.

Second derivatives. Of the derivatives of derivatives, here are the three which are most important for us:

$$\nabla\times(\nabla f) = 0, \tag{1.56}$$

$$\nabla\cdot(\nabla\times\mathbf{A}) = 0, \tag{1.57}$$

$$\nabla\times(\nabla\times\mathbf{A}) = \nabla(\nabla\cdot\mathbf{A}) - \nabla^2\mathbf{A}. \tag{1.58}$$

The first says that a gradient field has no curl, and the second that a curl has no divergence. Both follow from the fundamental theorems. By the gradient theorem, the circulation of ∇f around any small loop is zero, so by (1.23) its curl is zero. And by the curl theorem, the flux of $\nabla \times \mathbf{A}$ out of a small closed surface is zero, because a closed surface has no boundary curve, so by (1.21) its divergence is zero. In the third identity, $\nabla^2\mathbf{A}$ is the Laplacian of a vector field, which in Cartesian coordinates is the Laplacian of each component.

These three identities will come up throughout the book. Because the curl of a gradient field is zero, a field written as the gradient of a function, like $\mathbf{E} = -\nabla V$, automatically has no curl (Chapter 5). Because the divergence of a curl is zero, a field written as a curl, like $\mathbf{B} = \nabla \times \mathbf{A}$, automatically has no divergence (Chapter 12); the same identity shows why Ampère's law needs Maxwell's extra term (Exercise P2.1). The curl of a curl turns up whenever the vector potential is involved, and in the derivation of the wave equation (Section 16.5).

Integration by parts. Integrating the product rule (1.52) over a volume and using the divergence theorem on the left side gives

$$\int_V \mathbf{A}\cdot(\nabla f)\,d\tau = \oint_S f\,\mathbf{A}\cdot d\mathbf{a} - \int_V f\,(\nabla\cdot\mathbf{A})\,d\tau, \tag{1.59}$$

the three-dimensional version of integration by parts. It moves a derivative from one factor to the other, at the cost of a surface term. We will often use this in the limit of a very large volume, where the surface term goes to zero because the fields die out fast enough far away. Section 5.9.2 uses it to find the energy stored in an electric field.

Example 1.8

Prove the product rule (1.52).

Solution: In Cartesian components, $f\mathbf{A} = fA_x\hat{\mathbf{x}} + fA_y\hat{\mathbf{y}} + fA_z\hat{\mathbf{z}}$, so by (1.20) and the ordinary product rule,

$$\begin{aligned}\nabla\cdot(f\mathbf{A}) &= \frac{\partial(fA_x)}{\partial x} + \frac{\partial(fA_y)}{\partial y} + \frac{\partial(fA_z)}{\partial z}\\ &= f\left(\frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}\right) + \left(A_x\frac{\partial f}{\partial x} + A_y\frac{\partial f}{\partial y} + A_z\frac{\partial f}{\partial z}\right)\\ &= f\,(\nabla\cdot\mathbf{A}) + \mathbf{A}\cdot(\nabla f).\end{aligned}$$

(Why is a proof in Cartesian components enough for every coordinate system? Answer in footnote.[10]) The other identities can be proved the same way, one component at a time. Those with cross products take more writing, but no new ideas.

[10] Each side is a scalar whose value at a point does not depend on the coordinates used to compute it: the divergence, for example, is the flux out per unit volume, (1.21). So if the two sides are equal in Cartesian coordinates, they are equal in any.

Reality check: Take $f = x$ and $\mathbf{A} = x\hat{\mathbf{x}}$. The left side is $\nabla\cdot(x^2\hat{\mathbf{x}}) = 2x$. On the right, $\nabla\cdot\mathbf{A} = 1$ and $\nabla f = \hat{\mathbf{x}}$, so $f(\nabla\cdot\mathbf{A}) + \mathbf{A}\cdot(\nabla f) = x + x = 2x$.

1.8 Summary

This summary collects the vector calculus mathematics that the rest of the book uses, so that you can come back to it as a reference. The gradient, divergence, curl, and Laplacian in cylindrical and spherical coordinates are also printed on page 940, near the end of the book, and a few integrals that come up again and again in script-r problems are on the last page, below the physical constants.

Coordinates. The spherical coordinates are (r,θ,ϕ), and the cylindrical ones (s,ϕ,z). The angle θ is measured from the positive z axis, and ϕ from the positive x axis.

$x = r\sin\theta\cos\phi,\ y = r\sin\theta\sin\phi,\ z = r\cos\theta$	spherical to Cartesian
$x = s\cos\phi,\ y = s\sin\phi,\ z = z$	cylindrical to Cartesian
$r = \sqrt{x^2+y^2+z^2},\ s = \sqrt{x^2+y^2} = r\sin\theta$	the distances from the origin and from the z axis
$\theta = \cos^{-1}(z/r)$	the safe way to find θ, which runs from 0 to π
$\cos\phi = x/s,\ \sin\phi = y/s$	both together fix ϕ: `atan2(y, x)` in Python, `ArcTan[x, y]` in Mathematica; $\tan^{-1}(y/x)$ alone gets the quadrant wrong when $x<0$
$\mathbf{r} = x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}}$ $= r\hat{\mathbf{r}}$ $= s\hat{\mathbf{s}} + z\hat{\mathbf{z}}$	the position vector, with no $\hat{\boldsymbol{\theta}}$ or $\hat{\boldsymbol{\phi}}$ terms

Unit vectors. Each unit vector points the way the point moves when its coordinate alone increases. In terms of $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ (Section 1.1.4),

$$\begin{aligned}
\hat{\mathbf{r}} &= \sin\theta\cos\phi\,\hat{\mathbf{x}} + \sin\theta\sin\phi\,\hat{\mathbf{y}} + \cos\theta\,\hat{\mathbf{z}},\\
\hat{\boldsymbol{\theta}} &= \cos\theta\cos\phi\,\hat{\mathbf{x}} + \cos\theta\sin\phi\,\hat{\mathbf{y}} - \sin\theta\,\hat{\mathbf{z}},\\
\hat{\boldsymbol{\phi}} &= -\sin\phi\,\hat{\mathbf{x}} + \cos\phi\,\hat{\mathbf{y}},\\
\hat{\mathbf{s}} &= \cos\phi\,\hat{\mathbf{x}} + \sin\phi\,\hat{\mathbf{y}},
\end{aligned}$$

and the other way around,

$$\begin{aligned}
\hat{\mathbf{x}} &= \sin\theta\cos\phi\,\hat{\mathbf{r}} + \cos\theta\cos\phi\,\hat{\boldsymbol{\theta}} - \sin\phi\,\hat{\boldsymbol{\phi}} = \cos\phi\,\hat{\mathbf{s}} - \sin\phi\,\hat{\boldsymbol{\phi}},\\
\hat{\mathbf{y}} &= \sin\theta\sin\phi\,\hat{\mathbf{r}} + \cos\theta\sin\phi\,\hat{\boldsymbol{\theta}} + \cos\phi\,\hat{\boldsymbol{\phi}} = \sin\phi\,\hat{\mathbf{s}} + \cos\phi\,\hat{\boldsymbol{\phi}},\\
\hat{\mathbf{z}} &= \cos\theta\,\hat{\mathbf{r}} - \sin\theta\,\hat{\boldsymbol{\theta}}.
\end{aligned}$$

Both sets are right-handed, for two unit vectors at the same point: $\hat{\mathbf{r}} \times \hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\phi}}$, $\hat{\boldsymbol{\theta}} \times \hat{\boldsymbol{\phi}} = \hat{\mathbf{r}}$, and $\hat{\boldsymbol{\phi}} \times \hat{\mathbf{r}} = \hat{\boldsymbol{\theta}}$; $\hat{\mathbf{s}} \times \hat{\boldsymbol{\phi}} = \hat{\mathbf{z}}$, $\hat{\boldsymbol{\phi}} \times \hat{\mathbf{z}} = \hat{\mathbf{s}}$, and $\hat{\mathbf{z}} \times \hat{\mathbf{s}} = \hat{\boldsymbol{\phi}}$. Unlike $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$, the unit vectors $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\phi}}$, and $\hat{\mathbf{s}}$ change direction from point to point, so convert them to $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ before integrating (Colton's First Commandment, Section 3.5.1).

Line, area, and volume elements.

$d\boldsymbol{\ell} = dr\,\hat{\mathbf{r}} + r\,d\theta\,\hat{\boldsymbol{\theta}} + r\sin\theta\,d\phi\,\hat{\boldsymbol{\phi}}$	the line element in spherical coordinates
$d\boldsymbol{\ell} = ds\,\hat{\mathbf{s}} + s\,d\phi\,\hat{\boldsymbol{\phi}} + dz\,\hat{\mathbf{z}}$	the line element in cylindrical coordinates
$d\tau = r^2\sin\theta\,dr\,d\theta\,d\phi = s\,ds\,d\phi\,dz$	the volume elements; the factors $r^2\sin\theta$ and s are the Jacobians
$d\mathbf{a} = R^2\sin\theta\,d\theta\,d\phi\,\hat{\mathbf{r}}$	the outward area element on a sphere of radius R
$d\Omega = \sin\theta\,d\theta\,d\phi$	the element of solid angle, da/R^2 on a sphere; a whole sphere is 4π

Derivatives. In Cartesian coordinates,

$$\nabla f = \frac{\partial f}{\partial x}\hat{\mathbf{x}} + \frac{\partial f}{\partial y}\hat{\mathbf{y}} + \frac{\partial f}{\partial z}\hat{\mathbf{z}}, \qquad \nabla\cdot\mathbf{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z},$$

$$\nabla\times\mathbf{A} = \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ A_x & A_y & A_z \end{vmatrix}, \qquad \nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}.$$

The cylindrical and spherical forms are in Section 1.4.6 and on page 940. What the derivatives mean:

∇f	points uphill; its magnitude is the steepest slope
$\nabla\cdot\mathbf{A}$	nonzero where field lines begin or end
$\nabla\times\mathbf{A}$	nonzero where a small loop has a net circulation (a paddle wheel turns)
$\nabla^2 f = \nabla\cdot\nabla f$	compares f with its average nearby

The fundamental theorems.

$\int_{\mathbf{a}}^{\mathbf{b}} \nabla f\cdot d\boldsymbol{\ell} = f(\mathbf{b}) - f(\mathbf{a})$	gradient theorem: the same for every path
$\int_{\mathscr{V}} (\nabla\cdot\mathbf{A})\,d\tau = \oint_{\mathscr{S}} \mathbf{A}\cdot d\mathbf{a}$	divergence theorem: $d\mathbf{a}$ points outward
$\int_{\mathscr{S}} (\nabla\times\mathbf{A})\cdot d\mathbf{a} = \oint_{\mathscr{P}} \mathbf{A}\cdot d\boldsymbol{\ell}$	curl theorem: a right-hand rule links $d\mathbf{a}$ and the direction around $\mathscr{P}$

The separation vector and delta functions.

$\boldsymbol{\mathscr{r}} = \mathbf{r} - \mathbf{r}',\ \mathscr{r} = \lvert\mathbf{r} - \mathbf{r}'\rvert,\ \hat{\boldsymbol{\mathscr{r}}} = \boldsymbol{\mathscr{r}}/\mathscr{r}$	from a source point $\mathbf{r}'$ to the field point $\mathbf{r}$ (Chapter 3)
$\nabla\left(\frac{1}{\mathscr{r}}\right) = -\frac{\hat{\boldsymbol{\mathscr{r}}}}{\mathscr{r}^2}, \quad \nabla'\left(\frac{1}{\mathscr{r}}\right) = +\frac{\hat{\boldsymbol{\mathscr{r}}}}{\mathscr{r}^2}$	∇' differentiates with respect to the source coordinates
$\int f(x)\,\delta(x - x_0)\,dx = f(x_0)$	the sifting property; the range must include x_0
$\delta^3(\mathbf{r}) = \delta(x)\,\delta(y)\,\delta(z)$	units of m^{-3}; a point charge at $\mathbf{r}'$ has $\rho = q\,\delta^3(\mathbf{r} - \mathbf{r}')$
$\nabla \cdot \left(\frac{\hat{\boldsymbol{\mathscr{r}}}}{\mathscr{r}^2}\right) = 4\pi\,\delta^3(\mathbf{r} - \mathbf{r}')$	where the derivative formulas report zero, be careful!
$\nabla^2\left(\frac{1}{\mathscr{r}}\right) = -4\pi\,\delta^3(\mathbf{r} - \mathbf{r}')$	the same fact, written as a Laplacian

Identities. Here f is a scalar field and **A**, **B**, and **C** are vector fields (or, in the first two lines, just vectors); Section 1.7 has the rest.

$\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C}) = \mathbf{B} \cdot (\mathbf{C} \times \mathbf{A}) = \mathbf{C} \cdot (\mathbf{A} \times \mathbf{B})$	the scalar triple product
$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = \mathbf{B}\,(\mathbf{A} \cdot \mathbf{C}) - \mathbf{C}\,(\mathbf{A} \cdot \mathbf{B})$	the BAC-CAB rule
$\nabla \times (\nabla f) = 0$	a gradient field has no curl
$\nabla \cdot (\nabla \times \mathbf{A}) = 0$	a curl has no divergence
$\nabla \times (\nabla \times \mathbf{A}) = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2\mathbf{A}$	the curl of a curl
$\nabla \cdot (f\mathbf{A}) = f\,(\nabla \cdot \mathbf{A}) + \mathbf{A} \cdot (\nabla f)$	with the divergence theorem, also gives integration by parts in three dimensions, (1.59)
$\nabla \times (f\mathbf{A}) = f\,(\nabla \times \mathbf{A}) - \mathbf{A} \times (\nabla f)$	a product rule for the curl
$\nabla \cdot (\mathbf{A} \times \mathbf{B}) = \mathbf{B} \cdot (\nabla \times \mathbf{A}) - \mathbf{A} \cdot (\nabla \times \mathbf{B})$	used for the energy carried by fields (Chapter 20)

Approximations. These two series are the usual tools for taking limits, such as the field far from a charge distribution, and for numerical methods.

$f(x_0 + h) = f(x_0) + f'(x_0)\,h + \frac{1}{2}f''(x_0)\,h^2 + \cdots$	the Taylor series, useful when h is small
$(1 + x)^n \approx 1 + nx$	the binomial approximation, for $\lvert x\rvert \ll 1$ and any power n; for example, $(1 + x)^{-1/2} \approx 1 - x/2$

Exercises

Exercises for 1.1 Coordinate systems

Warm-up questions Read Section 1.1, then answer before class

W1.1 A classmate reasons that since the spherical coordinates of a point are (r,θ,ϕ), its position must be $\mathbf{r} = r\,\hat{\mathbf{r}} + \theta\,\hat{\boldsymbol{\theta}} + \phi\,\hat{\boldsymbol{\phi}}$, just as $\mathbf{r} = x\,\hat{\mathbf{x}} + y\,\hat{\mathbf{y}} + z\,\hat{\mathbf{z}}$ in Cartesian coordinates. Explain what is wrong. In the correct expression, $\mathbf{r} = r\,\hat{\mathbf{r}}$, where is the information about θ and ϕ?

W1.2 True or false? Give a one-sentence reason for each. (a) $\hat{\boldsymbol{\theta}}$ points the same way at every point of the xy plane. (b) $\hat{\boldsymbol{\phi}}$ points the same way at every point of the x axis, other than the origin. (c) $\hat{\mathbf{s}}$ points the same way at every point of a line parallel to the z axis.

W1.3 In your own words, what is a scale factor? Why is the scale factor for ϕ in spherical coordinates $r\sin\theta$, and not r?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P1.1 Find the spherical coordinates (r,θ,ϕ) and the cylindrical coordinates (s,ϕ,z) of each of the following points. Give the angles in degrees, and watch the quadrants.

(a) $(x,y,z) = (1,1,\sqrt{2})$

(b) $(x,y,z) = (-1,1,-\sqrt{2})$

(c) $(x,y,z) = (0,-2,0)$

(d) $(x,y,z) = (-\sqrt{3},-1,2)$

P1.2 Sketch the point $(x,y,z) = (-1,-1,\sqrt{2})$ and the four unit vectors $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\phi}}$, and $\hat{\mathbf{s}}$ at that point. Then find r, θ, ϕ, and s, express the four unit vectors in terms of $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$, and check that each one points the way your sketch shows.

P1.3 This exercise checks the unit vectors (1.7) and obtains the inverse relations (1.8).

(a) Show that $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\phi}}$, as given by (1.7), each have length one and are perpendicular to each other, and that $\hat{\mathbf{r}} \times \hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\phi}}$.

(b) Any vector can be written in terms of its projections onto three perpendicular unit vectors; for example, $\hat{\mathbf{x}} = (\hat{\mathbf{x}}\cdot\hat{\mathbf{r}})\,\hat{\mathbf{r}} + (\hat{\mathbf{x}}\cdot\hat{\boldsymbol{\theta}})\,\hat{\boldsymbol{\theta}} + (\hat{\mathbf{x}}\cdot\hat{\boldsymbol{\phi}})\,\hat{\boldsymbol{\phi}}$. Use this to write $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ in terms of $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\phi}}$.

(c) Do the same in cylindrical coordinates: write $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ in terms of $\hat{\mathbf{s}}$ and $\hat{\boldsymbol{\phi}}$.

P1.4 The angle θ and the unit vector $\hat{\boldsymbol{\theta}}$ can be defined relative to any polar axis $\hat{\mathbf{p}}$, not just $\hat{\mathbf{z}}$ (Figure 1.9). For each of the following polar axes and points, sketch the axis, the point, and $\hat{\boldsymbol{\theta}}$, and then find θ and $\hat{\boldsymbol{\theta}}$: (i) $\hat{\mathbf{p}} = \hat{\mathbf{z}}$ and the point $(0, \sqrt{3}, 1)$; (ii) $\hat{\mathbf{p}} = -\hat{\mathbf{z}}$ and the point $(2,0,0)$; (iii) $\hat{\mathbf{p}} = \hat{\mathbf{x}}$ and the point $(1,0,-1)$; (iv) $\hat{\mathbf{p}} = -\hat{\mathbf{y}}$ and the point $(1,1,0)$.

Exercises for 1.2 Scalar and vector fields

Warm-up questions Read Section 1.2, then answer before class

W1.4 This section compares the electric field and the electric potential with the flow of heat and the temperature. In what ways are they similar? Different?

W1.5 You are given a field written in spherical components, $\mathbf{A} = A_r\,\hat{\mathbf{r}} + A_\theta\,\hat{\boldsymbol{\theta}} + A_\phi\,\hat{\boldsymbol{\phi}}$, with $A_r = 1$ and $A_\theta = A_\phi = 0$ at every point. Is $\mathbf{A}$ the same vector at every point? Describe what it looks like.

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

Exercises for 1.3 Integrating in three dimensions: Jacobians

Warm-up questions Read Section 1.3, then answer before class

W1.6 A cone with its tip at the origin is the surface $\theta = \alpha$ in spherical coordinates. Which two coordinates vary over it? What are the lengths of the two edges of a small patch of it, and what is its area element da?

W1.7 Which coordinate system would you choose for each of these integrals, and why? (a) Find the total charge of a cube of side $2a$ centered on the origin, with $\rho = \rho_0 r^2/a^2$, where r is the distance from the origin. (b) Find the mass of a ball of radius R centered on the origin, whose density is proportional to z^2.

W1.8 Example 1.3 found $\int \hat{\mathbf{r}}\,da = \pi R^2\,\hat{\mathbf{z}}$ over the upper half of a sphere. Without integrating, what is $\int \hat{\mathbf{r}}\,da$ over the whole sphere? Explain.

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P1.5 Calculate the Jacobian determinants

$$\frac{\partial(x,y,z)}{\partial(s,\phi,z)} \quad \text{and} \quad \frac{\partial(x,y,z)}{\partial(r,\theta,\phi)}$$

directly from (1.4) and (1.2), and check that they agree with the volume elements (1.13).

P1.6 A solid cylinder of radius R and length L has a charge density that grows linearly with distance from its axis, $\rho = \rho_0\, s/R$ (Figure 1.31). Find its total charge. Is it more or less than the charge of a cylinder of the same size with a uniform density ρ_0, and is that reasonable?

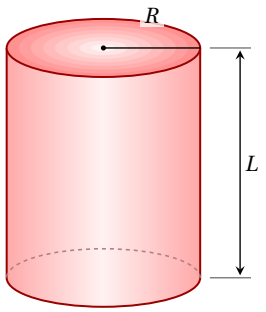


Figure 1.31 The cylinder of Exercise P1.6. The shading darkens where the charge density is larger.

P1.7 Let $\mathcal{V}$ be the upper half of a ball of radius R centered on the origin (the half with $z \geq 0$; Figure 1.32). Evaluate the following integrals. Remember that $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\phi}}$ are not constant: they have hidden θ and ϕ dependences. First put them in terms of the constant unit vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$, which *can* be moved outside the integrals (Colton's First Commandment, Section 3.5.1).

(a) $\displaystyle\int_{\mathcal{V}} \hat{\mathbf{r}} \cos\theta \, d\tau$

(b) $\displaystyle\int_{\mathcal{V}} \hat{\boldsymbol{\theta}} \cos\theta \, d\tau$

(c) $\displaystyle\int_{\mathcal{V}} \hat{\boldsymbol{\phi}} \cos\theta \, d\tau$

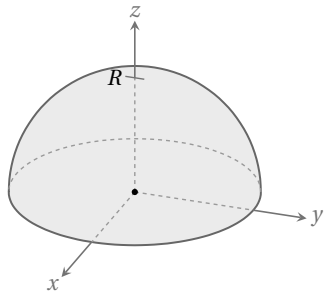


Figure 1.32 The half ball of Exercise P1.7, the region $\mathcal{V}$.

Exercises for 1.4 Derivatives in vector calculus: gradient, divergence, curl, Laplacian

Warm-up questions Read Section 1.4, then answer before class

W1.9 Water flows steadily through a pipe that narrows, and it moves faster where the pipe is narrower. (a) Is the divergence of the water's velocity positive where the water speeds up? (Water does not compress.) (b) The water also moves faster along the middle of the pipe than near its walls. Would a small paddle wheel near a wall, with its axis parallel to the wall and perpendicular to the flow, turn?

W1.10 Write Gauss's law, $\nabla \cdot \mathbf{E} = \rho/\varepsilon_0$, as a sentence about electric field lines. A classmate says that since a static electric field has no curl, its field lines can never curve. Is that right? Explain.

W1.11 The Laplacian $\nabla^2 f$ compares f at a point with the average of f around it. Is $\nabla^2 f$ positive or negative at the bottom of a bowl-shaped f?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P1.8 Calculate the gradient (a vector field) of each of the following (scalar) functions.

(a) $f(x,y,z) = xy + y^2 + z^4$

(b) $f(s,\phi,z) = sz\sin\phi$

(c) $f(r,\theta,\phi) = r\cos\theta$. Your answer should simplify to something very simple. Why?

P1.9 Calculate the divergence (a scalar field) of each of the following (vector) functions.

(a) $\mathbf{A}(x,y,z) = x^2 y\,\hat{\mathbf{x}} + y^3\,\hat{\mathbf{y}} + z^4\,\hat{\mathbf{z}}$

(b) $\mathbf{A}(s,\phi,z) = s^2\sin\phi\,\hat{\mathbf{s}}$

(c) $\mathbf{A}(r,\theta,\phi) = r^2\cos\theta\,\hat{\mathbf{r}} + r^2\sin\theta\,\hat{\boldsymbol{\phi}}$

P1.10 Calculate the curl (a vector field) of each of the three vector functions of Exercise P1.9.

P1.11 Section 1.4.2 said that the inverse square is special. This exercise shows how.

(a) Show that away from the origin the divergence of $\hat{\mathbf{r}}/r^n$ is $(2-n)/r^{n+1}$, so that among radial fields of this form only the inverse square has zero divergence.

(b) Find the curl of $s^n\,\hat{\boldsymbol{\phi}}$ away from the z axis, and show that among circulating fields of this form only $\hat{\boldsymbol{\phi}}/s$ has zero curl.

P1.12 For each of the following fields, sketch the field in the xy plane. Then predict from your sketch where the divergence is positive, negative, or zero, by imagining a small box at several places and comparing how much flows out of it with how much flows in, rather than by asking whether the arrows spread apart. Then calculate the divergence and compare.

(a) $\mathbf{A} = y^2\,\hat{\mathbf{y}}$

(b) $\mathbf{A} = x\,\hat{\mathbf{x}} - y\,\hat{\mathbf{y}}$

P1.13 For each of the following fields, sketch the field in the xy plane, predict from your sketch where a small paddle wheel would turn and in which direction, and then calculate the curl and compare.

(a) $\mathbf{A} = x^2\,\hat{\mathbf{y}}$

(b) $\mathbf{A} = (x\,\hat{\mathbf{x}} + y\,\hat{\mathbf{y}})/(x^2+y^2)$

(c) $\mathbf{A} = s^2\,\hat{\boldsymbol{\phi}}$

Exercises for 1.5 The fundamental theorems

Warm-up questions Read Section 1.5, then answer before class

W1.12 When you push a box around a closed path on a floor, friction does negative work on it all the way around, so the total work is not zero. Use the gradient theorem to explain why the friction force cannot be written as $-\nabla U$ for any function U.

W1.13 The divergence theorem uses a closed surface, and the curl theorem a surface that is not closed. Suppose you wanted to invent a new curl theorem, one for the flux of $\nabla \times \mathbf{A}$ out of a closed surface such as a sphere. What is the boundary curve of a closed surface, and what would your new theorem say about that flux?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P1.14 Let $f = x^2 z + y^2$. Calculate $\int \nabla f \cdot d\boldsymbol{\ell}$ from $(0,0,0)$ to $(1,1,1)$ along each of the following paths (Figure 1.33), and compare both answers with $f(1,1,1) - f(0,0,0)$.

(a) The straight line between the two points.

(b) The path that goes along the x axis to $(1,0,0)$, then parallel to the y axis to $(1,1,0)$, and then parallel to the z axis to $(1,1,1)$.

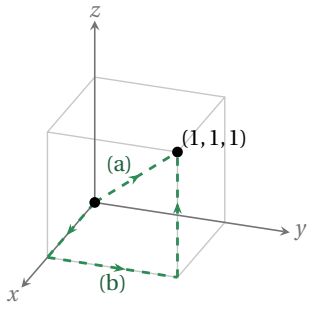


Figure 1.33 The two paths of Exercise P1.14.

P1.15 Check the divergence theorem for each of the following (Figure 1.34).

(a) The field $\mathbf{v} = r^2 \hat{\mathbf{r}}$ over a ball of radius R centered on the origin.

(b) The field $\mathbf{v} = \mathbf{r} = x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}}$ over the cube $0 \le x \le 1$, $0 \le y \le 1$, $0 \le z \le 1$.

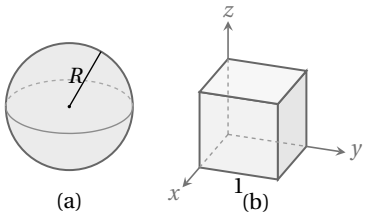


Figure 1.34 The two regions of Exercise P1.15: (a) the ball; (b) the cube.

P1.16 Check the curl theorem for the field $\mathbf{A} = x^2 y\hat{\mathbf{x}} + y^3\hat{\mathbf{y}} + z^4\hat{\mathbf{z}}$ of Exercise P1.9(a), on the square $0 \le x \le 1$, $0 \le y \le 1$ in the xy plane, with $d\mathbf{a}$ in the $+z$ direction (Figure 1.35).

Exercises for 1.6 The Dirac delta function

Warm-up questions Read Section 1.6, then answer before class

W1.14 A classmate says that $\delta(x)$ is infinite at $x = 0$ and zero everywhere else, so $2\delta(x)$ and $\delta(x)$ must be the same thing, since twice infinity is still infinity. What is wrong with this argument?

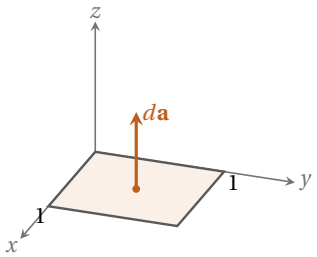


Figure 1.35 The square of Exercise P1.16 and the direction of $d\mathbf{a}$.

W1.15 A computer algebra program reports that the curl of $\hat{\boldsymbol{\phi}}/s$ is zero, and a classmate concludes that the circulation of $\hat{\boldsymbol{\phi}}/s$ around every closed loop is zero. For which loops is the conclusion right, and for which is it wrong? Why does the program miss the difference?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P1.17 Evaluate the following integrals.

(a) $\displaystyle\int_{-2}^{2} (x^3+1)\,\delta(x-1)\,dx$

(b) $\displaystyle\int_{0}^{\infty} e^{-x}\,\delta(x+2)\,dx$

(c) $\displaystyle\int_{0}^{\pi/4} \cos x\,\delta(x-\pi/3)\,dx$

(d) $\displaystyle\int (x^2+y^2+z^2)\,\delta^3(\mathbf{r}-\mathbf{a})\,d\tau$, over all space, with $\mathbf{a}=(1,2,3)$

(e) $\displaystyle\int_{\mathcal{V}} \delta^3(\mathbf{r}-\mathbf{b})\,d\tau$, where $\mathcal{V}$ is the cube $0\le x,y,z\le 2$ and $\mathbf{b}=(1,1,3)$

P1.18 A "point charge" located at (x_0,y_0,z_0) can be thought of as a volume charge density which has delta functions in the x, y, and z directions that serve to localize it in each of the three dimensions:

$$\rho(x,y,z) = q\,\delta(x-x_0)\,\delta(y-y_0)\,\delta(z-z_0).$$

(a) What does this equation imply about the units of the delta function? Do they match the units implied by the area of the delta function needing to equal one, $\int \delta(x)\,dx = 1$? Explain.

(b) Similarly, the ρ associated with a line charge density has delta functions in two dimensions, and the ρ associated with a surface charge density has a delta function in one dimension. For each of the following, describe what the charge distribution looks like, and make a quick sketch of it: (i) $\rho(x,y,z) = c_1\,\delta(x)\,\delta(y+3)$; (ii) $\rho(r,\theta,\phi) = c_2\,\delta(r-3)$; (iii) $\rho(s,\phi,z) = c_3\,\delta(z)\,\delta(s-2)$. (Distances are in meters.) Also find the units of the three constants, c_1, c_2, and c_3.

P1.19 The field $\hat{\boldsymbol{\phi}}/s$ has zero curl everywhere except on the z axis, where (1.47) puts a delta function. Check this a second way, with a field that is finite on the axis.

(a) Replace $\hat{\boldsymbol{\phi}}/s$ by the smooth field $s\hat{\boldsymbol{\phi}}/(s^2+\varepsilon^2)$, which is the same far from the axis but finite on it, and show that its curl is $2\varepsilon^2\,\hat{\mathbf{z}}/(s^2+\varepsilon^2)^2$.

(b) Show that the integral of this curl over the whole xy plane is $2\pi\hat{\mathbf{z}}$, whatever the value of ε.

P1.20 Consider the field $\mathbf{A} = (x\hat{\mathbf{x}} + y\hat{\mathbf{y}})/(x^2 + y^2)$, the field of Exercise P1.13(b).

(a) Sketch the field in the xy plane. Then predict from your sketch where the divergence is positive, negative, or zero, by imagining a small box at several places and comparing how much flows out of it with how much flows in, rather than by asking whether the arrows spread apart. Then calculate the divergence and compare.

(b) The formula cannot be used on the z axis, where the field is infinite. Find the flux of $\mathbf{A}$ out of a cylinder of radius s and height h centered on the z axis, and use it to show that $\nabla\cdot\mathbf{A} = 2\pi\,\delta(x)\,\delta(y)$.

P1.21 A field that stops abruptly at a surface has a curl or a divergence on that surface, even though the formulas give zero on both sides of it. Consider the field

$$\mathbf{F} = \begin{cases} F_0\,\hat{\mathbf{x}}, & y < 0, \\ 0, & y > 0, \end{cases}$$

which is uniform below the plane $y = 0$, points along that plane, and stops there (Figure 1.36a).

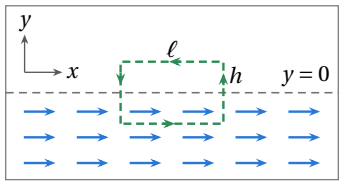
(a) Show that $\nabla\times\mathbf{F} = 0$ everywhere except on the plane $y = 0$.

(b) Find the circulation of $\mathbf{F}$ around a rectangle in the xy plane, with sides of length ℓ along x and h along y, centered on the plane $y = 0$ and traversed counterclockwise as seen from the $+z$ axis. Show that it does not depend on h, so that the circulation per unit area grows without limit as $h \to 0$.

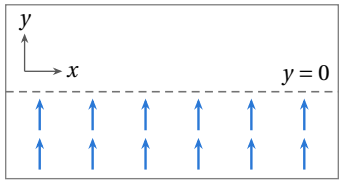
(c) Write $\mathbf{F}$ with the step function of Section 1.6.1, $\mathbf{F} = F_0\,\Theta(-y)\,\hat{\mathbf{x}}$, and show that $\nabla\times\mathbf{F} = F_0\,\delta(y)\,\hat{\mathbf{z}}$. Check that the curl theorem, applied to the rectangle of part (b), gives the circulation you found there.

(d) Now let the field point across the plane instead: $\mathbf{G} = G_0\,\hat{\mathbf{y}}$ for $y < 0$ and $\mathbf{G} = 0$ for $y > 0$ (Figure 1.36b). Show that $\mathbf{G}$ has no curl anywhere, but that $\nabla\cdot\mathbf{G} = -G_0\,\delta(y)$.

(e) In general, both the component of a field along a surface and the component across it can jump there. Which jump gives the field a curl on the surface, and which gives it a divergence?



(a) **F**



(b) **G**

Figure 1.36 (a) The field $\mathbf{F}$ of Exercise P1.21, which points along the plane $y = 0$ and stops there, and the rectangle of part (b). (b) The field $\mathbf{G}$ of part (d), which points across the plane. Both are seen from the $+z$ axis.

Exercises for 1.7 Vector identities

Warm-up questions Read Section 1.7, then answer before class

W1.16 In the three-dimensional integration by parts, (1.59), the surface term is often zero because the surface is taken far away. Take the surface to be a sphere of radius R. If $f\mathbf{A}$ falls off as $1/r^n$ far away, for which values of n does the surface term go to zero as $R \to \infty$?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P1.22 Using Cartesian components, show the following. (You will need the fact that the order of partial differentiation does not matter, for example $\partial^2 f/\partial x\,\partial y = \partial^2 f/\partial y\,\partial x$.)

(a) $\nabla \times (\nabla f) = 0$ for any scalar field f.

(b) $\nabla \cdot (\nabla \times \mathbf{A}) = 0$ for any vector field $\mathbf{A}$.

(c) The x component of $\nabla \times (\nabla \times \mathbf{A})$ equals the x component of $\nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A}$.

P1.23 Equation (1.59) is the three-dimensional version of integration by parts.

(a) Derive (1.59) from the product rule (1.52) and the divergence theorem.

(b) Check (1.59) for $f = z$ and $\mathbf{A} = \hat{\mathbf{z}}$, over a ball of radius R centered on the origin, by calculating each of the three integrals.

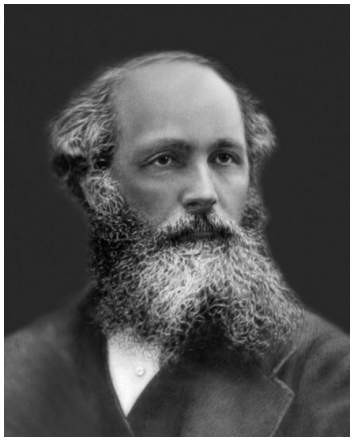
Exercises for 1.8 Summary

P1.24 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as the gradient of a function of two variables drawn as arrows over its contour plot, the divergence and curl of a vector field calculated numerically from its values on a grid, the field lines of a vector field traced numerically, the unit vectors of spherical coordinates drawn at points over a sphere, the curl theorem checked numerically on a curved surface, or a delta function approached as the limit of taller and narrower functions.

Chapter 2

Maxwell's Equations and the Helmholtz Theorem

Now that we have discussed the basics of vector calculus we are in a position where we can present Maxwell's equations, which are in large part what the rest of the book is about.



James Clerk Maxwell (1831–1879, Scottish) Maxwell wrote his first scientific paper, on drawing oval curves, at 14; it was read to the Royal Society of Edinburgh by James Forbes, because Maxwell was thought too young to present it. After Cambridge he held chairs at Aberdeen and at King's College London, and in 1871 he became the first Cavendish Professor of Physics at Cambridge. He won the Adams Prize (1859) for showing that the rings of Saturn must consist of many small particles, gave the first demonstration of color photography (1861), and derived the distribution of the speeds of molecules in a gas. His theory of the electromagnetic field (Section 2.4) united electricity, magnetism, and light. He died of abdominal cancer at 48. Einstein wrote, "One scientific epoch ended and another began with James Clerk Maxwell." (Wikipedia)

2.1 What the equations are

Maxwell's equations describe the divergence and curl of the electric and magnetic fields $\mathbf{E}$ and $\mathbf{B}$. In these equations, ρ (a scalar field) is the charge density and $\mathbf{J}$ (a vector field) describes the current density. The constants ε_0 and μ_0 come from the choice of units; Chapter 17 discusses where they come from.

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0} \qquad \text{(Gauss's law)} \tag{2.1}$$

$$\nabla \cdot \mathbf{B} = 0 \qquad \text{(Gauss's law for magnetism)} \tag{2.2}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \qquad \text{(Faraday's law)} \tag{2.3}$$

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \mu_0 \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} \qquad \text{(Ampère–Maxwell law)} \tag{2.4}$$

2.2 What the equations mean

The right-hand sides of the equations tell us what can produce electric and magnetic fields, and the left-hand sides tell us the behavior of the fields so produced.

Gauss's law: Charges/charge densities ρ can produce electric fields with divergence. The electric field lines thus produced will stream out of (if positive) or into (if negative) the regions of space containing the charges/charge densities.

Faraday's law: A changing magnetic field can produce an electric field with curl. The electric field lines thus produced will curl around the direction of negative magnetic field change.



Hermann von Helmholtz (1821–1894, German) Trained as a physician, Helmholtz published one of the founding statements of the conservation of energy, *Über die Erhaltung der Kraft* (1847). He measured the speed of nerve signals (1849), invented the ophthalmoscope (1851), and wrote standard works on the physiology of vision and of hearing. The theorem of this chapter appears in his 1858 paper on vortex motion in fluids, though Stokes had described the decomposition in 1849. In 1879 Helmholtz proposed a Berlin Academy prize for detecting an effect that Maxwell's theory predicted; his student Heinrich Hertz judged the problem too difficult at the time, but in 1886–88 Hertz produced and detected electromagnetic waves. Helmholtz had been raised to the nobility in 1883. (Wikipedia)

Gauss's law for magnetism: There is nothing we know of that can produce magnetic fields with divergence. In other words, as far as anyone knows, there are no magnetic monopoles.

Ampère–Maxwell law: There are two things which can produce magnetic fields, and each of these things produces a field with curl. Currents/current densities $\mathbf{J}$ are one of those things: the magnetic field lines thus produced will curl around the direction of the currents/current densities. And finally, a changing electric field will also produce a magnetic field, with the magnetic field lines curling around the direction of (positive) electric field change.[1]

2.3 The Helmholtz theorem

Maxwell's equations specify only the divergence and the curl of each field. Is that enough to determine the fields? The answer is given by the **Helmholtz theorem**:

> If the divergence and the curl of a vector field $\mathbf{F}$ are known everywhere, and $\mathbf{F}$ goes to zero far from the origin, then $\mathbf{F}$ is uniquely determined.

(The divergence and curl must themselves go to zero far away, faster than $1/r^2$.) This is the three-dimensional version of a familiar fact from one-variable calculus. If you know the derivative $f'(x)$ everywhere, you know $f(x)$ up to a constant of integration, and a boundary condition, such as the value of f at one point, fixes the constant. Here the divergence and curl play the role of f', and the condition that $\mathbf{F}$ vanish far away plays the role of the boundary condition.

The theorem also says how to construct $\mathbf{F}$. If $\nabla\cdot\mathbf{F} = D(\mathbf{r})$ and $\nabla\times\mathbf{F} = \mathbf{C}(\mathbf{r})$, then

$$\mathbf{F} = -\nabla U + \nabla\times\mathbf{W}, \qquad \text{where} \quad U(\mathbf{r}) = \frac{1}{4\pi}\int \frac{D(\mathbf{r}')}{\mathscr{r}}\,d\tau', \quad \mathbf{W}(\mathbf{r}) = \frac{1}{4\pi}\int \frac{\mathbf{C}(\mathbf{r}')}{\mathscr{r}}\,d\tau', \tag{2.5}$$

with $\mathscr{r} = |\mathbf{r}-\mathbf{r}'|$ and the integrals over all space. (Since the divergence of a curl is zero, $\mathbf{C}$ must itself have zero divergence, or no such field exists.) Exercise P2.2 shows that this $\mathbf{F}$ has the right divergence and curl, using $\nabla^2(1/\mathscr{r}) = -4\pi\,\delta^3(\mathbf{r}-\mathbf{r}')$ from (1.46), and that it is the only such field that goes to zero far away. Writing a field as a curl-free part, $-\nabla U$, plus a divergence-free part, $\nabla\times\mathbf{W}$, is called the **Helmholtz decomposition**.

For static fields the theorem says something very concrete. In electrostatics, $\nabla\cdot\mathbf{E} = \rho/\varepsilon_0$ and $\nabla\times\mathbf{E} = 0$, so the electric field is determined by the charge density alone. Equation (2.5), with $D = \rho/\varepsilon_0$ and $\mathbf{C} = 0$, gives $\mathbf{E} = -\nabla U$ with $U = (1/4\pi\varepsilon_0)\int(\rho/\mathscr{r})\,d\tau'$, which is Coulomb's law in another form (Section 5.5). In magnetostatics, $\nabla\cdot\mathbf{B} = 0$ and $\nabla\times\mathbf{B} = \mu_0\mathbf{J}$, and the theorem determines $\mathbf{B}$ from the current density alone, as the Biot–Savart law of Section 10.5 does. For fields that change with time the situation is subtler, because the curls in Faraday's law and the Ampère–Maxwell law involve the fields themselves. Maxwell's equations

[1] Section 16.3 looks more closely at what it means for a changing electric field to produce a magnetic field.

are then best read as telling how the fields evolve: given the fields at one instant, and ρ and $\mathbf{J}$ at all times, they determine the fields at all later times. Either way, Maxwell's equations give a *complete* picture of the electric and magnetic fields, insofar as they themselves are complete — which we believe they are, within the framework of classical (non-quantum) physics.

2.4 A short history of the equations

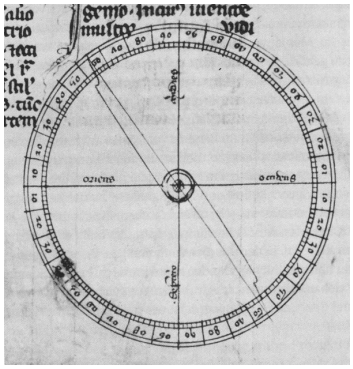
Maxwell's equations carry Maxwell's name, but much of the physics in them was discovered by others, and they were not written in the form above until after his death. Table 2.1 lists the main steps.

1269	Peregrinus	magnetic poles come in pairs
1785	Coulomb	inverse-square force between charges
1820	Ørsted	an electric current deflects a compass needle
1820–26	Ampère	forces between currents
1831	Faraday	electromagnetic induction
1834	Lenz	direction of the induced current
1835	Gauss	Gauss's law
1845	Neumann	mathematical law of induction
1855	Maxwell	the circuital law (Ampère's law)
1861–62	Maxwell	displacement current; light is an electromagnetic wave
1865, 1873	Maxwell	*A Dynamical Theory*; the *Treatise*
1885	Heaviside	the four equations in vector form
1886–88	Hertz	radio waves produced and detected

Table 2.1 The main steps toward Maxwell's equations.

The oldest piece is $\nabla \cdot \mathbf{B} = 0$. In 1269 Petrus Peregrinus de Maricourt described the north and south poles of a lodestone, and he found that when he cut one in two, each piece had a north and a south pole of its own. Magnetic poles come in pairs, which means no magnetic monopoles — what we now write as $\nabla \cdot \mathbf{B} = 0$. Gauss's law follows from Coulomb's inverse-square law (Chapter 4), which Coulomb established with his torsion balance in 1785; Gauss stated the law in 1835, and Lagrange had found the same mathematical result in 1773, in a study of the gravitational attraction of ellipsoids.

In 1820 Hans Christian Ørsted found that an electric current deflects a compass needle, the first clear demonstration that electricity and magnetism are connected. Ampère took up the subject at once, and over the next six years he worked out the forces between currents. The circuital law that carries his name, which relates the magnetic field around a loop to the current through it, was first written down by Maxwell, in 1855. Faraday discovered electromagnetic induction in 1831; Lenz found the rule for the direction of the induced current (the minus sign in Faraday's law) in 1834; and Franz Neumann gave induction its first mathematical formulation in 1845.



Petrus Peregrinus de Maricourt (fl. 1269, French) Little is known of Peregrinus's life beyond his one surviving work on magnetism, a letter to a friend that he finished on 8 August 1269 in the camp of the French army besieging Lucera, in southern Italy, where he was serving as an engineer. He took his name from the village of Maricourt, in Picardy. The letter is the earliest surviving written account of the poles of a magnet. It explains how to find the north and south poles of a lodestone, shows that like poles repel and unlike poles attract, and describes a floating compass and a new, pivoted one. It was copied by hand for nearly three centuries and first printed in 1558. The picture is a drawing of a compass from a fourteenth-century copy of the letter (Bodleian Library, Oxford, MS Ashmole 1522). (Wikipedia)

Maxwell started from Faraday's picture of lines of force. His first paper on the subject, "On Faraday's Lines of Force" (1855–56), gave that picture a mathematical form, and Faraday wrote to him in 1857, "I was at first almost frightened when I saw such mathematical force made to bear upon the subject and then wondered to see that the subject stood it so well." Later, while Maxwell was professor at King's College London (1860–65), he often attended lectures at the Royal Institution, where Faraday worked, and the two men, nearly forty years apart in age, came into regular contact.

Maxwell's own contribution was to put all of this into a single mathematical theory, and to add one term. In "On Physical Lines of Force" (1861–62) he built a mechanical model of the field. A magnetic field was made of tiny spinning "molecular vortices," and between the vortices were still smaller particles, turning like the idle wheels of a machine, whose motion was the electric current. When he made the medium elastic, so that the particles could be displaced slightly, the model gave him a new term, the *displacement current*, the $\mu_0\varepsilon_0\partial\mathbf{E}/\partial t$ term of (2.4), which he added to Ampère's law. Without it the equations contradict the conservation of charge (Exercise P2.1). With it they predict waves in the electric and magnetic fields, and in 1862 Maxwell calculated that waves in his medium would travel at the measured speed of light: "we can scarcely avoid the inference," he wrote, "that light consists in the transverse undulations of the same medium which is the cause of electric and magnetic phenomena." In "A Dynamical Theory of the Electromagnetic Field" (1865) he set the vortices and idle wheels aside, keeping only a medium filling space and the equations, and *A Treatise on Electricity and Magnetism* (1873) set out the complete theory.



Oliver Heaviside (1850–1925, English) Heaviside was almost entirely self-taught. After working as a telegraph operator in Newcastle upon Tyne from 1870 to 1874, he held no regular position again: he did his research alone, supported first by his brother and later by friends and a government pension. He developed a vector notation of his own, close to the one Gibbs was developing at the same time, and used it to put Maxwell's theory into the form of Section 2.1. In papers of 1885–87 in *The Electrician* he showed that long cables would carry signals with less distortion if inductance were added at intervals along them, and he coined the words *inductance* and *impedance*. The step function of Section 1.6.1 carries his name: he used it to find the current in a circuit after a switch is closed, in the operational calculus he devised for solving differential equations. After radio signals first crossed the Atlantic in 1901, he proposed that a conducting layer high in the atmosphere reflects radio waves; that layer is part of what is now called the ionosphere. He was elected a Fellow of the Royal Society in 1891, and the Institution of Electrical Engineers made him the first recipient of its Faraday Medal. (Wikipedia)

You would not recognize Maxwell's equations in Maxwell's papers, though. Written out one component at a time, his 1865 paper has twenty equations in twenty unknowns. The vector calculus of Chapter 1 was developed independently by Oliver Heaviside and Josiah Willard Gibbs, and in 1885 Heaviside used it to condense Maxwell's theory into the four equations we use today. (Heaviside wrote the curl as "curl", Maxwell's name for it; the $\nabla\cdot$ and $\nabla\times$ notation is Gibbs's.) Finally, in 1886–88 Heinrich Hertz generated and detected electromagnetic waves in his laboratory at Karlsruhe. His waves, from under a meter to several meters long, are what we now call radio waves (at the time, "Hertzian waves"). He showed that they travel at about the speed of light and can be reflected, refracted, and polarized, just as light can. Maxwell had identified light as an electromagnetic wave a quarter of a century earlier, on theoretical grounds; Hertz's experiments are what persuaded most physicists that the theory was right.

2.5 Summary

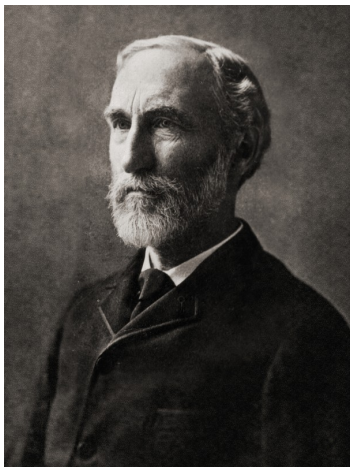
Maxwell's equations give the divergence and the curl of **E** and **B**; by the Helmholtz theorem, that is enough to determine the fields. The rest of the book works out what the equations mean, a piece at a time: Chapters 3 through 9 for static electric fields, Chapters 10 through 13 for static magnetic fields, and Chapters 15 and 16

for the terms with time derivatives.

Maxwell's equations

$\nabla \cdot \mathbf{E} = \dfrac{\rho}{\varepsilon_0}$	Gauss's law: charges are sources and sinks of $\mathbf{E}$
$\nabla \cdot \mathbf{B} = 0$	Gauss's law for magnetism: no magnetic monopoles
$\nabla \times \mathbf{E} = -\dfrac{\partial \mathbf{B}}{\partial t}$	Faraday's law: a changing $\mathbf{B}$ makes an $\mathbf{E}$ with curl
$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \mu_0 \varepsilon_0 \dfrac{\partial \mathbf{E}}{\partial t}$	the Ampère–Maxwell law: currents and a changing $\mathbf{E}$ make a $\mathbf{B}$ with curl

The Helmholtz theorem. A field that goes to zero far away is determined by its divergence and its curl, through (2.5).



J. Willard Gibbs (1839–1903, American) Gibbs lived nearly all his life in New Haven, Connecticut. His Yale doctorate of 1863, on the design of gear teeth, was the first American doctorate in engineering. After three years of study in Paris, Berlin and Heidelberg he became professor of mathematical physics at Yale in 1871, unpaid for the first nine years. His memoir *On the Equilibrium of Heterogeneous Substances* (1876–78) founded chemical thermodynamics, and his *Elementary Principles in Statistical Mechanics* (1902) gave that subject its modern form. The vector analysis he printed privately for his students in 1881 and 1884, which his student Edwin Bidwell Wilson made into a textbook in 1901, is the one this book uses, and the overshoot of a Fourier series at a jump (Figure 6.4) is named after him. (Wikipedia)

Exercises

Exercises for 2.1 What the equations are

Warm-up questions Read Section 2.1, then answer before class

W2.1 Two of the four equations give divergences and two give curls. Which are which? In statics, where nothing changes with time, what do the four equations become?

W2.2 Suppose magnetic monopoles were discovered: particles that are sources of magnetic field in the way that charges are sources of electric field. Which of the four equations would have to change, and what would appear on its right side?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P2.1 Take the divergence of both sides of the Ampère–Maxwell law, (2.4), and use (1.57) and Gauss's law to show that

$$\nabla \cdot \mathbf{J} = -\frac{\partial \rho}{\partial t}.$$

This is the *continuity equation* for electric charge, the local form of charge conservation (Section 20.1). Then show that without Maxwell's term $\mu_0 \varepsilon_0 \partial \mathbf{E}/\partial t$ the same steps would give $\nabla \cdot \mathbf{J} = 0$, which is false wherever charge is piling up, as it does on the plates of a capacitor while it charges.

Exercises for 2.2 What the equations mean

Warm-up questions Read Section 2.2, then answer before class

W2.3 A classmate says that, by Faraday's law, an electric field is always produced by a changing magnetic field. Explain what is wrong, and give an example of an electric field with no changing magnetic field anywhere.

W2.4 In your own words, what can produce a magnetic field, according to the Ampère–Maxwell law? Does each of them produce a field with divergence, or a field with curl?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

Exercises for 2.3 The Helmholtz theorem

Warm-up questions Read Section 2.3, then answer before class

W2.5 Give an example of a field that is not zero anywhere but has zero divergence and zero curl everywhere. Why does it not contradict the Helmholtz theorem? In the comparison with one-variable calculus, what does such a field correspond to?

W2.6 In electrostatics the Helmholtz theorem determines $\mathbf{E}$ from ρ alone. Why can it not, by itself, give $\mathbf{E}$ and $\mathbf{B}$ from ρ and $\mathbf{J}$ when the fields change with time?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P2.2 This exercise proves the Helmholtz theorem of Section 2.3. Let $D(\mathbf{r})$ and $\mathbf{C}(\mathbf{r})$ be given functions that go to zero faster than $1/r^2$ far away, with $\nabla\cdot\mathbf{C}=0$, and define U and $\mathbf{W}$ by (2.5).

(a) Show that the field $\mathbf{F}=-\nabla U+\nabla\times\mathbf{W}$ has $\nabla\cdot\mathbf{F}=D$. Use (1.57) and (1.46); the derivatives act on $\mathbf{r}$, not $\mathbf{r}'$, so they can be taken inside the integrals.

(b) Use (1.56) and (1.58) to show that $\nabla\times\mathbf{F}=\nabla(\nabla\cdot\mathbf{W})-\nabla^2\mathbf{W}$, and show that $-\nabla^2\mathbf{W}=\mathbf{C}$. (It can also be shown that $\nabla\cdot\mathbf{W}=0$, using $\nabla\cdot\mathbf{C}=0$ and the divergence theorem; you may take this as given.)

(c) Now suppose that two fields $\mathbf{F}_1$ and $\mathbf{F}_2$ have the same divergence and the same curl everywhere, and both go to zero far away. Use (1.58) to show that each Cartesian component of their difference satisfies Laplace's equation everywhere. Section 7.1.2 shows that a solution of Laplace's equation is equal at each point to its average over any sphere centered on that point, (7.4). Use this fact, with very large spheres, to argue that the difference is zero everywhere, so that the field is unique.

Exercises for 2.4 A short history of the equations

Warm-up questions Read Section 2.4, then answer before class

W2.7 What did Maxwell add to the equations, and what were the two main consequences of the addition?

W2.8 Maxwell had concluded in 1862 that light is an electromagnetic wave, yet most physicists were persuaded only after 1888. What did Hertz's experiments show that Maxwell's calculation could not?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

Part II

Electrostatics

We had for some Time been of Opinion, that the Electrical Fire was not created by Friction, but collected, being an Element diffused among, and attracted by other Matter, particularly by Water and Metals.

— Benjamin Franklin,
letter to Peter Collinson (1747)*

*B. Franklin to Peter Collinson, 25 May 1747 (dated 1 September 1747 in the 1751, 1754, and 1760 printings and 11 July in the 1769 and 1774 printings), in *The Papers of Benjamin Franklin*, Vol. 3, from the manuscript; https://founders.archives.gov/documents/Franklin/01-03-02-0059. The printed editions read "being really an Element."

Chapter 3

Coulomb's Law

Electrostatics asks one question: given a collection of charges that are not moving, what electric field do they produce? This chapter sets up that question, reviews what you learned about electric force and electric field in introductory physics, and then develops the most general way to answer it, which is to add up the fields produced by every piece of the charge distribution. The method is called the *script-r method*, after the symbol at its center. It always works, although it is not always easy, and the next two chapters develop two powerful shortcuts: Gauss's law and the electric potential. A numerical version of the method, for distributions where the integrals cannot be done by hand, closes the chapter.



Charles-Augustin de Coulomb (1736–1806, French) A military engineer, Coulomb spent eight years in Martinique supervising the construction of Fort Bourbon before turning to research in mechanics and electricity. His 1779 study of friction, *Théorie des machines simples*, is the origin of what engineers still call Coulomb friction. In 1785 he used a torsion balance of his own design to show that the force between two charged spheres varies as the inverse square of their separation. The inverse-square law had been suspected before him — Joseph Priestley proposed it in 1767 — and Henry Cavendish had measured it in the early 1770s, but Cavendish never published his results, and Coulomb's careful measurements are what established the law. (Wikipedia)

3.1 Charge and Coulomb's law

Everything in this book is built on electric charge, so it is worth collecting what is known about it. These facts are established by experiment, and in this book we take them as starting points.

- **There are two kinds of charge**, which we call positive and negative. Like charges repel and opposite charges attract. The names are a convention due to Benjamin Franklin, which is why the electron, discovered by J. J. Thomson in 1897, ended up with a negative charge.

- **Charge is quantized.** Every charge found in isolation is an integer multiple of the elementary charge,

$$e = 1.602\,176\,634 \times 10^{-19}\,\mathrm{C},$$

  which since 2019 is an exact, defined value (Section 17.4). The proton has charge $+e$ and the electron $-e$. Quarks carry $\pm e/3$ or $\pm 2e/3$, but they are never found in isolation.

- **Charge is conserved.** The total charge of an isolated system never changes. Conservation is also *local*: the charge inside any volume can change only by charge flowing in or out through the volume's surface, so charge cannot



Benjamin Franklin (1706–1790, American) Franklin was apprenticed at twelve to his brother, a printer in Boston, and became a successful printer and publisher in Philadelphia. He took up electricity in the mid-1740s, after Peter Collinson, a London merchant, sent a glass tube to Philadelphia's Library Company. In letters to Collinson he described electricity as a single fluid that is collected, not created, by rubbing: a body with more than its normal share is charged plus, or positively, and one with less, minus. The words positive, negative, charge, and battery entered physics with him. In 1752 an experiment he had proposed, to show that lightning is electrical, was carried out in France, and he made his own test with a kite; the lightning rod is his invention. His letters were published in London as *Experiments and Observations on Electricity* (1751), and the Royal Society gave him its Copley Medal in 1753. (Wikipedia)

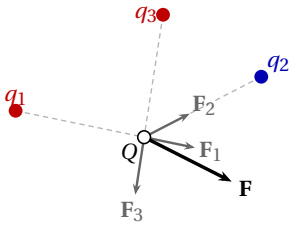


Figure 3.1 Three source charges and a test charge Q (positive charges red, negative blue). The gray arrows are the forces that the source charges would exert on Q one at a time, and the black arrow, their vector sum, is the force on Q.

disappear in one place and reappear somewhere else. When charge is created, as in the production of an electron–positron pair, equal amounts of positive and negative charge are created together.

- **Charge is a scalar,** and it is *relativistically invariant*: observers moving relative to one another all measure the same charge. This will matter in Section 25.10.

- **Charges exert forces on one another.** All charges exert and feel electric forces; moving charges also exert and feel magnetic forces. And the forces obey the principle of *superposition*, discussed next.

3.1.1 The fundamental problem, and how we simplify it

The basic problem of electromagnetism can be stated simply. Suppose you have a set of **source charges** $q_1, q_2, \ldots, q_N$, and you know where they are at all times. What force do they exert on another charge Q, often called the **test charge**? In general that is a hard problem. The force on Q depends not only on where the source charges are, but also on their velocities and accelerations. Worse, information about the source charges travels at the speed of light, so what matters is not where they are now but where they were at some earlier time. And a real object contains an enormous number of charges.

We will get around these difficulties with a series of simplifications, of several different kinds.

1. **Superposition** (a property of nature). The force that each source charge exerts on Q does not depend on the presence of the other source charges. So the forces can be calculated one at a time and then added as vectors (Figure 3.1).

2. **Stationary source charges** (a restriction we impose). If the source charges do not move, the force depends only on their positions. This is *electrostatics*, the subject of this and the next several chapters. (The test charge is allowed to move.)

3. **Charge densities** (an approximation). Instead of tracking every electron and proton, we describe a charged object by a smooth charge density (Section 3.4).

4. **Field first, force second** (a restriction we impose). We find the electric field produced by the source charges, and then the force on any test charge placed in it. This assumes that the test charge does not cause the source charges to move.

5. **Symmetry** (when available). Many problems have a symmetry that tells us the direction of the field, or which terms of a calculation must vanish, before we begin.

3.1.2 Coulomb's law

For two stationary point charges, the force is given by **Coulomb's law**. In magnitude, the force between a charge q and a charge Q a distance d apart is

$$F = \frac{1}{4\pi\varepsilon_0}\frac{qQ}{d^2}, \tag{3.1}$$

where $\varepsilon_0 = 8.854\times10^{-12}\,\mathrm{C^2/(N\,m^2)}$ is the **permittivity of free space**. The combination $1/4\pi\varepsilon_0 = 8.988\times10^9\,\mathrm{N\,m^2/C^2}$ is sometimes given its own symbol, k_e or just k, but this book will always write out $1/4\pi\varepsilon_0$. The direction of the force is given by "like charges repel, opposite charges attract." For example, two 1 C charges placed 1 m apart push each other apart with a force of 8.99×10^9 N.[1]

Coulomb's law is an experimental result. It closely resembles Newton's law of gravitation, and like that law it was found by measurement, not derived. In this book it is the starting point from which the rest of electrostatics follows.

To use Coulomb's law in three dimensions we need its vector form, and for that we need to say where the charges are. Figure 3.2 shows the three vectors that appear in every problem in this chapter:

- $\mathbf{r}'$, the **source point**, the location of the charge q producing the force;
- $\mathbf{r}$, the **field point**, the location of the charge Q feeling the force (or, later, the point where we want to know the field);
- $\boldsymbol{\mathscr{r}} = \mathbf{r} - \mathbf{r}'$, the **separation vector**, which points *from* the source point *to* the field point.

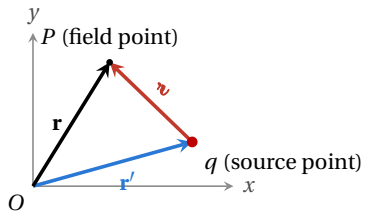


Figure 3.2 The three vectors of electrostatics. The source point $\mathbf{r}'$ (blue) locates the charge; the field point $\mathbf{r}$ (black) is where the force or field is wanted; and the separation vector $\boldsymbol{\mathscr{r}} = \mathbf{r} - \mathbf{r}'$ (red) points from the source point to the field point.

The symbol $\boldsymbol{\mathscr{r}}$ is a script r, a notation introduced by David Griffiths in his *Introduction to Electrodynamics*. Its magnitude is written $\mathscr{r}$, and the unit vector in its direction is $\hat{\boldsymbol{\mathscr{r}}} = \boldsymbol{\mathscr{r}}/\mathscr{r}$. (As everywhere in this book, a symbol without boldface or a hat means the magnitude of the corresponding vector.) The force on Q is then

$$\mathbf{F} = \frac{1}{4\pi\varepsilon_0}\frac{qQ}{\mathscr{r}^2}\hat{\boldsymbol{\mathscr{r}}}. \tag{3.2}$$

If q and Q have the same sign, the force points along $\hat{\boldsymbol{\mathscr{r}}}$, away from q; if they have opposite signs, it points back toward q. The signs of the charges take care of the direction automatically.

There is an equivalent form of (3.2) that is usually easier to calculate with. Since $\hat{\boldsymbol{\mathscr{r}}} = \boldsymbol{\mathscr{r}}/\mathscr{r}$,

$$\mathbf{F} = \frac{1}{4\pi\varepsilon_0}\frac{qQ\boldsymbol{\mathscr{r}}}{\mathscr{r}^3}. \tag{3.3}$$

The cube in the denominator does not make this an inverse-cube law. One factor of $\mathscr{r}$ cancels the length of the vector $\boldsymbol{\mathscr{r}}$ in the numerator, and what remains falls off as $1/\mathscr{r}^2$. The advantage of (3.3) is that $\boldsymbol{\mathscr{r}}$ is simply $\mathbf{r} - \mathbf{r}'$, which can be written down directly in components, while finding $\hat{\boldsymbol{\mathscr{r}}}$ requires an extra step.

☞ The unit vector $\hat{\boldsymbol{\mathscr{r}}}$ is *not* $\hat{\mathbf{r}} - \hat{\mathbf{r}}'$. Subtract the vectors first, then divide by the length of the result.

[1] One coulomb is a huge amount of charge. You would never find that much excess charge in one place, but it makes the numbers in examples easy to follow.

A common mistake is to write the direction of the force as $\hat{\mathbf{r}} - \hat{\mathbf{r}}'$, the difference of the two unit vectors. That is not a unit vector, and it does not point from q to Q either. The correct order of operations is to subtract the position vectors, $\boldsymbol{\imath} = \mathbf{r} - \mathbf{r}'$, and only then divide by the length, $\imath = |\mathbf{r} - \mathbf{r}'|$.

3.2 The electric field and superposition

If several source charges $q_1, q_2, \ldots, q_N$ act on Q, superposition says that the total force is the vector sum of the individual Coulomb forces,

$$\mathbf{F} = \frac{1}{4\pi\varepsilon_0} \sum_{i=1}^{N} \frac{q_i Q \boldsymbol{\imath}_i}{\imath_i^3}, \tag{3.4}$$

where $\boldsymbol{\imath}_i = \mathbf{r} - \mathbf{r}'_i$ is the separation vector from the ith charge to the field point. Every term contains the test charge Q as a factor. That suggests dividing it out, and the result is the **electric field**,

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \sum_{i=1}^{N} \frac{q_i \boldsymbol{\imath}_i}{\imath_i^3}, \tag{3.5}$$

so that the force on any charge Q placed at $\mathbf{r}$ is

$$\mathbf{F} = Q\mathbf{E}. \tag{3.6}$$

In words, the electric field is the force per unit charge that a test charge would feel at $\mathbf{r}$. More carefully, it is that force per charge in the limit where the test charge is so small that it does not disturb the source charges. The force describes the actual effect on a charge that is present; the field describes the possible effect, the force there would be if a charge were placed at that point. The units of the field are N/C, which is the same as V/m. The direction of $\mathbf{E}$ is the direction a *positive* test charge would be pushed: away from positive charges and toward negative ones. For a single point charge q at $\mathbf{r}'$, the sum has one term,

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \frac{q\boldsymbol{\imath}}{\imath^3}. \tag{3.7}$$

In electrostatics, the field can be regarded as a convenient piece of bookkeeping: everything it predicts could also be found from the forces directly. But when charges move, a change in the field travels outward at the speed of light, and in the meantime the field carries energy and momentum of its own. Light itself is a traveling electric and magnetic field. So the field is a physical thing, not just a calculational device, and later chapters will treat it that way.

The most important point about (3.5) is that the fields of the individual charges add as *vectors*. You cannot simply add their magnitudes.

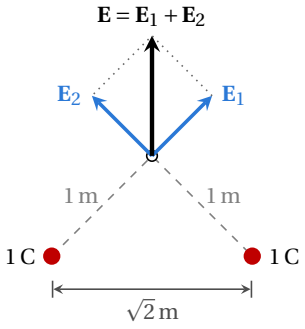
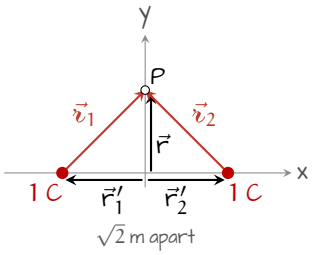


Figure 3.3 Example 3.1: the fields of two 1 C charges at a point 1 m from each. $\mathbf{E}_1$, from the left charge, and $\mathbf{E}_2$, from the right charge, are perpendicular, and their sum points straight up.

Example 3.1

Two 1 C charges are $\sqrt{2}$ m apart. Find the electric field at a point 1 m from each of them, as in Figure 3.3.

Solution:



The first step is a sketch of the charges and the field point P, which also helps us choose coordinates that make the numbers simple: the charges on the x axis at $x = \mp 1/\sqrt{2}\,\text{m}$, so that P is on the y axis, 1 m from each charge. Then we draw the vectors: $\mathbf{r}$ from the origin to P, $\mathbf{r}'_1$ and $\mathbf{r}'_2$ from the origin to the charges, and the separation vectors from the charges to P.

The vectors, with the left charge's in the first column and the right charge's in the second, are

$$\mathbf{r} = \frac{1}{\sqrt{2}}\hat{\mathbf{y}}\,\text{m}$$

$$\mathbf{r}'_1 = -\frac{1}{\sqrt{2}}\hat{\mathbf{x}}\,\text{m} \qquad \mathbf{r}'_2 = \frac{1}{\sqrt{2}}\hat{\mathbf{x}}\,\text{m}$$

$$\boldsymbol{\mathscr{r}}_1 = \mathbf{r} - \mathbf{r}'_1 = \frac{1}{\sqrt{2}}(\hat{\mathbf{x}} + \hat{\mathbf{y}})\,\text{m} \qquad \boldsymbol{\mathscr{r}}_2 = \mathbf{r} - \mathbf{r}'_2 = \frac{1}{\sqrt{2}}(-\hat{\mathbf{x}} + \hat{\mathbf{y}})\,\text{m}$$

$$\mathscr{r}_1 = 1\,\text{m} \qquad \mathscr{r}_2 = 1\,\text{m}.$$

Then (3.5) gives

$$\begin{aligned}\mathbf{E} &= \frac{1}{4\pi\varepsilon_0}\left(\frac{q_1\boldsymbol{\mathscr{r}}_1}{\mathscr{r}_1^3} + \frac{q_2\boldsymbol{\mathscr{r}}_2}{\mathscr{r}_2^3}\right) \\ &= \left(8.99\times10^9\,\frac{\text{N m}^2}{\text{C}^2}\right)\frac{1\,\text{C}}{(1\,\text{m})^3}\,\frac{1}{\sqrt{2}}\left[(\hat{\mathbf{x}}+\hat{\mathbf{y}}) + (-\hat{\mathbf{x}}+\hat{\mathbf{y}})\right]\text{m} \\ &= \sqrt{2}\,(8.99\times10^9\,\text{N/C})\,\hat{\mathbf{y}} = (1.27\times10^{10}\,\text{N/C})\,\hat{\mathbf{y}}.\end{aligned}$$

Each charge alone would produce a field of $8.99\times10^9\,\text{N/C}$ at this point, but the total is not twice that: the two fields are perpendicular, their x components cancel, and only their y components add.

Reality check: The direction makes sense: P is equally far from two equal charges, so nothing favors $+x$ over $-x$, and both charges, below P, push a positive test charge away, upward.

The $\boldsymbol{\mathscr{r}}/\mathscr{r}^3$ form handles signs automatically, which is a real advantage when the field point can be on either side of a charge. The next example shows this.

Example 3.2

A charge $+3q$ sits at the origin and a charge $-2q$ at $(0, a, 0)$. Find the electric field at every point on the x axis.

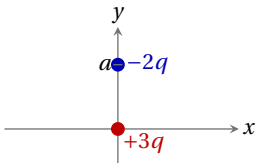
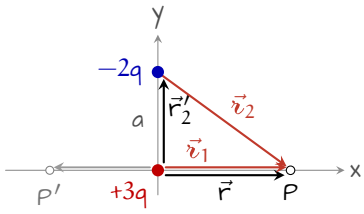


Figure 3.4 Example 3.2: a charge $+3q$ at the origin and $-2q$ at $(0, a, 0)$. The field is wanted everywhere on the x axis.

Solution:



To see what is being asked, we draw the two charges and a field point P on the x axis, at $\mathbf{r} = x\hat{\mathbf{x}}$, where x can be positive or negative (a second field point, P', on the negative side, is drawn in gray). The vector $\mathbf{r}'_2$ goes from the origin to the charge on the y axis ($\mathbf{r}'_1 = 0$), and the separation vectors point from each charge to the field point.

The vectors, with the charge at the origin in the first column and the one on the y axis in the second, are

$$\mathbf{r} = x\hat{\mathbf{x}}$$

$$\mathbf{r}'_1 = 0 \qquad \mathbf{r}'_2 = a\hat{\mathbf{y}}$$

$$\boldsymbol{\imath}_1 = \mathbf{r} - \mathbf{r}'_1 = x\hat{\mathbf{x}} \qquad \boldsymbol{\imath}_2 = \mathbf{r} - \mathbf{r}'_2 = x\hat{\mathbf{x}} - a\hat{\mathbf{y}}$$

$$\imath_1 = |x| \qquad \imath_2 = \sqrt{x^2 + a^2}.$$

(Why is $\imath_1 = |x|$, and not x? Answer in footnote.[2]) Then

$$\mathbf{E}(x) = \frac{q}{4\pi\varepsilon_0}\left[\frac{3x\hat{\mathbf{x}}}{|x|^3} - \frac{2\left(x\hat{\mathbf{x}} - a\hat{\mathbf{y}}\right)}{\left(x^2 + a^2\right)^{3/2}}\right].$$

The first term needs care. For $x > 0$ it is $3\hat{\mathbf{x}}/x^2$, pointing in the $+x$ direction, away from the positive charge. For $x < 0$ it is $-3\hat{\mathbf{x}}/x^2$, pointing in the $-x$ direction, again away from the charge. Had we started from (3.2) and written the field of the first charge as $3q\hat{\mathbf{x}}/(4\pi\varepsilon_0 x^2)$, we would have been right for $x > 0$ and wrong for $x < 0$. The $\boldsymbol{\imath}/\imath^3$ form, with $\imath = |x|$, gives the right direction everywhere without any extra thought.

Reality check: Both charges are on the y axis, so the arrangement is symmetric under $x \to -x$, and the field at $-x$ must be the mirror image of the field at x: the same y component and the opposite x component. It is: the two $\hat{\mathbf{x}}$ terms change sign with x, and the $\hat{\mathbf{y}}$ term does not.

[2] A length is never negative, and x is negative to the left of the origin.

3.3 Field vectors and field lines

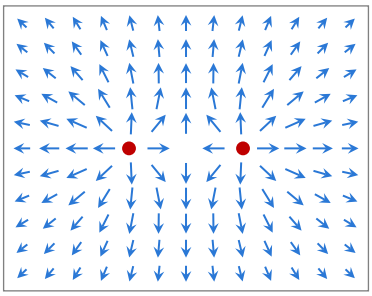


Figure 3.5 The field of two equal positive charges, drawn as arrows on a grid. The lengths are compressed (each is proportional to $|\mathbf{E}|^{0.35}$) so that the distant arrows remain visible; with lengths strictly proportional to $|\mathbf{E}|$, the arrows near the charges would be enormous and the distant ones too short to see.

A number together with a direction is a vector, and is often drawn as an arrow. An electric field assigns such an arrow to every point in space: the arrow at a point shows the direction of the field there, and its length shows the strength. So one way to picture a field is to draw its arrows on a grid of points, as in Figure 3.5. You can only ever draw a sample of the arrows, since there is one at every point, but the collection of arrows *is* the electric field, in a very basic sense.

Field arrows can be computed one point at a time from (3.5), so they are well suited to computer plots. For sketches by hand, the more useful picture is made of **field lines**. If you start from a picture of the arrows, on paper or in your head, a field line is traced as follows.

1. Pick a point to start at.
2. Move a short distance in the direction of the arrow there, drawing as you go.
3. You are now at a slightly different point, with a slightly different arrow. Move a short distance in the direction of the new arrow.
4. Repeat until you reach a charge or are far away from all the charges.

A field line is therefore tangent to the field at every point along it. Figure 3.6 shows field lines traced this way by a computer, for three pairs of charges.

As with the arrows, there are infinitely many field lines, and a sketch shows only enough of them to make the pattern clear. The following facts, which all follow from the definition and from Coulomb's law, are what make a sketch correct.

1. Field lines point away from positive charges and toward negative charges. They begin on positive charges and end on negative charges, or else they go off to infinity (or come in from infinity), or at a point where the field is zero (see fact 4). If you compare the field to a flow of water, positive charges are the faucets and negative charges are the drains.
2. The number of lines leaving or entering a charge is proportional to the size of the charge. About eight lines for each unit of charge is a good choice for a sketch.
3. Where the lines are close together, the field is strong; where they are far apart, it is weak. Strictly, this is true of the lines in three dimensions; in a drawing confined to a plane it holds only qualitatively.
4. Field lines never cross. If two lines crossed, the field at the crossing point would have two different directions, which is impossible. The one exception is a point where the field is zero, which has no direction at all. Lines can meet at such a **null point**, as in Figure 3.6(a) and (b).

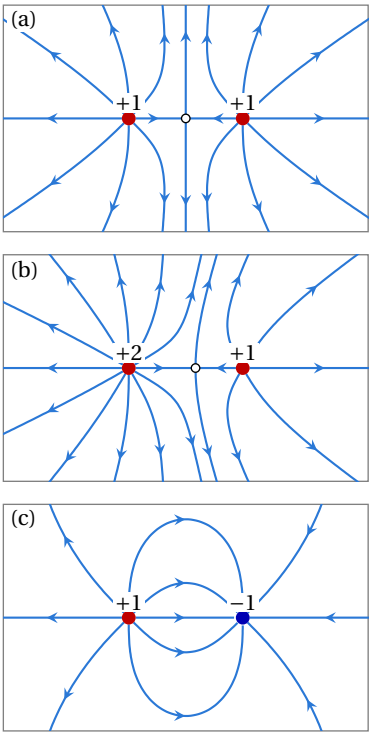


Figure 3.6 Field lines, traced from each positive charge, for (a) two equal positive charges, (b) charges +2 and +1, and (c) charges +1 and −1. The open circles mark points where the field is zero. In (a) and (b) the lines that leave the charges along the axis end at the zero-field point, and two new lines leave it in the perpendicular directions. Animation.

5. Close to any one charge, the lines look like those of that charge alone: straight lines going radially out of it or into it.

6. Far away from all of the charges, the lines look like those of a single point charge equal to the *net* charge, located somewhere among the actual charges. In Figure 3.6(b), the lines far away go radially outward as if from a charge +3. If the net charge is zero, as in (c), the far field is weaker and has a different pattern; Section 8.3.1 describes it.

A good way to sketch field lines is to use the last two facts first. Draw short radial stubs around each charge, with the right number of stubs for each, and decide what the lines should look like far away. Then connect the stubs smoothly, without crossing, so that the far-away pattern comes out right. It also helps to locate any null points. For the +2 and +1 charges of Figure 3.6(b), a distance d apart, the fields cancel on the line between them at the point where $2/x_1^2 = 1/x_2^2$, with x_1 and x_2 the distances to the two charges. That point is a distance $d/(1+\sqrt{2}) \approx 0.41\,d$ from the smaller charge.

3.4 Continuous charge distributions

A charged object contains an enormous number of individual charges, far too many to add up one at a time. Instead, we describe the charge by a density, just as in mechanics we describe the mass of an object by a mass density rather than by listing its atoms. Three kinds of density are used, depending on the shape of the object (Figure 3.7):

- a **volume charge density** ρ (units C/m^3), for charge spread through a volume;
- a **surface charge density** σ (units C/m^2), for charge spread over a surface so thin that its thickness does not matter;
- a **line charge density** λ (units C/m), for charge spread along a line or curve so thin that its width does not matter.

We have already seen the first of these briefly, in Example 1.2. Each is defined locally, as the charge in a small piece divided by the size of the piece: $\rho = dq/d\tau$, $\sigma = dq/da$, and $\lambda = dq/d\ell$. A density can vary from place to place, so it is a function of position, such as $\rho(\mathbf{r})$. The charge in a small piece of the object is then

$$dq = \rho\, d\tau, \qquad dq = \sigma\, da, \qquad dq = \lambda\, d\ell, \tag{3.8}$$

and the total charge is the integral of the density over the object,

$$Q = \int_{\mathcal{V}} \rho\, d\tau, \qquad Q = \int_{\mathcal{S}} \sigma\, da, \qquad Q = \int_{\mathcal{L}} \lambda\, d\ell. \tag{3.9}$$

Strictly speaking, a charge density is an approximation. Charge is carried by discrete particles, so the true charge density is zero almost everywhere and infinite at the location of each particle — a sum of delta functions (Section 1.6.2). A smooth density is what you get by averaging over regions that are large compared to the spacing between atoms but small compared to the object. The approximation is an excellent one for anything much bigger than an atom, and it turns sums over 10^{23} particles into integrals, which is a great simplification. Physics often makes approximations of this kind, and the important thing is to know when they apply.

The delta function also lets point charges be described by a density when that is convenient. A point charge q at $\mathbf{r}'$ has

$$\rho(\mathbf{r}) = q\,\delta^3(\mathbf{r}-\mathbf{r}'), \tag{3.10}$$

and integrating this density over any volume that contains $\mathbf{r}'$ gives q, as it should.

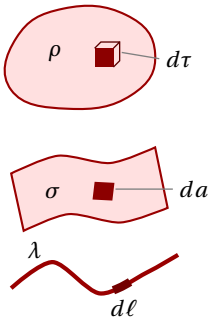


Figure 3.7 Charge spread through a volume, over a surface, and along a line, with densities ρ, σ, and λ. The charge in a small piece (darker) is $dq = \rho\,d\tau$, $\sigma\,da$, or $\lambda\,d\ell$.

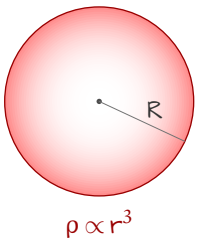
Example 3.3

A sphere of radius R, centered on the origin, has the charge density

$$\rho = \frac{\rho_0}{R^3}\left(x^2+y^2+z^2\right)^{3/2}.$$

Find its total charge.

Solution:



To visualize the charge distribution, we sketch the sphere, shaded darker where the density is larger. Since $x^2+y^2+z^2 = r^2$, the density is $\rho = \rho_0\, r^3/R^3$: it depends only on the distance from the center, and it grows toward the surface. The density is given in Cartesian coordinates, but the object is a sphere, and the limits of integration are much simpler in spherical coordinates. The shape of the object is what should decide the coordinates.

With $d\tau = r^2 \sin\theta\, dr\, d\theta\, d\phi$,

$$\begin{aligned} Q &= \int_0^{2\pi}\int_0^{\pi}\int_0^{R} \frac{\rho_0\, r^3}{R^3}\, r^2 \sin\theta\, dr\, d\theta\, d\phi \\ &= \frac{\rho_0}{R^3}\left(\int_0^R r^5\, dr\right)\left(\int_0^{\pi}\sin\theta\, d\theta\right)\left(\int_0^{2\pi} d\phi\right) \\ &= \frac{\rho_0}{R^3}\cdot\frac{R^6}{6}\cdot 2\cdot 2\pi = \frac{2\pi}{3}\,\rho_0 R^3. \end{aligned}$$

(Why does the r integral contain r^5, when the density goes as r^3? Answer in footnote.[3])

[3] The volume element $d\tau$ supplies the other r^2: a thin shell farther from the center has more volume.

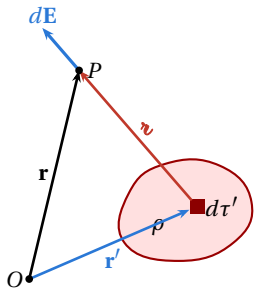


Figure 3.8 A small piece $d\tau'$ of a charge distribution, at the source point $\mathbf{r}'$, and its contribution $d\mathbf{E}$ to the field at the field point P, along $\hat{\boldsymbol{\mathscr{r}}}$. The script-r method adds up the contributions of all the pieces.

Reality check: A uniform sphere with density ρ_0 would have charge $\frac{4}{3}\pi\rho_0 R^3$. Our density is smaller than ρ_0 everywhere inside, so the charge should be smaller, and it is: exactly half as much.

3.5 Solving for E when given ρ: script-r method

We can now find the field of any distribution of charge. Divide the distribution into small pieces. Each piece carries a charge dq and acts like a point charge, so it produces a field given by (3.7) with q replaced by dq (Figure 3.8). By superposition, the total field is the sum of these contributions, and since there are infinitely many infinitesimal pieces the sum becomes an integral. With dq from (3.8), the field of a line, surface, or volume charge is

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\int_{\mathscr{L}} \frac{\lambda(\mathbf{r}')\boldsymbol{\mathscr{r}}}{\mathscr{r}^3}\,d\ell', \tag{3.11}$$

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\int_{\mathscr{S}} \frac{\sigma(\mathbf{r}')\boldsymbol{\mathscr{r}}}{\mathscr{r}^3}\,da', \tag{3.12}$$

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\int_{\mathscr{V}} \frac{\rho(\mathbf{r}')\boldsymbol{\mathscr{r}}}{\mathscr{r}^3}\,d\tau', \tag{3.13}$$

where, as before, $\boldsymbol{\mathscr{r}} = \mathbf{r}-\mathbf{r}'$ and $\mathscr{r} = |\mathbf{r}-\mathbf{r}'|$. Each of these can equally well be written with $\hat{\boldsymbol{\mathscr{r}}}/\mathscr{r}^2$ in place of $\boldsymbol{\mathscr{r}}/\mathscr{r}^3$. Like the point-charge formula, they are all usually called Coulomb's law, since they express the same physics: inverse-square fields, added up by superposition. Together with $\mathbf{F} = Q\mathbf{E}$, they solve every problem in electrostatics, at least in principle. Using them is called the **script-r method**.

It helps to be clear about which quantities vary in these integrals. The field point $\mathbf{r}$ is wherever we want to know the field; it is chosen before the integral begins and stays fixed while it is done. The source point $\mathbf{r}'$ runs over all the pieces of the charge distribution, and the primed coordinates that describe it are the variables of integration. That is why the density is written $\rho(\mathbf{r}')$ and the volume element $d\tau'$: they belong to the source. When the integral is done, the primed variables are gone and the unprimed ones are left behind, so the answer is a function of the field point, $\mathbf{E}(\mathbf{r})$. A density given in a problem as, say, $\lambda = k\cos\phi$ means $\lambda = k\cos\phi'$ once it is inside the integral.

Equations (3.11)–(3.13) are not yet equations you can integrate. They are more like blueprints: they tell you how to construct the integral for a particular problem. The construction always follows the same steps.

1. Draw a picture. Mark an arbitrary piece of the charge distribution, not one at a special location such as an end or the center, and draw $\mathbf{r}$, $\mathbf{r}'$, and $\boldsymbol{\mathscr{r}}$. If you can, use symmetry to find the direction of $\mathbf{E}$ before calculating anything. It is always a good idea to know something about the answer before you start.

2. Write down $\mathbf{r}$, $\mathbf{r}'$, $\boldsymbol{\mathscr{r}} = \mathbf{r} - \mathbf{r}'$, and $\mathscr{r}$, each on a line of its own, and the line, area, or volume element $d\ell'$, da', or $d\tau'$. The element is the size of one piece of the source (a length, an area, or a volume), and $\lambda\, d\ell'$, $\sigma\, da'$, or $\rho\, d\tau'$ is the charge on that piece. An angle is not a length: changing θ' by $d\theta'$ moves the source point a distance $r'\, d\theta'$, not $d\theta'$ (Section 1.3).

3. Write the charge density in terms of the primed variables, using the density, not the total charge. Then set the limits: integrate over the source, counting each piece of it once.

4. Put the pieces into the blueprint, and write any unit vector that changes direction in terms of $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ (see below).

5. Do the integrals, one for each component.

6. Check the result: its direction, its units, and its limits. In particular, far away from any charge distribution with a net charge Q, the field must approach that of a point charge, $Q/4\pi\varepsilon_0 r^2$.

Steps 2 and 3 are where most of the physics is. Once they are done correctly, there is an excellent chance of getting the problem right; the rest is mathematics.

3.5.1 Colton's First Commandment

! **Colton's First Commandment:** Thou shalt not move a variable unit vector outside the integral.

Step 4 deserves a section of its own. The integrals in (3.11)–(3.13) are integrals of vectors, and an integral of a vector is done one component at a time. That is straightforward if the vector is written in terms of $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$, which point in the same direction everywhere: they are constants, and they can be taken outside the integral like any other constant. But the unit vectors of spherical and cylindrical coordinates are *not* constants. The unit vector $\hat{\mathbf{s}}$, for example, points away from the z axis, which is a different direction at every value of ϕ. Its length is constant, but its direction is not:

$$\hat{\mathbf{s}} = \cos\phi\,\hat{\mathbf{x}} + \sin\phi\,\hat{\mathbf{y}}. \tag{3.14}$$

Similarly, $\hat{\boldsymbol{\phi}} = -\sin\phi\,\hat{\mathbf{x}} + \cos\phi\,\hat{\mathbf{y}}$, and

$$\hat{\mathbf{r}} = \sin\theta\cos\phi\,\hat{\mathbf{x}} + \sin\theta\sin\phi\,\hat{\mathbf{y}} + \cos\theta\,\hat{\mathbf{z}}. \tag{3.15}$$

Taking a variable unit vector outside an integral gives wrong answers. For example, $\int_0^{2\pi} \hat{\mathbf{s}}\, d\phi$ is *not* $2\pi\,\hat{\mathbf{s}}$; with (3.14) it is

$$\int_0^{2\pi} \left(\cos\phi\,\hat{\mathbf{x}} + \sin\phi\,\hat{\mathbf{y}}\right) d\phi = \hat{\mathbf{x}} \int_0^{2\pi} \cos\phi\, d\phi + \hat{\mathbf{y}} \int_0^{2\pi} \sin\phi\, d\phi = 0,$$

which makes sense: unit vectors pointing in every direction around a circle add up to nothing. This is the rule called **Colton's First Commandment**:

Thou shalt not move a variable unit vector outside the integral.

Breaking it is the most common mathematical mistake students make in this course. Whenever $\hat{\mathbf{s}}$, $\hat{\boldsymbol{\phi}}$, $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$, or a primed version of one of them appears inside an integral, replace it with its Cartesian form, and only then take $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ outside. You can still integrate over cylindrical or spherical coordinates (and usually should, if that makes the limits simple); it is only the unit vectors that must be Cartesian.

3.5.2 Worked examples

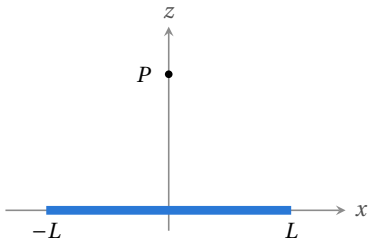
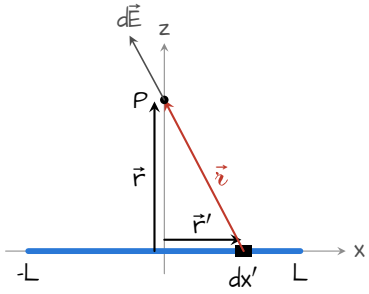


Figure 3.9 Example 3.4: a line of charge from $-L$ to L on the x axis, with the field point P on the z axis.

Example 3.4

A line of charge with constant line charge density λ runs along the x axis from $x = -L$ to $x = L$. Find the electric field at a point a distance z from the center of the line, on the z axis (Figure 3.9).

Solution:



To start out, we identify the starting equation, (3.11):

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0}\int_{\mathcal{L}} \frac{\lambda(\mathbf{r}')\boldsymbol{\imath}}{\imath^3}\, d\ell'.$$

To use the equation we need to draw in an arbitrary chunk of charge (in black), the vectors $\mathbf{r}$ (from the origin to the field point P) and $\mathbf{r}'$ (from the origin to the arbitrary chunk of charge), and $d\ell'$ (the width of the arbitrary chunk of charge). The separation vector $\boldsymbol{\imath} = \mathbf{r} - \mathbf{r}'$ then points from the chunk to P.

The pieces of the blueprint, (3.11), are

$$\mathbf{r} = z\hat{\mathbf{z}}$$
$$\mathbf{r}' = x'\hat{\mathbf{x}}$$
$$\boldsymbol{\imath} = \mathbf{r} - \mathbf{r}' = z\hat{\mathbf{z}} - x'\hat{\mathbf{x}}$$
$$\imath = \sqrt{z^2 + x'^2},$$

and $d\ell' = dx'$. The source point is written with x', not x. (Why x', and not x? Answer in footnote.[4])

Before integrating, the direction. For every piece of the line at x' there is a matching piece at $-x'$; their fields have the same z components and opposite x components. So

[4] It is the source that moves along the line during the integral; x would refer to the field point, which is fixed.

for $\lambda > 0$ and $z > 0$ the field points in the $+z$ direction. The density is the constant λ, so

$$\mathbf{E} = \frac{\lambda}{4\pi\varepsilon_0}\int_{-L}^{L} \frac{z\hat{\mathbf{z}} - x'\hat{\mathbf{x}}}{\left(z^2 + x'^2\right)^{3/2}}\, dx'.$$

The unit vectors here are already Cartesian. In the $\hat{\mathbf{x}}$ term the integrand is an odd function of x', and the integral over the symmetric interval is zero, just as the symmetry argument said. In the $\hat{\mathbf{z}}$ term, z and $\hat{\mathbf{z}}$ belong to the field point and come outside the integral:

$$\mathbf{E} = \frac{\lambda z\hat{\mathbf{z}}}{4\pi\varepsilon_0}\int_{-L}^{L} \frac{dx'}{\left(z^2 + x'^2\right)^{3/2}}.$$

The integral can be done with the substitution $x' = z\tan\alpha$, or looked up in a table, or done by computer:

$$\int_{-L}^{L} \frac{dx'}{\left(z^2 + x'^2\right)^{3/2}} = \frac{x'}{z^2\sqrt{z^2 + x'^2}}\bigg|_{-L}^{L} = \frac{2L}{z^2\sqrt{z^2 + L^2}}.$$

Therefore

$$\mathbf{E} = \frac{1}{4\pi\varepsilon_0}\frac{2\lambda L}{z\sqrt{z^2 + L^2}}\hat{\mathbf{z}}. \tag{3.16}$$

Reality check: For $z < 0$ the factor z in the denominator is negative, and the field points in the $-z$ direction, away from the (positive) line, as it should. Far away, for $z \gg L$, the square root becomes $\sqrt{z^2} = z$ (taking $z > 0$), and

$$\mathbf{E} \approx \frac{1}{4\pi\varepsilon_0}\frac{2\lambda L}{z^2}\hat{\mathbf{z}} = \frac{1}{4\pi\varepsilon_0}\frac{Q}{z^2}\hat{\mathbf{z}},$$

the field of a point charge $Q = 2\lambda L$, which is the total charge of the line.

Close to the line, for $L \gg z$, the square root in (3.16) becomes L, and

$$\mathbf{E} \approx \frac{1}{4\pi\varepsilon_0}\frac{2\lambda}{z}\hat{\mathbf{z}} = \frac{\lambda}{2\pi\varepsilon_0 z}\hat{\mathbf{z}}. \tag{3.17}$$

This is the field of an infinitely long line of charge. Example 4.3 will obtain the same result much more easily with Gauss's law; but Gauss's law cannot give (3.16), because the finite line does not have enough symmetry.

Example 3.5

A disk of radius R lies in the xy plane, centered on the origin. It carries the surface charge density $\sigma = k/s$, where s is the distance from the center and k is a positive constant. Find the electric field at a point on the z axis.

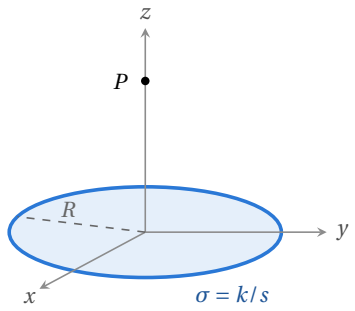


Figure 3.10 Example 3.5: a disk of radius R in the xy plane, with surface charge density $\sigma = k/s$, and the field point P on the z axis.

Solution:

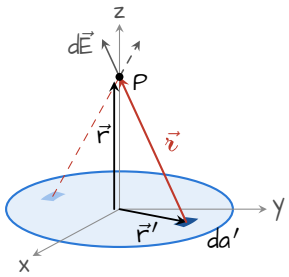


Figure 3.10 shows what we are given. To it we add an arbitrary patch of charge da', the vectors $\mathbf{r}$ (to P), $\mathbf{r}'$ (to the patch) and $\boldsymbol{\mathscr{r}}$ (from the patch to P), and the patch's partner on the opposite side of the center, with the field each one makes at P. The components of these two fields parallel to the disk cancel, so by symmetry the field points in the $+z$ direction (for $z > 0$).

The pieces of the blueprint, (3.12), for the patch da' in the sketch are

$$\begin{aligned}
\mathbf{r} &= z\,\hat{\mathbf{z}} \\
\mathbf{r}' &= s'\,\hat{\mathbf{s}}' = s'\left(\cos\phi'\,\hat{\mathbf{x}} + \sin\phi'\,\hat{\mathbf{y}}\right) \\
\boldsymbol{\mathscr{r}} &= \mathbf{r} - \mathbf{r}' = -s'\cos\phi'\,\hat{\mathbf{x}} - s'\sin\phi'\,\hat{\mathbf{y}} + z\,\hat{\mathbf{z}} \\
\mathscr{r} &= \sqrt{z^2 + s'^2}.
\end{aligned}$$

The unit vector $\hat{\mathbf{s}}'$ varies with ϕ', so by Colton's First Commandment it is written in Cartesian components right away, on the second line. Here $\mathscr{r}$ comes from the Pythagorean theorem, since $\cos^2\phi' + \sin^2\phi' = 1$. The area element is $da' = s'\,ds'\,d\phi'$. (Why $s'\,ds'\,d\phi'$, and not $s\,d\phi\,dz$? Answer in footnote.[5]) In primed variables the density is $\sigma = k/s'$. Putting the pieces together,

$$\mathbf{E} = \frac{1}{4\pi\varepsilon_0}\int_0^{2\pi}\int_0^R \frac{k}{s'}\,\frac{-s'\cos\phi'\,\hat{\mathbf{x}} - s'\sin\phi'\,\hat{\mathbf{y}} + z\,\hat{\mathbf{z}}}{\left(z^2 + s'^2\right)^{3/2}}\,s'\,ds'\,d\phi'.$$

The ϕ' integrals of the $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ terms are $\int_0^{2\pi}\cos\phi'\,d\phi' = 0$ and $\int_0^{2\pi}\sin\phi'\,d\phi' = 0$, as the symmetry argument said they would be. In the $\hat{\mathbf{z}}$ term nothing depends on ϕ', so the ϕ' integral gives 2π, and the two factors of s' cancel:

$$\mathbf{E} = \frac{2\pi k z\,\hat{\mathbf{z}}}{4\pi\varepsilon_0}\int_0^R \frac{ds'}{\left(z^2 + s'^2\right)^{3/2}}.$$

This is the same integral as in Example 3.4, with different limits:

$$\int_0^R \frac{ds'}{\left(z^2 + s'^2\right)^{3/2}} = \frac{R}{z^2\sqrt{z^2 + R^2}}.$$

Therefore

$$\mathbf{E} = \frac{kR}{2\varepsilon_0\, z\sqrt{z^2 + R^2}}\,\hat{\mathbf{z}}. \tag{3.18}$$

Reality check: Far from the disk, $z \gg R$, (3.18) becomes $kR/(2\varepsilon_0 z^2)$. The total charge on the disk is $Q = \int_0^{2\pi}\int_0^R (k/s')\,s'\,ds'\,d\phi' = 2\pi kR$, so this is $Q/4\pi\varepsilon_0 z^2$, the field of a point charge, as expected.

[5] The disk is flat, like the end cap of a cylinder, and a patch of it has sides ds' and $s'\,d\phi'$. The element $s\,d\phi\,dz$ belongs to the curved side of a cylinder.

Close to the disk, $z \to 0$, the field grows without limit, like $1/z$ (Figure 3.11). That might seem surprising, since a uniformly charged disk has a finite field close to its center. But this density is not uniform: $\sigma = k/s$ itself becomes infinite at the center, and the field near the center reflects that. The total charge is still finite, so the infinite density is not a problem in itself.

The next tutorial works through a third example in the same way, one step at a time. It is the kind of problem that is worth doing yourself before reading further.

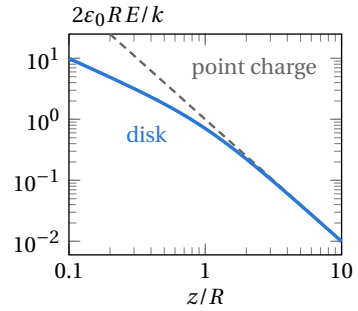


Figure 3.11 The field on the axis of the disk of Example 3.5, (3.18) (blue), and the field of a point charge with the same total charge, $Q = 2\pi kR$ (dashed), in units of $k/2\varepsilon_0 R$. Far away they agree; close to the disk the field grows as $1/z$, not $1/z^2$.

Tutorial 3.1 Coulomb's law for the electric field

Printable version

Background

Coulomb's law states that the electric field produced by a point charge q, located at the point $\mathbf{r}' = (x', y', z')$, at an observation point $\mathbf{r} = (x, y, z)$, is

$$\mathbf{E}(\mathbf{r}) = \frac{q}{4\pi\varepsilon_0}\frac{\boldsymbol{\mathscr{r}}}{\mathscr{r}^3},$$

where $\boldsymbol{\mathscr{r}} = \mathbf{r} - \mathbf{r}'$ is the vector that points from the charge to the observation point and $\mathscr{r}$ is its magnitude. This is an inverse square law because $\boldsymbol{\mathscr{r}}/\mathscr{r}$ is a unit vector in the $\boldsymbol{\mathscr{r}}$ direction; consequently the field falls off as $1/\mathscr{r}^2$.

If several point charges produce the field, then by superposition

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\sum_i \frac{q_i \boldsymbol{\mathscr{r}}_i}{\mathscr{r}_i^3},$$

where $\boldsymbol{\mathscr{r}}_i = \mathbf{r} - \mathbf{r}'_i$ points from the ith charge to the field point. For a line, surface or volume charge density the sum becomes an integral:

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\int \frac{\lambda(\mathbf{r}')\boldsymbol{\mathscr{r}}}{\mathscr{r}^3}\,d\ell',$$

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\int_{\mathscr{S}} \frac{\sigma(\mathbf{r}')\boldsymbol{\mathscr{r}}}{\mathscr{r}^3}\,da',$$

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\int_{\mathscr{V}} \frac{\rho(\mathbf{r}')\boldsymbol{\mathscr{r}}}{\mathscr{r}^3}\,d\tau'.$$

Situation 1

A point charge q is located at $(6, -2, 0)$ m. An observer is located at $(-2, 5, 0)$ m.

(a) On graph paper, draw x and y axes and mark the origin, the charge and the observer. Draw and label the vectors $\mathbf{r}$, $\mathbf{r}'$ and $\boldsymbol{\mathscr{r}}$.

(b) Write out $\mathbf{r}$, $\mathbf{r}'$ and $\boldsymbol{\mathscr{r}}$ in terms of the unit vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$ and $\hat{\mathbf{z}}$.

(c) What is the electric field at $\mathbf{r}$?

Situation 2

Two point charges, each with charge q, are located on the x axis at the points shown in Figure 3.12. The observer is at an arbitrary location (x, y), marked as point P. The goal is to find the field at P.

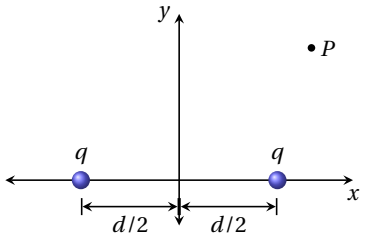


Figure 3.12 Tutorial 3.1, Situation 2: two point charges q on the x axis and an observation point P.

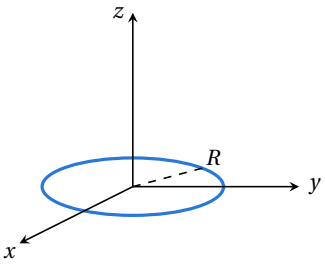


Figure 3.13 Tutorial 3.1, Situation 3: a ring of radius R in the xy plane.

(a) Draw the vectors $\mathbf{r}$, $\mathbf{r}'_1$, $\boldsymbol{\mathscr{r}}_1$, $\mathbf{r}'_2$ and $\boldsymbol{\mathscr{r}}_2$ on a copy of the figure.

(b) Write out the five vectors in Cartesian components using unit-vector notation, as well as the magnitudes $\mathscr{r}_1$ and $\mathscr{r}_2$.

(c) Write down a general equation for the electric field produced by two point charges. Fill in the equation with the values you have just written out to get the answer.

Situation 3

A ring of charge of radius R is centered on the origin and lies in the xy plane, as in Figure 3.13. Its linear charge density is given in cylindrical coordinates by $\lambda = \lambda_0 \cos\phi$. The goal is to find the field at an arbitrary point on the z axis.

(a) Visualize the situation: draw positive and negative charges on the ring to show how and where the charge is located for the given λ. Use your picture to determine by symmetry the direction of the field on the z axis, and explain how you know.

(b) Describe what you have to work with in terms of givens, knowns, unknowns, target variables, and so on.

(c) What will be your starting equation? Briefly describe the underlying principles and concepts at work in that equation.

(d) Mark an arbitrary chunk of charge on the ring, then draw the vectors $\mathbf{r}$, $\mathbf{r}'$ and $\boldsymbol{\mathscr{r}}$ for that chunk producing a field at an arbitrary point on the z axis. Which of these change during the integral, and how?

(e) Using cylindrical coordinates and cylindrical unit vectors, write down expressions for $\mathbf{r}$, $\mathbf{r}'$, $\boldsymbol{\mathscr{r}}$, $\mathscr{r}$ and $d\ell'$.

(f) Fill in your starting equation with those expressions and the given charge density, then do the integral. Along the way, keep Colton's First Commandment: thou shalt not move a variable unit vector outside the integral. Instead, turn any variable unit vectors into constant ones ($\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, $\hat{\mathbf{z}}$); only then can they come out of the integral. You should find three separate integrals, one for each component, two of which are zero by symmetry.

(g) How can you check whether your answer is reasonable? Consider limiting cases, units, comparison to other problems, and so forth.

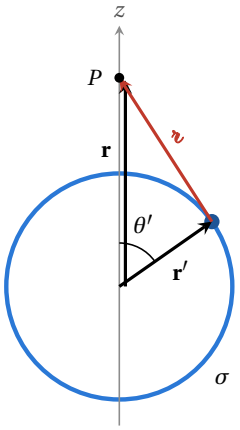


Figure 3.14 A spherical shell of radius R with uniform surface charge density σ, a small patch of it at polar angle θ', a field point P on the z axis, and the vectors $\mathbf{r}$, $\mathbf{r}'$, and $\boldsymbol{\mathscr{r}}$. Their lengths, z, R, and $\mathscr{r}$, are the three sides of a triangle.

3.5.3 The uniformly charged spherical shell

As a last example, consider a thin spherical shell of radius R with a uniform surface charge density σ, and find the field at a point on the z axis, a distance z from the center (Figure 3.14). By symmetry, the field points along $\hat{\mathbf{z}}$. The pieces of the

blueprint are

$$\begin{aligned}
\mathbf{r} &= z\,\hat{\mathbf{z}} \\
\mathbf{r}' &= R\,\hat{\mathbf{r}}' = R\left(\sin\theta'\cos\phi'\,\hat{\mathbf{x}} + \sin\theta'\sin\phi'\,\hat{\mathbf{y}} + \cos\theta'\,\hat{\mathbf{z}}\right) \\
\boldsymbol{\mathscr{r}} &= \mathbf{r} - \mathbf{r}' = -R\sin\theta'\cos\phi'\,\hat{\mathbf{x}} - R\sin\theta'\sin\phi'\,\hat{\mathbf{y}} + \left(z - R\cos\theta'\right)\hat{\mathbf{z}} \\
\mathscr{r} &= \sqrt{R^2\sin^2\theta' + (z - R\cos\theta')^2} = \sqrt{z^2 + R^2 - 2zR\cos\theta'},
\end{aligned} \tag{3.19}$$

and $da' = R^2\sin\theta'\,d\theta'\,d\phi'$. The source point is written $R\,\hat{\mathbf{r}}'$ rather than $r'\,\hat{\mathbf{r}}'$, because every piece of the shell is the same distance R from the center, and $\hat{\mathbf{r}}'$, a variable unit vector, is written in Cartesian components on the same line, (3.15), as Colton's First Commandment requires. It would be tempting to write $\mathscr{r} = \sqrt{z^2 + R^2}$, but that is wrong: $\mathbf{r}$ and $\mathbf{r}'$ are not perpendicular, so the Pythagorean theorem does not apply to them. The last form of $\mathscr{r}$ is the law of cosines for the triangle in Figure 3.14. The ϕ' integrals of the $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ components of $\boldsymbol{\mathscr{r}}$ vanish, the ϕ' integral of everything else gives 2π, and what remains is

$$\mathbf{E} = \frac{2\pi\sigma R^2}{4\pi\varepsilon_0}\,\hat{\mathbf{z}}\int_0^\pi \frac{\left(z - R\cos\theta'\right)\sin\theta'}{\left(z^2 + R^2 - 2zR\cos\theta'\right)^{3/2}}\,d\theta'. \tag{3.20}$$

This integral can be done by hand with a substitution, although it takes some care; Exercise P3.5 walks through it. The result is

$$\mathbf{E} = \begin{cases} 0 & \text{inside the shell,} \\ \dfrac{1}{4\pi\varepsilon_0}\dfrac{Q}{z^2}\,\hat{\mathbf{z}} & \text{outside the shell,} \end{cases} \tag{3.21}$$

where $Q = 4\pi R^2\sigma$ is the total charge of the shell. Since the z axis can be pointed in any direction, the field outside at any point $\mathbf{r}$ is $Q\,\hat{\mathbf{r}}/4\pi\varepsilon_0 r^2$: exactly the field of a point charge Q at the center.

Both results deserve some thought. Outside, the shell is indistinguishable from a point charge, not just far away (as must be true of any charge distribution with a nonzero net charge) but at every distance. Inside, the field is zero everywhere, not just at the center. To see why that is plausible, consider a point inside the shell near one side (Figure 3.15). The nearby part of the shell is close, so its field is strong, but it is a small part of the shell. The far side is farther away, so its field is weaker, but there is more of it. For an inverse-square force the two effects cancel exactly. These results will appear again in Example 4.2, where Gauss's law gives them in a few lines. They also give the field of a uniformly charged solid sphere, by superposition: think of the sphere as a set of nested shells. At a point outside, every shell acts like a point charge at the center, so the whole sphere does too.

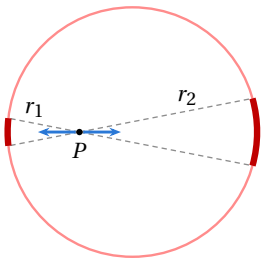


Figure 3.15 A point P inside a charged spherical shell, and a thin double cone through it. The far patch of the shell is larger than the near one, in the ratio r_2^2/r_1^2, and farther away, so the two patches give fields (blue) of equal size and opposite direction at P.

Tips for Mathematica and Python. Integrals like (3.20) are good candidates for symbolic software, but they come with a lesson about assumptions.

- In Mathematica, `Integrate` applied to (3.20) with no further information either takes a long time or returns a `ConditionalExpression` containing terms like $(z-R)/\sqrt{(R-z)^2}$. Telling it only that $z>0$ and $R>0$ gives $(1-\text{Sign}[R-z])/z^2$. The difficulty is that $\sqrt{(R-z)^2}$ is $|R-z|$, and the program cannot simplify that without knowing which of R and z is bigger. Wrapping the integral in `Assuming[z > R && R > 0, ...]` gives $2/z^2$, and with $0<z<R$ it gives 0, which are the two cases of (3.21).

- In Python, `sympy.integrate` has the same difficulty. Declaring the symbols with `positive=True` is necessary but not enough: the result still contains $\sqrt{R^2-2Rz+z^2}$ and $\sqrt{R^2+2Rz+z^2}$, which sympy will not reduce to $|z-R|$ and $z+R$. One approach is to substitute `Abs(z - R)` and `z + R` for the two square roots by hand, and then call `refine` with `Q.gt(z, R)` or `Q.lt(z, R)` (followed by `simplify`), which gives $2/z^2$ and 0.

- With either program, a quick numerical check (`NIntegrate` in Mathematica, `scipy.integrate.quad` in Python) at one value of z inside and one outside is a reliable check that the symbolic result has been simplified correctly.

3.6 Numerical method: direct integration

The examples so far were chosen so that the integrals can be done by hand, and in each one the field point was on an axis of symmetry. Move the field point off the axis, or give the charge distribution a less convenient shape, and the integrals usually have no closed-form answer. The integrals can still be done numerically, and the most direct way to do so uses the same physics that led to (3.11)–(3.13) in the first place.

The idea is to go back one step, from the integral to the sum. Instead of infinitesimal pieces, divide the charge distribution into a finite number N of small chunks, treat each chunk as a point charge, and add up their fields with (3.5):

1. Divide the source into N small chunks. For chunk i, find the location of its center, $\mathbf{r}'_i$, and its charge, Δq_i, which is the density at the center times the chunk's length, area, or volume: $\lambda\,\Delta\ell$, $\sigma\,\Delta a$, or $\rho\,\Delta\tau$.

2. For the field point $\mathbf{r}$, find $\boldsymbol{\imath}_i = \mathbf{r} - \mathbf{r}'_i$ and $\imath_i$ for each chunk.

3. Add up the contributions:

$$\mathbf{E}(\mathbf{r}) \approx \frac{1}{4\pi\varepsilon_0}\sum_{i=1}^{N}\frac{\Delta q_i\,\boldsymbol{\imath}_i}{\imath_i^3}. \tag{3.22}$$

4. Increase N (for example, by doubling the number of divisions in each direction) and repeat, until the answer stops changing at the accuracy you need.

The last step is not optional. A single calculation with one value of N gives a number, but no indication of whether the number is right; only the convergence as N increases tells you that. In the language of numerical analysis, this procedure is the midpoint rule applied to each component of the integral. Because the calculation is done with numbers in Cartesian components from the start, Colton's First Commandment takes care of itself.

One caution: the method assumes that every chunk is small compared to its distance from the field point. If the field point is on the charge distribution or very close to it, the nearby chunks need to be much smaller than the distance, and a field point exactly at a chunk's center gives a division by zero.

As an example, consider a hollow cylindrical shell — only the curved side, which I like to call the *wrapper*, without end caps — of radius $R = 0.1\,\text{m}$ and length $L = 0.2\,\text{m}$, centered on the origin with its axis along z, carrying a uniform surface charge density $\sigma = 1.0 \times 10^{-7}\,\text{C/m}^2$. We want the field at the point $(0, 0.4, 0.4)\,\text{m}$, off the axis, where there is no simple closed form. Divide the wrapper into N_ϕ strips in the ϕ direction and N_z strips in the z direction, as in Figure 3.16. The chunk centered at angle ϕ_i and height z_j is at $\mathbf{r}'_{ij} = R\cos\phi_i\,\hat{\mathbf{x}} + R\sin\phi_i\,\hat{\mathbf{y}} + z_j\,\hat{\mathbf{z}}$ and has charge $\Delta q = \sigma R\,\Delta\phi\,\Delta z$, with $\Delta\phi = 2\pi/N_\phi$ and $\Delta z = L/N_z$. Table 3.1 shows the results of (3.22) for increasing numbers of chunks, $N = N_\phi N_z$.

With a single chunk, the whole cylinder is treated as one point charge at the center of that chunk, which happens to be at $(-R, 0, 0)$, and the result even has a spurious x component. By 16×16 chunks, the answer has converged to four significant figures: $\mathbf{E} = (0,\ 246.8,\ 250.8)\,\text{N/C}$, with magnitude $351.9\,\text{N/C}$. From 8×8 on, the errors shrink by about a factor of four each time N is multiplied by four (Figure 3.17). (They shrink faster at first, because in the ϕ direction the cylinder closes on itself, and the midpoint rule converges very quickly for such a periodic integrand.) A useful sanity check is the field of a point charge: the cylinder's total charge is $Q = 2\pi RL\sigma = 1.26 \times 10^{-8}\,\text{C}$, and a point charge Q at the origin would produce $(0,\ 249.6,\ 249.6)\,\text{N/C}$ at this point, within about one percent of the converged answer. That is about as close as one should expect, since the field point is less than three times as far from the center as the cylinder is across.

Most scientific software also has built-in numerical integrators: for example, `NIntegrate` in Mathematica, and `dblquad` and `tplquad` in Python's `scipy.integrate` module. They do the same job as the chunk-and-sum method, but adaptively: they put more points where the integrand changes quickly and fewer where it does not, and they estimate their own error. Applied to the cylinder, either one reproduces the converged values above. A few tips:

- Integrate each Cartesian component as a separate integral. Writing the integrand in Cartesian components from the start also keeps Colton's First Commandment.
- In `dblquad`, the function to be integrated takes its arguments in the order (inner variable, outer variable), which is the reverse of what many people expect, and the limits are given outer variable first. A test on an integral whose answer you know catches this kind of mistake quickly.

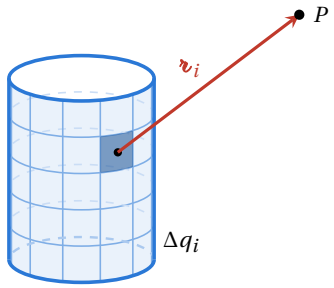


Figure 3.16 The curved surface of a finite cylinder divided into chunks, for the direct-integration method. Each chunk is treated as a point charge Δq_i at its center, and the fields of all the chunks are added at the field point P.

$N_\phi \times N_z$	E_x	E_y	E_z
1×1	59.6	238.3	238.3
2×2	0.0	248.5	266.4
4×4	0.0	246.3	250.6
8×8	0.0	246.7	250.8
16×16	0.0	246.8	250.8
32×32	0.0	246.8	250.8

Table 3.1 The field of the charged wrapper at $(0, 0.4, 0.4)\,\text{m}$, in N/C, from (3.22) with $N_\phi \times N_z$ chunks.

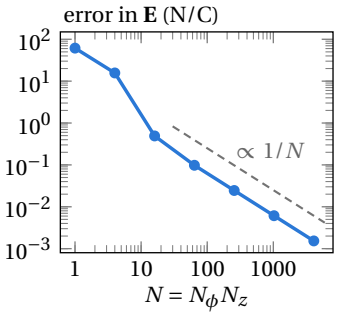


Figure 3.17 The error of the chunk-and-sum field of the cylinder, $|\mathbf{E}(N) - \mathbf{E}|$, against the number of chunks N, with $N_\phi = N_z$; $\mathbf{E}$ is the converged value. The dashed line has slope -1: from 8×8 on, the error is proportional to $1/N$.

- In Python, the chunk-and-sum method itself is easy to write without explicit loops, using `numpy` arrays for the chunk positions and charges; in Mathematica, `Table` and `Total` do the same.
- If the field point lies on the charge distribution itself, the integrand is singular there, and the field may not even be finite (on a line of charge, it is not). Mathematica's `NIntegrate` can be told about singular points with its `Exclusions` option.

The same chunk-and-sum method will be used again for the electric potential in Section 5.5 and for magnetic fields in Section 10.6.

3.7 Summary

The electric field of a set of stationary charges follows from two experimental facts, Coulomb's law and superposition. Everything else in this chapter is a way of applying them: to a few point charges, to continuous distributions of charge, and, when the integrals cannot be done by hand, numerically.

Charge, force, and field. The vector $\boldsymbol{\mathscr{r}} = \mathbf{r} - \mathbf{r}'$ points from the source point $\mathbf{r}'$ to the field point $\mathbf{r}$.

$Q = \int \lambda\, d\ell, \quad \int \sigma\, da, \quad \int \rho\, d\tau$	total charge from a line, surface, or volume charge density
$\mathbf{F} = \dfrac{1}{4\pi\varepsilon_0}\dfrac{qQ\boldsymbol{\mathscr{r}}}{\mathscr{r}^3}$	Coulomb's law: the force of q on Q
$\mathbf{F} = Q\mathbf{E}$	the force on a charge Q placed in a field $\mathbf{E}$
$\mathbf{E}(\mathbf{r}) = \dfrac{1}{4\pi\varepsilon_0}\sum_i \dfrac{q_i \boldsymbol{\mathscr{r}}_i}{\mathscr{r}_i^3}$	the field of point charges, by superposition

The script-r method. The field of a continuous charge distribution,

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\int \frac{\lambda(\mathbf{r}')\boldsymbol{\mathscr{r}}}{\mathscr{r}^3}\, d\ell', \qquad \frac{1}{4\pi\varepsilon_0}\int \frac{\sigma(\mathbf{r}')\boldsymbol{\mathscr{r}}}{\mathscr{r}^3}\, da', \qquad \frac{1}{4\pi\varepsilon_0}\int \frac{\rho(\mathbf{r}')\boldsymbol{\mathscr{r}}}{\mathscr{r}^3}\, d\tau'.$$

Integrate over the primed (source) variables; the answer is a function of the field point. Write down $\mathbf{r}$, $\mathbf{r}'$, $\boldsymbol{\mathscr{r}}$, $\mathscr{r}$, and the element $d\ell'$, da', or $d\tau'$ separately before assembling the integral, and use symmetry to find the direction of the field first when you can. *Colton's First Commandment:* thou shalt not move a variable unit vector outside the integral — write $\hat{\mathbf{s}}$, $\hat{\mathbf{r}}$, and the rest in terms of $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ first.

Results worth remembering. A line of charge from $-L$ to L, on its perpendicular bisector, has $E = 2\lambda L/\left(4\pi\varepsilon_0 z\sqrt{z^2+L^2}\right)$, which becomes $\lambda/2\pi\varepsilon_0 z$ for an infinite line. A uniformly charged spherical shell has zero field inside and the field of a point charge at its center outside.

Field lines begin on positive charges and end on negative charges (or at infinity, or at a point where $\mathbf{E} = 0$), are closer together where the field is stronger, and never cross, except that they can meet at points where $\mathbf{E} = 0$. Near a charge they look like that charge's field alone; far away they look like the field of the net charge.

Direct integration. Divide the source into N chunks, treat each as a point charge, add their fields, and increase N until the answer stops changing.

Exercises

Exercises for 3.1 Charge and Coulomb's law

Warm-up questions Read Section 3.1, then answer before class

W3.1 Moving charges feel magnetic forces, as this section reminds us, and yet the second simplification allows the test charge to move; only the source charges must stay still. Why does the motion of the test charge not change the force on it?

W3.2 A charge q is at $\mathbf{r}' = a\hat{\mathbf{x}}$, and a test charge Q is at $\mathbf{r} = 2a\hat{\mathbf{y}}$. Find $\boldsymbol{\mathscr{r}}$, $\mathscr{r}$, and $\hat{\boldsymbol{\mathscr{r}}}$. Then work out $\hat{\mathbf{r}} - \hat{\mathbf{r}}'$, the difference of the two unit vectors. Is it a unit vector? Does it point in the same direction as $\hat{\boldsymbol{\mathscr{r}}}$?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

Exercises for 3.2 The electric field and superposition

Warm-up questions Read Section 3.2, then answer before class

W3.3 In your own words, why is the electric field defined with a test charge so small that it does not disturb the source charges? Describe a situation in which a large test charge Q would feel a force $\mathbf{F}$ for which $\mathbf{F}/Q$ is not the field that was there before Q arrived.

W3.4 Suppose the right-hand charge of Example 3.1 were $-1\,\mathrm{C}$ instead of $+1\,\mathrm{C}$. Without redoing the calculation, sketch $\mathbf{E}_1$ and $\mathbf{E}_2$ at the field point and decide which way the total field points. Is its magnitude larger than before, smaller, or the same?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

Exercises for 3.3 Field vectors and field lines

Warm-up questions Read Section 3.3, then answer before class

W3.5 The third fact about field lines (close lines, strong field) is strictly true only in three dimensions. To see why, let N lines leave a point charge evenly in all directions. How does the number of lines per unit area crossing a sphere of radius r depend on r? In a drawing in a plane, how does the number of lines per unit length crossing a circle of radius r depend on r? Which falls off as E does?

W3.6 True or false? Give a reason for each. (a) Wherever a sketch shows no field line, the field is zero. (b) Far from a charge $+2q$ and a charge $-q$, the field lines look like those of a single charge $+q$. (c) A positive test charge released from rest at a point on a curved field line moves along that line.

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P3.1 Sketch the electric field lines in the xy plane for this three-charge configuration (Figure 3.18): $+q$ at $(1,0)$, $-q$ at $(0,1)$, and $-q$ at $(0,-1)$. Use about eight lines per charge. Before connecting the lines, decide what they look like close to each charge and far away from all three, and look for points where the field is zero.

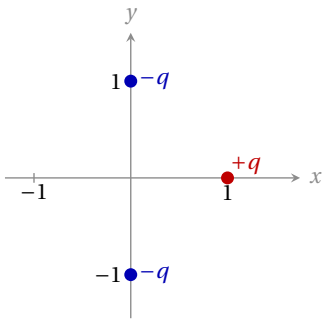


Figure 3.18 The three charges of Exercise P3.1.

P3.2 True or false? Explain your reasoning.

(a) Field lines can never cross. (Consider points where $\mathbf{E} = 0$.)

(b) A charge $3q$ at $(1,0,0)$ and a charge $-q$ at $(-1,0,0)$ produce an electric field that is zero at some point on the x axis (not counting infinity). If this is true, find the point.

Exercises for 3.4 Continuous charge distributions

Warm-up questions Read Section 3.4, then answer before class

W3.7 Which density, ρ, σ, or λ, would you use to describe the charge on each of these, and what does your choice assume? (a) A long, thin charged wire. (b) A charged metal ball, whose charge sits in a very thin layer at its surface. (c) A thundercloud, whose charge is carried by its many water droplets and ice crystals.

W3.8 Charge is quantized (Section 3.1), and yet the charge found by integrating a smooth density, such as the $(2\pi/3)\rho_0 R^3$ of Example 3.3, can have any value, not just a whole number times e. Why is there no contradiction? About how many elementary charges make up a charge of 1 nC?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P3.3 Determine the amount of charge in each of the following objects, by integrating the given charge densities. In each case, also give the units of the constant k.

(a) A sphere of charge of radius R, with $\rho = k/r$. (Is the infinite density at the center a problem?)

(b) A spherical shell of charge of radius R, with $\sigma = k\cos^2\theta$.

(c) A line of charge extending from $x = -a$ to $x = +a$ on the x axis, with $\lambda = kx^2$.

Exercises for 3.5 Solving for E when given ρ: script-r method

Warm-up questions Read Section 3.5, then answer before class

W3.9 Evaluate $\int_0^\pi \hat{\mathbf{s}}\, d\phi$, which adds up the unit vectors $\hat{\mathbf{s}}$ of the points of a half circle around the z axis. Why is the answer not $\pi\hat{\mathbf{s}}$? Sketch a few of the unit vectors, and use the sketch to check the direction and the size of your result.

W3.10 Setting up Example 3.4, a classmate draws the piece of charge at the center of the line, $x' = 0$, and so writes $\mathbf{r}' = 0$, $\boldsymbol{\mathscr{r}} = z\hat{\mathbf{z}}$, and $\mathscr{r} = z$. What field does that lead to, and what went wrong? For which field points is the result nevertheless close to right?

W3.11 The double cone of Figure 3.15 explains why the field is zero everywhere inside a uniformly charged spherical shell. Try the same argument at a point inside a uniformly charged ring, in the plane of the ring, with a thin double wedge through the point in place of the double cone. How do the lengths of the two pieces of ring that the wedge cuts out depend on their distances from the point? Is the field zero there? If not, which way does it point? (Compare how a piece's length and its distance enter its field.)

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

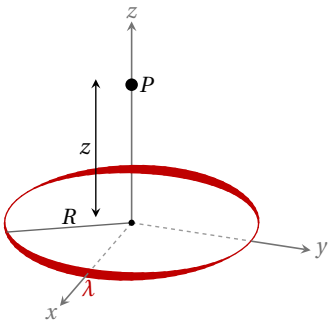


Figure 3.19 The ring of Exercise P3.4 and the field point P. The thickness of the ring shows its charge density.

P3.4 A ring of charge of radius R, centered on the origin in the xy plane, carries the line charge density $\lambda = k\cos^2\phi$ (Figure 3.19).

(a) Find the electric field at a point a distance z above the center of the ring. Write down $\mathbf{r}$, $\mathbf{r}'$, $\boldsymbol{\mathscr{r}}$, and $\mathscr{r}$ first. Two cautions: $\mathbf{r}'$ involves a cylindrical unit vector, which you should write in terms of $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$; and you need an expression for $d\ell'$, the length of a small piece of the ring at an arbitrary location.

(b) Show that your answer is the same as the field of a *uniformly* charged ring with the same total charge, and explain why that is true on the z axis, even though the charge on this ring is not uniform.

P3.5 In this problem you derive the field of a uniformly charged spherical shell, (3.21).

(a) Starting from (3.12) and the quantities given in Section 3.5.3, obtain (3.20). Show explicitly what happens to the $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ components.

(b) Do the θ' integral by hand. The substitution $w = z^2 + R^2 - 2zR\cos\theta'$, which is u^2, works well. Take $z > 0$, and remember that $\sqrt{(z-R)^2} = |z-R|$. Show that the result is (3.21), both for $z < R$ and for $z > R$.

(c) Find the field at $z = R$ exactly, on the surface of the shell. (Go back to the integral rather than taking a limit of your answer to (b).) Compare it with the fields just inside and just outside the shell.

(d) Do the integral with Mathematica, Python, or other software, and show how you handled the difference between $z < R$ and $z > R$.

Exercises for 3.6 Numerical method: direct integration

Warm-up questions Read Section 3.6, then answer before class

W3.12 With a single chunk, the field of the charged wrapper in Table 3.1 has $E_x = 59.6\,\text{N/C}$, but the converged field has $E_x = 0$. Where did the x component come from? What symmetry of the wrapper and the field point makes E_x zero, and why do all the grids with 2, 4, 8, ... chunks around the cylinder give $E_x = 0.0$?

W3.13 A classmate's program gives the field of the wrapper at $(0, 0.4, 0.4)\,\text{m}$, converged to four figures, with 64×64 chunks. They then use the same chunks to find the field at a point 1 mm outside the wrapper's surface. Should they trust that result? Estimate the size of a chunk, and explain.

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P3.6 A hollow cylindrical shell of radius R and length L, centered on the origin with its axis along z, carries a uniform surface charge density σ (Figure 3.20). The shell includes only the curved wrapper, not the end caps.

(a) Write down the integrals you would use to find the x, y, and z components of the electric field at an arbitrary point (x, y, z). The integrals should contain only fundamental constants, the given quantities, the constant unit vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$, and the variables of integration.

(b) For a point on the z axis, do the integrals by hand.

(c) Take $\sigma = 1.0 \times 10^{-7}\,\text{C/m}^2$, $R = 0.1\,\text{m}$, and $L = 0.2\,\text{m}$, and evaluate your result from (b) at $(0, 0, 0.4)\,\text{m}$. Then evaluate the integrals of (a) at the same point numerically, either with the chunk-and-sum method of Section 3.6 or with a built-in numerical integrator, and compare.

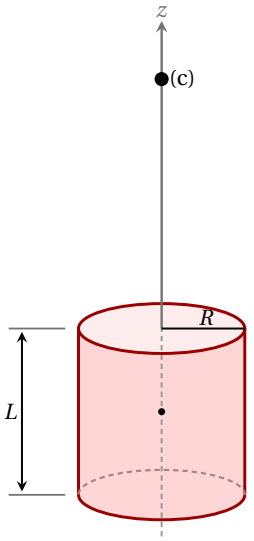


Figure 3.20 The cylindrical shell of Exercise P3.6, drawn with $L = 2R$, and the point of part (c).

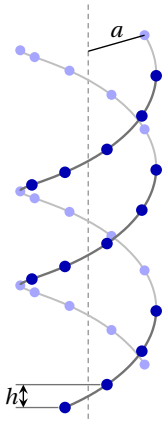


Figure 3.21 Fifteen base pairs of the model DNA molecule of Exercise P3.7, drawn to scale with its axis (dashed) vertical. Each strand's phosphates (blue) lie on a gray helix; the parts behind the axis are lighter.

P3.7 The DNA double helix is a charged object (Figure 3.21). Each nucleotide contains a phosphate group, and in water each phosphate carries a charge of $-e$; the phosphates lie along the two sugar–phosphate backbones, which wind around each other in a double helix. In this problem you find the electric field near a DNA molecule, using a simple model of its common B form.[6] Model each phosphate as a point charge $-e$. The phosphates lie on a cylinder of radius $a = 0.89\,\text{nm}$ about the axis of the molecule, one on each strand for every base pair. Successive base pairs are separated by $h = 0.34\,\text{nm}$ along the axis and rotated by $+36°$ about it (a right-handed helix), so that it makes one turn every 10 base pairs. At each base pair, the phosphate of the second strand is at the same height as that of the first and is rotated $+144°$ from it about the axis; the unequal spacing of the two strands creates the major and minor grooves of the molecule.

(a) Write a program, in any language, that builds a list of the charge positions for a molecule of N base pairs centered on the origin, with its axis along z, and that calculates $\mathbf{E}$ at any point by adding the fields of the point charges.

(b) Take $N = 200$. At a distance $s = 1.2\,\text{nm}$ from the axis, and at a height midway between two base pairs, calculate the radial component E_s of the field at 72 equally spaced values of the angle ϕ around the molecule, and plot E_s against ϕ. By what factor do the largest and smallest magnitudes differ? Why does the field depend on ϕ?

(c) Repeat (b) at $s = 2$, 3, and 5 nm. Compare the average of E_s over ϕ with the field of an infinite, uniform line charge with the same charge per unit length, $\lambda = -2e/h$, given by (3.17). Beyond roughly what distance does the helical structure stop mattering, in the sense that E_s varies by less than 1% with ϕ?

(d) At $s = 5\,\text{nm}$, how do your results change if you use $N = 100$ or $N = 400$ instead of 200? What determines how many base pairs you need?

(e) Compare the size of the field at $s = 1.2\,\text{nm}$ with the field at which air breaks down and sparks, about $3 \times 10^6\,\text{V/m}$. (Inside a cell, DNA is surrounded by salt water, and positive ions from the solution gather close to the molecule and cancel much of its field beyond about a nanometer. What you have calculated is the "bare" field, the field the molecule would produce without them.)

[6] The geometry is close to that of a standard coarse-grained model of B-DNA described in "Adsorption of monovalent and multivalent cations and anions on DNA molecules," E. Allahyarov, H. Löwen, and G. Gompper, Phys. Rev. E 68, 061903 (2003). https://doi.org/10.1103/PhysRevE.68.061903. Real B-DNA in solution has about 10.5 base pairs per turn rather than 10.

Exercises for 3.7 Summary

P3.8 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as the motion of a charged particle among a few fixed point charges, the positions at which several equal charges confined to the surface of a sphere come to rest, the field vectors of a dipole drawn in three dimensions, the field of a finite line of charge compared with that of an infinite one, the field of a charged ring at points off its axis, or the field of a uniformly charged square plate at points near it.

Chapter 4

Gauss's Law

Chapter 3 found the electric field of a charge distribution by adding up the fields of all of its pieces. That method always works, but the integrals can be long, as the uniformly charged spherical shell of Section 3.5.3 showed. This chapter takes a different approach. Instead of asking what the field is at a particular point, it asks what Coulomb's law implies about the *divergence* of the field. The answer is Gauss's law, the first of Maxwell's equations. It has two main uses. Given a field, it gives the charge density that produces it, with no integrals at all. And given a charge distribution with enough symmetry, it gives the field in a few lines. It also explains the basic facts about conductors, which close the chapter.



Carl Friedrich Gauss (1777–1855, German) Often ranked among the greatest mathematicians, Gauss made fundamental contributions to number theory, algebra, statistics, and differential geometry; his *Disquisitiones Arithmeticae*, written when he was 21, appeared in 1801. The same year he predicted where the newly discovered asteroid Ceres, which had been lost in the glare of the Sun, would reappear, and it was found within half a degree of that position. He directed the observatory at Göttingen from 1807 until his death. In the 1830s he worked with the physicist Wilhelm Weber on magnetism; together they built one of the first electromagnetic telegraphs (1833) and organized measurements of Earth's magnetic field at observatories around the world. The law named after him came out of his work on the gravitational attraction of ellipsoids (1835); Lagrange had found it in 1773, in the same context. The CGS unit of magnetic field, the gauss, is also named after him. (Wikipedia)

4.1 Divergence of Coulomb's law

Start with the field of a single point charge q at the source point $\mathbf{r}'$,

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\frac{q\hat{\boldsymbol{\imath}}}{\imath^2},$$

where $\boldsymbol{\imath} = \mathbf{r} - \mathbf{r}'$ is the separation vector of Chapter 3. What is its divergence? The divergence is taken with respect to the field point $\mathbf{r}$, with the source point held fixed. Away from the charge, this field is a constant times the field $\hat{\mathbf{r}}/r^2$ of Figure 1.12b, shifted to $\mathbf{r}'$, and its divergence is zero: its field lines spread out, but none of them begins or ends there. At the charge itself, all of its field lines begin (or, for a negative charge, end). Section 1.6.3 showed how to make this precise: the divergence is a delta function, (1.45),

$$\nabla\cdot\left(\frac{\hat{\boldsymbol{\imath}}}{\imath^2}\right) = 4\pi\,\delta^3(\mathbf{r}-\mathbf{r}').$$

This is one of the cases where the formula for the divergence in spherical coordinates, (1.30), gives the wrong answer, zero, so be careful with it. The divergence of the point charge's field is therefore

$$\nabla\cdot\mathbf{E} = \frac{q}{4\pi\varepsilon_0}4\pi\,\delta^3(\mathbf{r}-\mathbf{r}') = \frac{q\,\delta^3(\mathbf{r}-\mathbf{r}')}{\varepsilon_0}. \tag{4.1}$$

But $q\,\delta^3(\mathbf{r}-\mathbf{r}')$ is exactly the charge density of a point charge, (3.10). So for a point charge, $\nabla\cdot\mathbf{E}=\rho/\varepsilon_0$.

The same is true for any collection of charges, because of superposition. The field of several point charges is the sum of their individual fields, $\mathbf{E}=\mathbf{E}_1+\mathbf{E}_2+\cdots$, and the divergence of a sum is the sum of the divergences (the divergence is a *linear* operator). Each term gives the charge density of one charge divided by ε_0, so

$$\begin{aligned}\nabla\cdot\mathbf{E} &= \nabla\cdot\mathbf{E}_1+\nabla\cdot\mathbf{E}_2+\cdots\\ &= \frac{\rho_1}{\varepsilon_0}+\frac{\rho_2}{\varepsilon_0}+\cdots\\ &= \frac{\rho}{\varepsilon_0},\end{aligned}$$

with ρ the total charge density. For a continuous charge distribution the sum becomes an integral, but the argument is the same. Take the divergence of the blueprint (3.13), written with $\boldsymbol{\mathscr{r}}/\mathscr{r}^3=\hat{\boldsymbol{\mathscr{r}}}/\mathscr{r}^2$:

$$\nabla\cdot\mathbf{E}(\mathbf{r})=\nabla\cdot\left[\frac{1}{4\pi\varepsilon_0}\int\rho(\mathbf{r}')\,\frac{\hat{\boldsymbol{\mathscr{r}}}}{\mathscr{r}^2}\,d\tau'\right].$$

The divergence involves derivatives with respect to the unprimed coordinates, while the integral is over the primed ones, so the divergence can be moved inside the integral, where it acts only on $\hat{\boldsymbol{\mathscr{r}}}/\mathscr{r}^2$ (the charge density depends only on $\mathbf{r}'$). Then, with (1.45) and the sifting property (1.42),

$$\nabla\cdot\mathbf{E}(\mathbf{r})=\frac{1}{4\pi\varepsilon_0}\int\rho(\mathbf{r}')\,4\pi\,\delta^3(\mathbf{r}-\mathbf{r}')\,d\tau'=\frac{\rho(\mathbf{r})}{\varepsilon_0}.$$

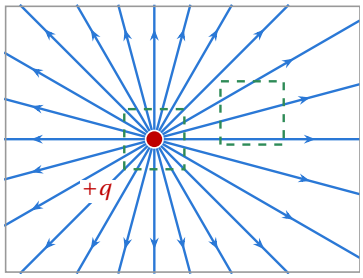


Figure 4.1 Field lines of a positive point charge q. Every line that enters the box on the right, where the field is strong, also leaves it, so the net outflow, and with it the divergence, is zero there. Only the box around the charge has lines leaving it that never entered it.

This is **Gauss's law**:

$$\nabla\cdot\mathbf{E}=\frac{\rho}{\varepsilon_0}. \qquad \text{(Gauss's law, differential form)} \quad (4.2)$$

With words, it says that the divergence of the electric field at any point is the charge density there, divided by ε_0. Electric field lines begin and end only where there is charge (Section 1.4.4). Where there is no charge, the divergence of $\mathbf{E}$ is zero, no matter how strong the field is (Figure 4.1).

Gauss's law is the first of Maxwell's equations (Section 2.1). We derived it from Coulomb's law, which holds only for stationary charges, but it turns out to hold for moving charges and changing fields as well. The electric field produced by a changing magnetic field (Chapter 15) has no divergence, so it adds nothing to the left side. Gauss's law alone does not determine $\mathbf{E}$, though. By the Helmholtz theorem (Section 2.3), a field that vanishes far away is determined by its divergence *and* its curl. Coulomb's law also says that the curl of an electrostatic field is zero, which is the subject of the next chapter.

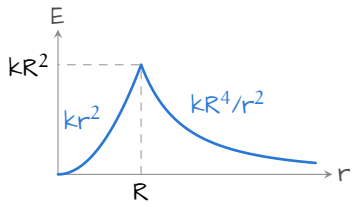
Example 4.1

In some region the electric field is

$$\mathbf{E} = \begin{cases} kr^2\,\hat{\mathbf{r}}, & r < R, \\ \dfrac{kR^4}{r^2}\,\hat{\mathbf{r}}, & r > R, \end{cases}$$

where k is a constant. What charge density produces it?

Solution:



Here what is given is a field, so the first sketch is a graph of its magnitude against r (drawn for $k > 0$). It rises as r^2 out to $r = R$, where it is kR^2, and falls off as $1/r^2$ beyond, and it is radial everywhere. By (4.2), $\rho = \varepsilon_0 \nabla \cdot \mathbf{E}$. The field has only an r component, so only the first term of the spherical divergence (1.30) survives.

Inside,

$$\begin{aligned} \rho &= \varepsilon_0 \left[\frac{1}{r^2}\frac{\partial (r^2 E_r)}{\partial r} + \cancel{\frac{1}{r\sin\theta}\frac{\partial (\sin\theta E_\theta)}{\partial \theta}} + \cancel{\frac{1}{r\sin\theta}\frac{\partial E_\phi}{\partial \phi}} \right] \\ &= \varepsilon_0 \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \cdot kr^2\right) \\ &= \varepsilon_0 \frac{4kr^3}{r^2} \\ &= 4\varepsilon_0 k\, r, \qquad r < R. \end{aligned}$$

Outside,

$$\begin{aligned} \rho &= \varepsilon_0 \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \cdot \frac{kR^4}{r^2}\right) \\ &= \varepsilon_0 \frac{1}{r^2}\frac{\partial}{\partial r}\left(kR^4\right) \\ &= 0, \qquad r > R. \end{aligned}$$

So the charge is confined to a ball of radius R, with a density that grows in proportion to r. The field outside the ball is far from zero, but there is no charge there. (How can there be no charge where the field is not zero? Answer in footnote.[1]) The field is also continuous at $r = R$, which, as Section 4.4 shows, means that there is no surface charge.

Reality check: By Coulomb's law $\varepsilon_0 E$ is a charge per unit area ($q/4\pi r^2$), so inside, $\rho = 4\varepsilon_0 E/r$ is a charge per unit volume: the units are right.

[1] Its field lines spread out, but none of them begins or ends outside the ball.

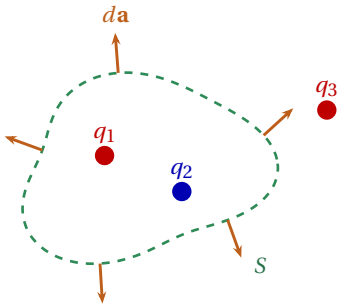


Figure 4.2 A closed surface S (dashed). Its area elements $d\mathbf{a}$ point outward, and only the charges inside it, here q_1 and q_2, make up q_{enc}; the field of q_3, outside, carries no net flux out of S.

4.2 Statement and physical meaning

Gauss's law has a second form, which is often more useful. Integrate (4.2) over a volume $\mathcal{V}$ bounded by a closed surface S. On the left, the divergence theorem (Section 1.5.3) turns the volume integral of the divergence into an integral over the surface; on the right, the volume integral of ρ is the total charge inside the surface, q_{enc}. The result is

$$\oint_S \mathbf{E}\cdot d\mathbf{a} = \frac{q_{\text{enc}}}{\varepsilon_0}, \qquad \text{(Gauss's law, integral form)} \quad (4.3)$$

where the area element $d\mathbf{a}$ points outward from the closed surface, which is called a **Gaussian surface** (Figure 4.2). The integral on the left is the **electric flux** through the surface. For any surface, open or closed, the flux is defined as

$$\Phi_E = \int_S \mathbf{E}\cdot d\mathbf{a}. \quad (4.4)$$

The dot product picks out the component of $\mathbf{E}$ perpendicular to the surface, so the flux adds up how much field passes *through* the surface, in the same way that the flux of the velocity of flowing water through a hoop is the volume of water passing through it per second. With words, Gauss's law says that *the electric flux out of any closed surface equals the charge enclosed by the surface, divided by* ε_0.

4.2.1 Counting field lines

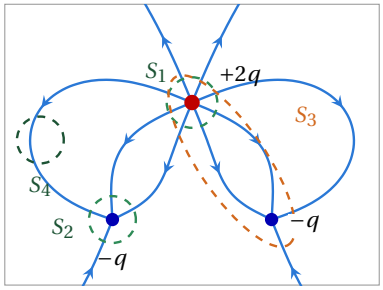


Figure 4.3 Field lines of $+2q$ and two charges $-q$: eight lines leave $+2q$ and four end on each $-q$. Lines that leave the frame return to one of the negative charges. The dashed curves are four Gaussian surfaces. Animation of one surface moving among these charges: the net number of lines leaving the surface changes only when a charge crosses the surface.

The flux has a simple picture in terms of field lines. Suppose the lines are drawn so that the number crossing a small area perpendicular to the field is proportional to the strength of the field (Section 3.3). Then the flux through a surface is proportional to the number of lines crossing it, with lines going out counted as positive and lines coming in as negative. Since field lines begin only on positive charges and end only on negative charges, the net number of lines leaving a closed surface counts the charge inside. A line that begins and ends outside the surface crosses it an even number of times, going in as often as it comes out, and adds nothing.

Figure 4.3 shows this for charges $+2q$, $-q$, and $-q$, drawn so that eight lines leave the $+2q$ and four end on each $-q$. (The figure shows only the lines that lie in the plane of the charges, but the counting works the same way in three dimensions.) Surface S_1, around the $+2q$, has eight lines going out: a flux proportional to $2q$. Surface S_2, around one $-q$, has four lines coming in: a flux proportional to $-q$. Surface S_3 encloses $+2q$ and one $-q$. Four of the lines leaving $+2q$ end on the $-q$ inside S_3, and the other four cross S_3 on their way to the other $-q$, so the net count is four lines out, matching the net charge $+q$ inside. (Two of the lines that end inside S_3 leave it and later come back in; each of them contributes one crossing out and one crossing in.) Surface S_4 encloses no charge, and every line that enters it also leaves, so the flux through it is zero, even though the field on it is not.

4.2.2 Gauss's law directly from Coulomb's law

Equation (4.3) also follows from Coulomb's law directly, without going through the divergence. Put a point charge q at the center of a sphere of radius R. On the sphere, $\mathbf{E} = q\hat{\mathbf{r}}/4\pi\varepsilon_0 R^2$ and $d\mathbf{a} = R^2 \sin\theta\, d\theta\, d\phi\, \hat{\mathbf{r}}$, so

$$\begin{aligned}\oint \mathbf{E}\cdot d\mathbf{a} &= \int_0^{2\pi}\int_0^{\pi} \frac{q}{4\pi\varepsilon_0 R^2} R^2 \sin\theta\, d\theta\, d\phi \\ &= \frac{q}{4\pi\varepsilon_0} 4\pi \\ &= \frac{q}{\varepsilon_0}.\end{aligned}$$

The radius dropped out: the area of the sphere grows as R^2, and the field falls as $1/R^2$. This is where the inverse-square law matters; for any other power of R, the flux would depend on the size of the sphere. Next, any other closed surface around the charge is crossed by the same field lines, so it has the same flux, while a closed surface that does not enclose the charge has as many lines going in as coming out, so its flux is zero. Finally, by superposition, the flux of the field of many charges is the sum of their individual fluxes, and only the enclosed charges contribute. Charges outside the surface change the field *on* the surface, but not the total flux through it.

4.2.3 Two forms, two uses

The two forms of Gauss's law, (4.2) and (4.3), contain the same information; the divergence theorem turns either one into the other. The integral form handles point, line, and surface charges naturally, since they simply contribute to q_{enc}, while in the differential form they need delta functions. In practice the two forms answer opposite questions. The differential form goes from field to charge: given $\mathbf{E}$, take its divergence and multiply by ε_0 to find ρ, as in Example 4.1. The integral form goes from charge to field: given ρ, it can sometimes be solved for $\mathbf{E}$, which is the subject of the next section.

Keep two cautions in mind. First, a zero flux does not mean a zero field. A closed box in a uniform field encloses no charge and has no net flux, but the field on its faces is not zero. The same is true of the box beside the point charge in Figure 4.1. Second, Gauss's law is true for every closed surface around every charge distribution, but it can be solved for the field only in a few special cases. The law contains the field inside an integral, and for most charge distributions there is no way to get it out.

4.3 Solving for E when given ρ: Gauss's law method (high symmetry)

To solve (4.3) for $\mathbf{E}$, the field has to be taken outside the integral, and that is possible only when the Gaussian surface can be divided into pieces on each of

which one of two things is true:

1. **E** is perpendicular to the surface (parallel to $d\mathbf{a}$), and its magnitude is the same everywhere on that piece. Then $\int \mathbf{E}\cdot d\mathbf{a}$ over that piece is simply $\pm EA$, where A is the piece's area.

2. **E** is parallel to the surface (perpendicular to $d\mathbf{a}$). Then the flux through that piece is zero.

This requires a great deal of symmetry, and you have to know the direction of the field, and what its magnitude can depend on, before you start. Gauss's law gives the magnitude, but the direction always comes from symmetry. The procedure is:

1. Use symmetry to answer two separate questions, each with its own reason: which components of **E** are zero, and which coordinates **E** can depend on (Section 4.3.2).

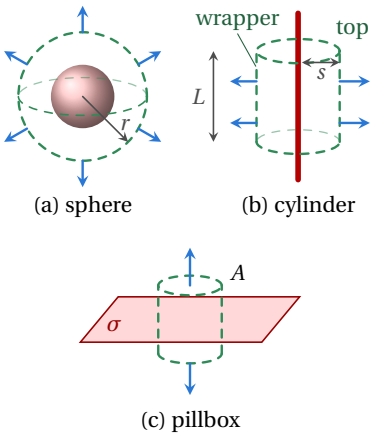
2. Choose a Gaussian surface that passes through the point where you want the field and that matches the symmetry, so that each piece of it satisfies condition 1 or 2 above. Draw it, and list its pieces.

3. Calculate the enclosed charge q_{enc}, by integrating the charge density if necessary.

4. Write down Gauss's law, simplify the left side to E times an area, solve for E, and give **E** its direction.

The left side is easy by design. The only hard part, if any, is calculating q_{enc}.

! **Colton's Second Commandment:** When using Gauss's law to determine electric fields, thou shalt always draw a Gaussian surface.



(a) sphere (b) cylinder (c) pillbox

Figure 4.4 The three kinds of Gaussian surface (dashed): (a) a sphere, for spherical symmetry; (b) a cylinder of radius s and length L, for an infinitely long line or cylinder; (c) a pillbox with faces of area A, for an infinite plane or slab.

4.3.1 Colton's Second Commandment

Step 2 deserves emphasis. Gauss's law is about a specific closed surface, and an argument that never says what that surface is cannot check whether the field really is perpendicular and constant on each piece of it, or whether every piece has been accounted for. In my experience, students who skip the drawing are the ones who end up applying the law where it does not work, or who leave out part of the surface. This is **Colton's Second Commandment**:

> *When using Gauss's law to determine electric fields, thou shalt always draw a Gaussian surface.*

Draw it explicitly, and account for every piece of it.

4.3.2 The three kinds of symmetry

There are only three kinds of symmetry for which this works, and therefore only three kinds of Gaussian surface (Figure 4.4).

- **Spherical symmetry**: the charge density depends only on r. The field has only an r component, and that component depends only on r: $\mathbf{E} = E(r)\,\hat{\mathbf{r}}$. The Gaussian surface is a concentric sphere.
- **Cylindrical symmetry**: the charge density depends only on s, and the distribution is infinitely long. The field has only an s component, and that component depends only on s: $\mathbf{E} = E(s)\,\hat{\mathbf{s}}$. The Gaussian surface is a coaxial cylinder of finite length L, with three pieces: top, bottom, and wrapper (the curved side).
- **Planar symmetry**: the charge density depends only on z, and the distribution is infinite in x and y. The field has only a z component, and that component depends only on z: $\mathbf{E} = E(z)\,\hat{\mathbf{z}}$. The Gaussian surface is a "pillbox", a short cylinder (or box) with faces of area A parallel to the plane.

Each item in this list answers the two questions of step 1 separately, and each answer needs its own reason. The first question is which components of **E** are zero. A component is zero if no piece of the charge contributes to it, if the pieces can be paired so that their contributions to it cancel, or if a rotation about a line through the field point would leave the charge distribution unchanged and yet reverse that component. The second question is which coordinates **E** can depend on. It cannot depend on a coordinate if moving the field point along that coordinate, by sliding it or by rotating it, leaves the charge distribution looking exactly the same from the new point.

For spherical symmetry, a rotation by 180° about the line from the center through the field point leaves the charge distribution unchanged, and therefore cannot change the field at that point; yet it would reverse E_θ and E_ϕ, which are perpendicular to that line. So only E_r is left. And the distribution looks the same from every point at the same distance from the center, so E_r depends only on r. Example 4.3 makes both arguments in full for an infinite line of charge. Keep the two questions apart, because the Gaussian surface needs both answers: the components tell you on which of its pieces **E** is perpendicular to the surface and on which it is parallel, and the dependence tells you whether **E** has the same magnitude all over a piece.

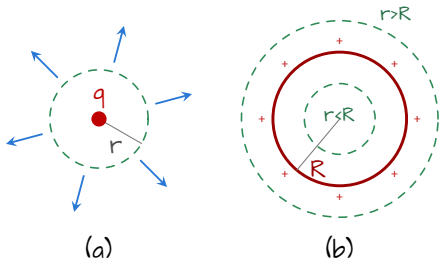
Note that the symmetry must be in the charge distribution, not only in the shape of the object. A sphere with $\rho = k\cos\theta$ is a sphere, but its field is not radial, and Gauss's law cannot be solved for it. (Figure 7.29 shows a similar field.)

4.3.3 Worked examples

Example 4.2

Use Gauss's law to find the field of (a) a point charge q, and (b) a uniformly charged spherical shell of radius R and total charge Q, both inside and outside.

Solution:



First we sketch what we are given, (a) the point charge and (b) the shell, and in each the Gaussian surface that its symmetry calls for. By symmetry the field of each points radially away from the center, and its magnitude can depend only on the distance r. So each Gaussian surface is a sphere of radius r centered on the charge: one for the point charge, and for the shell one inside it and one outside it.

(a) On the sphere,

$$\begin{aligned}\oint \mathbf{E}\cdot d\mathbf{a} &= \oint E\, da \\ &= E\oint da \\ &= E\,(4\pi r^2).\end{aligned}$$

(Why can E come out of the integral? Answer in footnote.[2]) The enclosed charge is q, so $E\,(4\pi r^2) = q/\varepsilon_0$, and

$$\mathbf{E} = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}\hat{\mathbf{r}},$$

which is Coulomb's law, as it must be.

(b) The shell has the same symmetry, so the Gaussian surface is again a concentric sphere of radius r, and the left side is again $E\,(4\pi r^2)$. Only the enclosed charge changes. Outside the shell ($r > R$) the sphere encloses all of the charge, Q; inside ($r < R$) it encloses none. So

$$\mathbf{E} = \begin{cases} 0, & r < R, \\ \dfrac{1}{4\pi\varepsilon_0}\dfrac{Q}{r^2}\hat{\mathbf{r}}, & r > R. \end{cases}$$

Reality check: This agrees with (3.21), which solved the same problem. However, the script-r method obtains the same answer only after a challenging setup and integral (Exercise P3.5).

Example 4.3

An infinitely long straight line of charge, with constant line charge density λ, lies along the z axis. Find the electric field everywhere.

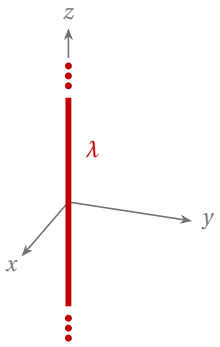
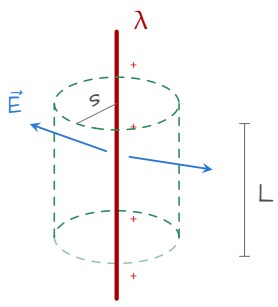


Figure 4.5 The infinite line of charge of Example 4.3, along the z axis.

[2] By symmetry, $\mathbf{E}$ has only an r component, so it is parallel to $d\mathbf{a}$, and that component depends only on r, so it has the same magnitude at every point of the sphere.

Solution:



We begin, as always, with a sketch of what is given: the line of charge along the z axis (drawn for positive λ) and its field, which points directly away from the line, along $\hat{\mathbf{s}}$, as the argument below shows. So the Gaussian surface, drawn dashed, is a cylinder of radius s and length L, coaxial with the line. Gauss's law will give the magnitude of the field, but its direction, and what its magnitude can depend on, have to come from symmetry, so here we make that argument in full (Figure 4.6).

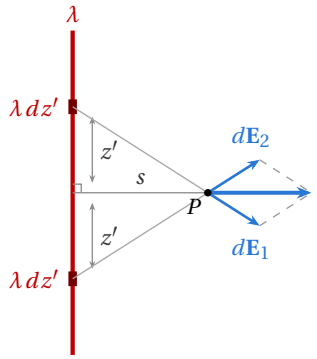


Figure 4.6 The symmetry argument of Example 4.3, in the plane of the line and P. The pieces $\lambda\,dz'$ above and below P give fields $d\mathbf{E}_1$ and $d\mathbf{E}_2$ of the same magnitude; their z components cancel, and their sum (thick) points directly away from the line.

Write the field at a point P in cylindrical components, $E_s\,\hat{\mathbf{s}} + E_\phi\,\hat{\boldsymbol{\phi}} + E_z\,\hat{\mathbf{z}}$. The line and P lie in one plane, and the field of each piece of the line points directly away from that piece, so it lies in that plane too; $\hat{\boldsymbol{\phi}}$ points out of the plane, so $E_\phi = 0$. Next, pair each piece of the line a distance z' above P with the piece the same distance below it. The two pieces are equally far from P, so their fields have the same magnitude: their z components cancel, and their s components add. The line is infinite, so every piece has a partner, and $E_z = 0$. Finally, the magnitude can depend only on s: every value of z looks the same, because the line is infinite, and every value of ϕ looks the same, because the line is straight. So $\mathbf{E} = E(s)\,\hat{\mathbf{s}}$.

The cylinder has three pieces:

$$\oint \mathbf{E}\cdot d\mathbf{a} = \int_{\text{top}} \mathbf{E}\cdot d\mathbf{a} + \int_{\text{bottom}} \mathbf{E}\cdot d\mathbf{a} + \int_{\text{wrapper}} \mathbf{E}\cdot d\mathbf{a}.$$

On the top and bottom, $\mathbf{E}$ is parallel to the surface, so those two integrals are zero. On the wrapper, $\mathbf{E}$ is perpendicular to the surface and has the same magnitude everywhere, so that integral is E times the wrapper's area, $2\pi sL$. The enclosed charge is the charge on a length L of the line, λL. So $E\,(2\pi sL) = \lambda L/\varepsilon_0$, and

$$\mathbf{E} = \frac{\lambda}{2\pi\varepsilon_0\, s}\,\hat{\mathbf{s}}.$$

Reality check: This is the same as (3.17), and the length L of the Gaussian cylinder cancels, as it should, since it was our choice.

For a line of finite length, this argument fails, even though the line still looks symmetric. Near the ends, the field is not perpendicular to the wrapper, it is not the same at every point of the wrapper, and some of it passes through the top and bottom (Figure 4.7). Gauss's law is still true for the finite line, but it cannot be solved for the field. For a long line, at points much closer to the line than to either end, the infinite-line result is a good approximation.

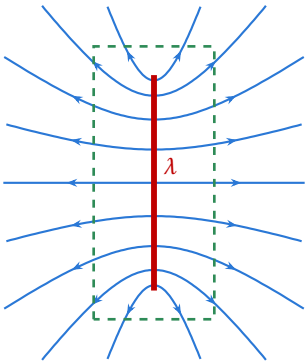


Figure 4.7 Field lines of a uniformly charged line of finite length, in a plane through the line, with a coaxial Gaussian cylinder (dashed). Near the ends the field is neither perpendicular to the wrapper nor the same all over it, and some of it leaves through the top and bottom. (The field lines are hyperbolas whose foci are the ends of the line.)

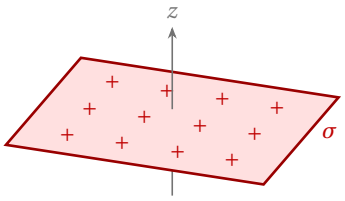
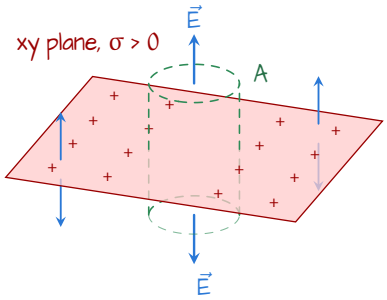


Figure 4.8 The charged plane of Example 4.4, the xy plane, with σ drawn positive.

Example 4.4

An infinite plane, the xy plane, carries a constant surface charge density σ. Find the electric field everywhere.

Solution:



To start out, we need to visualize the situation, so we draw the xy plane with positive charge density (it's OK to assume σ is positive for visualization purposes). Then we recognize from the picture that the electric field points directly away from the plane: along $+\hat{\mathbf{z}}$ above the plane and along $-\hat{\mathbf{z}}$ below it. By symmetry, its magnitude can depend only on the distance from the plane, $|z|$, and it is the same at z and $-z$.

The Gaussian surface, drawn dashed in the sketch as Colton's Second Commandment requires, is a pillbox with faces of area A at heights z and $-z$. On the wrapper, $\mathbf{E}$ is parallel to the surface, so there is no flux through it. On the top face, $\mathbf{E}$ points out of the pillbox, and the flux is EA. On the bottom face, $\mathbf{E}$ points down, and yet that flux is also $+EA$. (Why is the flux through the bottom face $+EA$, and not $-EA$? Answer in footnote.[3]) The enclosed charge is σA, so $2EA = \sigma A/\varepsilon_0$, and

$$\mathbf{E} = \begin{cases} +\dfrac{\sigma}{2\varepsilon_0}\hat{\mathbf{z}}, & z > 0, \\ -\dfrac{\sigma}{2\varepsilon_0}\hat{\mathbf{z}}, & z < 0. \end{cases} \tag{4.5}$$

The field does not depend on the distance from the plane at all.

Reality check: That fits the picture of field lines (Section 3.3): the lines of an infinite plane are straight and parallel, so they never spread apart, and their density stays the same.

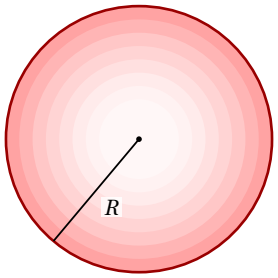


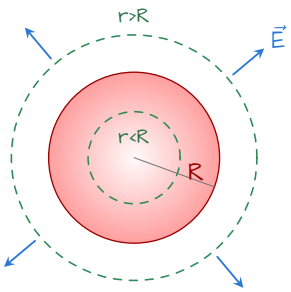
Figure 4.9 The ball of Example 4.5, in cross section, shaded darker where its charge density is larger.

Example 4.5

A ball of radius R has the charge density $\rho(r) = \rho_0\, r^2/R^2$, which is zero at the center and ρ_0 at the surface. Find the electric field inside and outside the ball.

[3] The outward normal of the bottom face points down too, so $\mathbf{E}\cdot d\mathbf{a}$ is positive there: what matters for the flux is whether the field leaves the surface, not which way it points.

Solution:



A sketch comes first: the ball, shaded darker toward its surface, where ρ is largest, and a concentric Gaussian sphere of radius r inside it and another outside it. The charge density depends only on r, so the field is radial and depends only on r. As in Example 4.2, the left side of Gauss's law is $E(4\pi r^2)$. This time, though, the right side takes some work: the enclosed charge has to be found by integrating the charge density.

Just because the left side is trivial does not mean that the right side is. Inside the ball ($r < R$),

$$q_{\text{enc}} = \int \rho \, d\tau' = \int_0^{2\pi}\int_0^{\pi}\int_0^{r} \frac{\rho_0 \, r'^2}{R^2} \, r'^2 \sin\theta' \, dr' \, d\theta' \, d\phi'$$
$$= \frac{4\pi\rho_0}{R^2}\int_0^{r} r'^4 \, dr' = \frac{4\pi\rho_0 \, r^5}{5R^2}.$$

Notice the primes: the problem gives the charge density as a function of r, but inside the integral it must be written in terms of the integration variable r', which runs from 0 up to the radius r of the Gaussian sphere. Then Gauss's law gives

$$E(4\pi r^2) = \frac{4\pi\rho_0 \, r^5}{5\varepsilon_0 R^2},$$

and so

$$\mathbf{E} = \frac{\rho_0 \, r^3}{5\varepsilon_0 R^2}\hat{\mathbf{r}}, \qquad r < R.$$

Outside the ball, the only change is in the enclosed charge: the upper limit of the integral becomes R. (Why is the upper limit R, and not r? Answer in footnote.[4]) This makes $q_{\text{enc}} = 4\pi\rho_0 R^3/5$, so we have

$$\mathbf{E} = \frac{\rho_0 \, R^3}{5\varepsilon_0 \, r^2}\hat{\mathbf{r}}, \qquad r > R.$$

Reality check: At $r = R$ both expressions give $\rho_0 R/5\varepsilon_0$, so the field is continuous (Figure 4.10). That is what you should expect, since there is no surface charge at $r = R$; Section 4.4 shows that the perpendicular field only has a discontinuity where there is a surface charge.

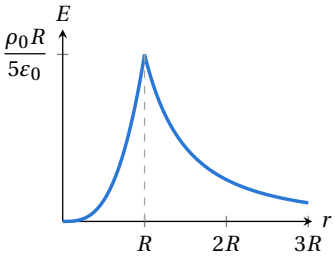


Figure 4.10 The field of the ball of Example 4.5. It grows as r^3 inside, falls as $1/r^2$ outside, and is continuous at $r = R$.

[4] The density is zero beyond R, so a larger sphere encloses no more charge.

4.3.4 Superposition

Gauss's law can sometimes be used even when the charge distribution as a whole has no useful symmetry, because the fields of its parts add. If each part has one of the three symmetries, find the field of each part with its own Gaussian surface and add the results.

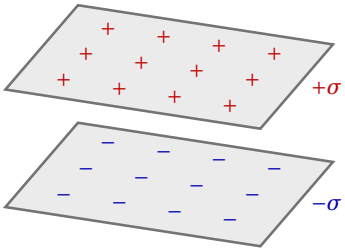
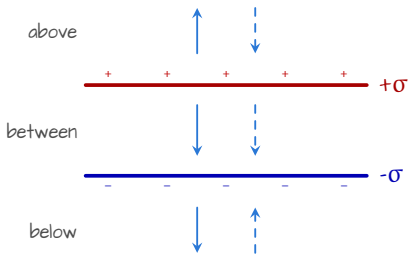


Figure 4.11 The two charged planes of Example 4.6; only a piece of each is drawn.

Example 4.6

Two infinite parallel planes carry surface charge densities $+\sigma$ (the upper one) and $-\sigma$ (the lower one). Find the field above, between, and below them.

Solution:



Before writing any equations, we draw the two planes edge-on and, in each of the three regions, the field that each plane makes by itself. By (4.5), each has magnitude $\sigma/2\varepsilon_0$ and points away from the positive plane (solid arrows) and toward the negative one (dashed arrows).

Above both planes, the field of the upper plane points up and the field of the lower plane points down, and they cancel. (Why do they cancel at every height? Answer in footnote.[5]) Below both planes they cancel again. Between the planes, both fields point down, from the positive plane toward the negative one, and they add:

$$\mathbf{E} = \begin{cases} 0, & \text{above both planes,} \\ -\dfrac{\sigma}{\varepsilon_0}\hat{\mathbf{z}}, & \text{between the planes,} \\ 0, & \text{below both planes.} \end{cases}$$

This is the field of an ideal parallel-plate capacitor (Section 5.8).

Reality check: A pillbox that encloses a piece of both planes encloses no net charge, and the flux through it is indeed zero, since the field is zero above and below the planes.

It is tempting to get this result instead by drawing a pillbox around just one of the planes and writing $2EA = \sigma A/\varepsilon_0$, as in Example 4.4. That gives $\sigma/2\varepsilon_0$, which is wrong. Gauss's law is still true for that pillbox, but the field is no longer the same on its two faces: it is σ/ε_0 on the face between the planes and zero on the other. The symmetry that made the two faces equal in Example 4.4 belonged to a single plane, and it is lost when a second plane is added. If you already know that the field outside is zero, then the pillbox around one plane does give the right answer, because only one face carries any flux.

[5] Each plane's field has the same magnitude, $\sigma/2\varepsilon_0$, at every distance from the plane, by (4.5).

The field of a charged cube is a good example of a case where Gauss's law cannot be used. The cube looks symmetric, but its field lines bend near its edges and corners so there is no closed surface on which the field is everywhere perpendicular to the surface and constant in magnitude. The same is true for a finite line charge, as Example 4.3 explained, and for a disk, like the one whose field on its axis was found by the script-r method in Example 3.5. For these, use the script-r method or its numerical version.

4.4 Boundary condition on $E^\perp$

Gauss's law also tells us how the field changes across any surface charge. Consider a small patch of a surface carrying the surface charge density σ, and let "above" be the side that a unit normal $\hat{\mathbf{n}}$ points toward. Draw a pillbox that straddles the surface, with faces of area A just above and just below it (Figure 4.12). The faces are small enough that the surface is flat across them and the field is constant over each. Now make the pillbox very short. The area of its wrapper goes to zero, and so does the flux through it (as long as the field is finite). The flux through the top face is $E^\perp_{\text{above}} A$ and the flux through the bottom face is $-E^\perp_{\text{below}} A$, where $E^\perp$ is the component of $\mathbf{E}$ along $\hat{\mathbf{n}}$; the bottom term has a minus sign because the outward normal of the bottom face is $-\hat{\mathbf{n}}$. The enclosed charge is σA, so $E^\perp_{\text{above}} A - E^\perp_{\text{below}} A = \sigma A/\varepsilon_0$, or

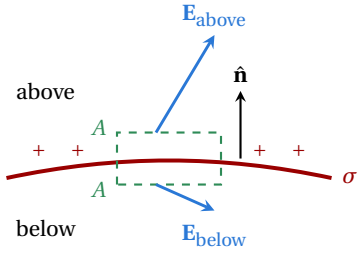


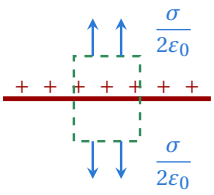
Figure 4.12 A short pillbox straddling a surface charge σ. Only the components of the fields along $\hat{\mathbf{n}}$ contribute flux through the faces, and they differ by σ/ε_0.

$$E^\perp_{\text{above}} - E^\perp_{\text{below}} = \frac{\sigma}{\varepsilon_0}. \tag{4.6}$$

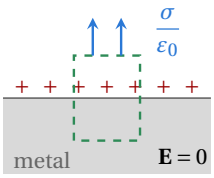
With words, when an electric field encounters a sheet of surface charge, the charge will create a discontinuity in the perpendicular component of the field, of magnitude σ/ε_0. A jump in the perpendicular component across a surface is a divergence concentrated on the surface (Exercise P1.21), and here, by Gauss's law, that divergence is the surface charge divided by ε_0. The infinite plane of Example 4.4 is an example: its field goes from $-\sigma/2\varepsilon_0$ to $+\sigma/2\varepsilon_0$. So is the spherical shell of Example 4.2: its field jumps from zero to $Q/4\pi\varepsilon_0 R^2 = \sigma/\varepsilon_0$. Conversely, a perpendicular field that is continuous, as in Examples 4.1 and 4.5, means there is no surface charge. The components of $\mathbf{E}$ *parallel* to the surface do not jump at all; showing this needs the curl of $\mathbf{E}$, which is in the next chapter (Section 5.2). Together with the continuity of the potential, these are the boundary conditions used in Section 7.1.1.

4.5 Conductors and induced charges

A **conductor** is a material in which charges are free to move. In a metal the mobile charges are electrons, a vast number of them, which can move anywhere within the metal but cannot easily leave it. In electrostatics nothing is moving. That has several consequences.

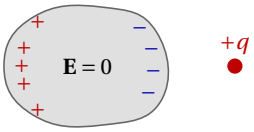


(a) insulating sheet

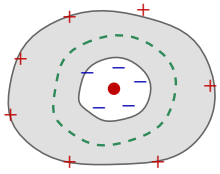


(b) conductor

Figure 4.13 (a) At an insulating sheet, flux leaves through both faces of the pillbox, and $E = \sigma/2\varepsilon_0$ on each side. (b) At the surface of a conductor, the bottom face is in the metal, where $\mathbf{E} = 0$, and all of the flux leaves through the top: $E = \sigma/\varepsilon_0$.



(a) neutral conductor



(b) charge in a cavity

Figure 4.14 (a) A neutral conductor near a point charge $+q$. Negative charge is induced on the near side and positive charge on the far side, and inside the metal the field is zero. (b) A neutral conductor with a cavity containing $+q$. The cavity wall carries $-q$ and the outer surface carries $+q$; the dashed Gaussian surface lies inside the metal.

1. **The field inside a conductor is zero.** If it were not, the free charges would move in response to it, and the situation would not be static. When a field is first applied, the charges do move, and they keep rearranging themselves until the field they produce cancels the applied field everywhere inside the metal. In a good conductor this takes a tiny fraction of a nanosecond.

2. **There is no net charge inside a conductor; any net charge lies on its surface.** This follows directly from (4.2): if $\mathbf{E} = 0$, then $\rho = \varepsilon_0 \nabla \cdot \mathbf{E} = 0$. Or, with the integral form, draw a Gaussian surface just inside the conductor's surface. The field is zero everywhere on it, so the enclosed charge must be zero. The excess charge has nowhere to go but the surface.

3. **Just outside a conductor, the field is perpendicular to the surface.** If the field had a component parallel to the surface, it would push the surface charges along the surface, and again the situation would not be static.

4. **Just outside a conductor, the field is $\mathbf{E} = (\sigma/\varepsilon_0)\,\hat{\mathbf{n}}$,** where σ is the local surface charge density and $\hat{\mathbf{n}}$ is the outward normal. This is (4.6) with $E^{\perp}_{\text{below}} = 0$; with a pillbox, the bottom face is inside the metal, where there is no field, and all of the flux goes through the top face (Figure 4.13b).

These facts hold only in the static case. A wire carrying a current has a field inside it; that is what pushes the current along (Section 19.2.1). The conductor is also an *equipotential*, which the next chapter shows (Section 5.7).

The field just outside a conductor, σ/ε_0, is twice the field of an insulating sheet with the same σ, $\sigma/2\varepsilon_0$, and the factor of two surprises many people. The difference is where the flux goes. At the insulating sheet, the flux from the charge in the pillbox leaves through both faces; at the conductor, it all leaves through the top face. The two results are also consistent with each other. An isolated thin conducting plate carrying a total charge per unit area σ_t puts half of it, $\sigma_t/2$, on each face, so the field just outside each face is $(\sigma_t/2)/\varepsilon_0 = \sigma_t/2\varepsilon_0$: the same as for an insulating sheet carrying σ_t.

4.5.1 Induced charge

Bring a positive point charge q near a neutral conductor (Figure 4.14a). The free electrons in the conductor are attracted toward q, so negative charge collects on the near side, leaving positive charge on the far side. This separation of charge is called **induced charge**. It continues until the field of the induced charges exactly cancels the field of q everywhere inside the conductor. The total induced charge is zero, since the conductor is neutral, but the negative part is closer to q than the positive part, so the net force on the conductor is *attractive*. This is why a charged object attracts an uncharged piece of metal. The field outside the conductor is complicated; for a few simple shapes, Chapter 7 finds it exactly.

Now put the charge q inside a cavity in the conductor (Figure 4.14b). Draw a Gaussian surface that surrounds the cavity but lies entirely inside the metal.

The field is zero everywhere on it, so the total charge it encloses is zero. That enclosed charge consists of q and whatever charge sits on the wall of the cavity, so *the cavity wall carries a total charge* $-q$, whatever the shape of the cavity and wherever the charge is inside it. If the conductor is neutral, its outer surface must then carry $+q$; if it has a net charge Q, the outer surface carries $Q+q$. The charge on the outer surface arranges itself as if the cavity and its contents were not there: its distribution, and the field outside the conductor, depend only on the shape of the outer surface and the total charge on it, not on where q is inside the cavity. This follows from the second uniqueness theorem (Section 7.1.4). For a spherical conductor, the outer charge is spread uniformly, and the field outside is that of a point charge $Q+q$ at the center, even if q is far off center.

Now consider charges outside the conductor. A cavity that contains no charge has no field inside it, whatever charges are outside the conductor. The charges on the conductor move until they cancel the external field everywhere inside the metal, and in doing so they make the whole conductor, including the wall of the cavity, a single equipotential (Section 5.7). An empty region enclosed by a single equipotential surface turns out to have that same potential everywhere inside it (as we will see in Section 7.1.2), and a constant potential means that there is no field. That is the principle of a **Faraday cage**, a conducting enclosure that shields its contents from external fields (Figure 4.15). Strictly, this is true only for static fields. Fields that change in time penetrate a conductor to some extent, to a depth that depends on the frequency and on the conductivity of the material, as Section 22.2.2 discusses. Shielding in the opposite direction is not automatic; as the last paragraph showed, a charge inside the cavity of an isolated conductor produces a field outside, through the charge it induces on the outer surface. If there are no other charges nearby, connecting the conductor to the ground (Section 7.1.1) drains that outer charge away, and the field outside disappears.

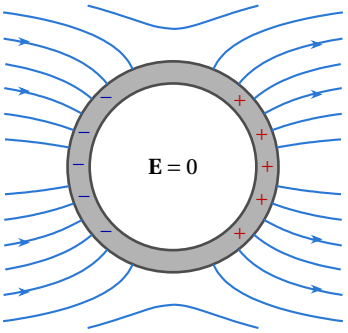


Figure 4.15 A hollow conducting sphere in a uniform field. The field lines outside end on the negative charge induced on the left and start again from the positive charge on the right, and the field inside the cavity is zero.

4.5.2 The force on a surface charge

The charge on the surface of a conductor sits in an electric field, so it feels a force. Which field acts on it is not obvious: just outside the surface the field is σ/ε_0, just inside it is zero, and the charge is right at the boundary. The answer turns out to be the average of the two, $\sigma/2\varepsilon_0$, so that the force per unit area is

$$\mathbf{f} = \frac{\sigma^2}{2\varepsilon_0}\,\hat{\mathbf{n}} = \frac{\varepsilon_0}{2}\,E^2\,\hat{\mathbf{n}}, \tag{4.7}$$

where $E = \sigma/\varepsilon_0$ is the field just outside and $\hat{\mathbf{n}}$ is the outward normal. Exercise P4.10 derives this. Since σ^2 is never negative, the force always points outward, whether the surface charge is positive or negative: the surface of a charged conductor is pulled out into the field. This force per unit area acts like a pressure, except that it pulls the surface outward instead of pushing on it, and it is usually called an electrostatic *stress*.

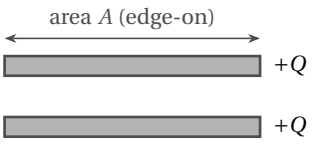


Figure 4.16 The two conducting plates of Example 4.7, edge-on.

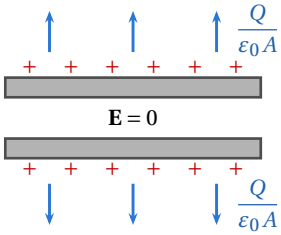
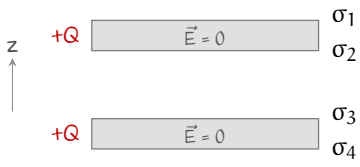


Figure 4.17 The answer to Example 4.7: all of the charge sits on the two outer faces, the field between the plates is zero, and the field outside them is $Q/\varepsilon_0 A$, pointing away.

Example 4.7

Two large, identical, parallel conducting plates, each of area A, each carry a charge $+Q$. Ignoring effects near the edges, find how the charge is distributed on their four faces, and the force between the plates.

Solution:



We start, as in Example 4.6, by drawing the plates edge-on, with their faces numbered 1 to 4 from top to bottom (faces 1 and 2 belong to the upper plate) and surface charge densities σ_1 to σ_4. Each face acts like an infinite plane, producing a field of magnitude $\sigma_i/2\varepsilon_0$ pointing away from it (for $\sigma_i > 0$).

Inside the upper plate the field must be zero. (Why must the field be zero inside the plate? Answer in footnote.[6]) There, face 1 is above and faces 2, 3, and 4 are below, so, taking up as positive,

$$\frac{1}{2\varepsilon_0}(-\sigma_1 + \sigma_2 + \sigma_3 + \sigma_4) = 0.$$

Inside the lower plate, faces 1, 2, and 3 are above and face 4 is below:

$$\frac{1}{2\varepsilon_0}(-\sigma_1 - \sigma_2 - \sigma_3 + \sigma_4) = 0.$$

Adding and subtracting these gives $\sigma_1 = \sigma_4$ and $\sigma_2 = -\sigma_3$. Each plate carries Q, so $\sigma_1 + \sigma_2 = Q/A$ and $\sigma_3 + \sigma_4 = Q/A$. Adding these, and using $\sigma_3 = -\sigma_2$ and $\sigma_4 = \sigma_1$, gives $2\sigma_1 = 2Q/A$. So

$$\sigma_1 = \sigma_4 = \frac{Q}{A}, \qquad \sigma_2 = \sigma_3 = 0.$$

All of the charge is on the outer faces (Figure 4.17). The field is zero between the plates and $Q/\varepsilon_0 A$ outside them. (The same steps for plates carrying any two charges Q_1 and Q_2 show that the two outer faces always carry equal charges, $(Q_1 + Q_2)/2$ each, and the two inner faces always carry equal and opposite charges; see Exercise P4.12.)

The force on the upper plate comes from the charge on face 1, the only charged face of that plate. By (4.7) it is $\sigma_1^2/2\varepsilon_0$ per unit area, pointing up, so

$$\mathbf{F} = \frac{\sigma_1^2}{2\varepsilon_0} A\hat{\mathbf{z}} = \frac{Q^2}{2\varepsilon_0 A}\hat{\mathbf{z}}.$$

The plates repel, as they should.

Reality check: The field of the lower plate at the upper plate is that of a sheet with total charge per area Q/A, namely $Q/2\varepsilon_0 A$, and that times the charge Q of the upper plate gives the same force.

[6] The plate is a conductor, and in electrostatics the field inside a conductor is zero; otherwise its free charges would move (Section 4.5).

4.6 Summary

Gauss's law follows from Coulomb's law and superposition, through one fact about the inverse-square field: its divergence is zero everywhere except at the charge, where it is a delta function. The law relates the field to the charge in two equivalent ways.

$\nabla \cdot \mathbf{E} = \dfrac{\rho}{\varepsilon_0}$	Gauss's law, differential form: field lines begin and end only on charge
$\Phi_E = \displaystyle\int_S \mathbf{E} \cdot d\mathbf{a}$	the electric flux through a surface
$\displaystyle\oint_S \mathbf{E} \cdot d\mathbf{a} = \frac{q_{\text{enc}}}{\varepsilon_0}$	Gauss's law, integral form: the flux out of a closed surface
$E^{\perp}_{\text{above}} - E^{\perp}_{\text{below}} = \dfrac{\sigma}{\varepsilon_0}$	the perpendicular field jumps across a surface charge
$\mathbf{f} = \dfrac{\sigma^2}{2\varepsilon_0}\hat{\mathbf{n}}$	the force per unit area (stress) on the surface of a conductor

The Gauss's law method. Use symmetry to answer two separate questions: which components of the field are zero, and which coordinates it can depend on; choose a Gaussian surface through the field point on which every piece has the field either perpendicular and constant or parallel; calculate q_{enc}; solve. There are only three kinds of symmetry (spherical, cylindrical, and planar), with a sphere, a cylinder, and a pillbox as the Gaussian surfaces. Superposition extends the method to combinations of these. *Colton's Second Commandment:* when using Gauss's law to determine electric fields, thou shalt always draw a Gaussian surface.

Results worth remembering. A point charge or a uniformly charged spherical shell, outside: $Q/4\pi\varepsilon_0 r^2$ (and zero inside the shell). An infinite line: $\lambda/2\pi\varepsilon_0 s$. An infinite plane: $\sigma/2\varepsilon_0$, independent of distance.

Conductors in electrostatics have $\mathbf{E} = 0$ and $\rho = 0$ inside; any net charge is on the surface, and the field just outside is $(\sigma/\varepsilon_0)\,\hat{\mathbf{n}}$, perpendicular to the surface. A charge q in a cavity induces $-q$ on the cavity wall. A neutral conductor near a charge acquires induced charges that cancel the field inside it, and it is attracted to the charge.

Exercises

Exercises for 4.1 Divergence of Coulomb's law

Warm-up questions Read Section 4.1, then answer before class

W4.1 The field lines in Figure 4.1 spread apart as they go outward, and a classmate concludes that the field must have a divergence everywhere, not only at the charge. The spreading is real. Where does it show up in the flux through the box on the right, and what makes up for it, so that the net flux out of the box is zero?

W4.2 The divergence of $\hat{\mathbf{r}}/r^2$ is zero everywhere except at the origin, and yet its flux out of any sphere centered on the origin is 4π. Explain in your own words how both statements can be true, and what they tell you about the divergence at the origin.

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P4.1 Suppose that the electric field in some region is found to be $\mathbf{E} = kr^5\hat{\mathbf{r}}$, where k is a constant.

(a) Use Gauss's law in differential form to find the charge density that produces this field.

(b) Find the total charge inside a sphere of radius R centered on the origin in two ways: by integrating the charge density you found in part (a), and by using Gauss's law in integral form.

P4.2 A line of charge with constant line charge density λ lies along the z axis. Its field is $\mathbf{E} = \lambda\hat{\mathbf{s}}/2\pi\varepsilon_0 s$ (Example 4.3), and its charge density can be written as $\rho = \lambda\delta(x)\delta(y)$ (Section 1.6.2). Use the divergence $\nabla\cdot(\hat{\mathbf{s}}/s) = 2\pi\delta(x)\delta(y)$ of Exercise P1.20 to show that the field of the line satisfies $\nabla\cdot\mathbf{E} = \rho/\varepsilon_0$, everywhere including the axis.

Exercises for 4.2 Statement and physical meaning

Warm-up questions Read Section 4.2, then answer before class

W4.3 Suppose that the field of a point charge fell off as $1/r^3$ instead of $1/r^2$. Would the flux through a sphere centered on the charge be the same for every radius? Would Gauss's law (4.3) still hold?

W4.4 Section 4.2.3 says that the two forms of Gauss's law "answer opposite questions." In your own words, what question does each form answer? Describe a situation in which you would use each one.

W4.5 The flux (4.4) through an open surface, such as a flat disk, has a sign that depends on a choice you have to make. What is that choice? Why does Gauss's law (4.3) leave no such choice open?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P4.3 A point charge Q sits at the center of the square base of a pyramid, which has four identical triangular sides (Figure 4.18). Without doing any integrals, find the electric flux through one of the triangular sides. Does your answer depend on the height of the pyramid? (Hint: what is the flux through the base?)

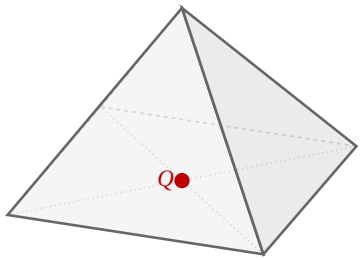


Figure 4.18 The pyramid and the point charge of Exercise P4.3.

P4.4 A point charge q is inside a cube whose faces are at $x, y, z = \pm a$ (Figure 4.19), at the off-center point $(0.3a, -0.5a, 0.2a)$.

(a) Calculate the flux through each of the six faces numerically, by integrating $\mathbf{E} \cdot d\mathbf{a}$ over each face (with Mathematica's `NIntegrate` or Python's `scipy.integrate.dblquad`; see the tips in Section 3.6). Express your answers as multiples of q/ε_0. Which face has the largest flux, and why does that make sense?

(b) Check that the six fluxes add up to q/ε_0.

(c) As a check of your code, put the charge at the center of the cube, where each face must have one sixth of the total flux.

(d) Move the charge to $(1.5a, 0, 0)$, outside the cube. Find the six fluxes again, and show that they now add up to zero. Which faces have negative flux?

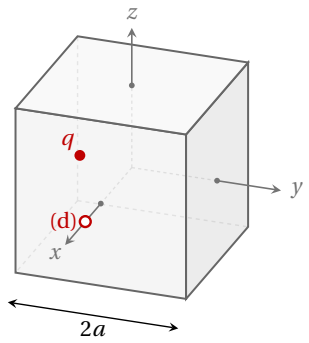


Figure 4.19 The cube and the point charge of Exercise P4.4, with the point of part (d) (open circle).

Exercises for 4.3 Solving for E when given ρ: Gauss's law method (high symmetry)

Warm-up questions Read Section 4.3, then answer before class

W4.6 Example 4.3 shows in full why the field of an infinite line has neither a ϕ component nor a z component. Make the corresponding argument for the infinite plane of Example 4.4: why can its field have no component parallel to the plane, and why can its magnitude depend only on the distance from the plane?

W4.7 Which of these could you solve with the Gauss's law method, and with what Gaussian surface? Give a one-sentence reason for each. (a) An infinitely long cylinder whose charge density depends on both s and ϕ. (b) An infinite slab, $-a < z < a$, whose charge density depends only on z and is the same at z and $-z$. (c) A uniformly charged spherical shell with a point charge inside it, away from its center.

W4.8 In Example 4.5, a classmate finds the charge inside the Gaussian sphere of radius r as $\rho(r)$ times the sphere's volume, $\frac{4}{3}\pi r^3$, instead of integrating. When is that shortcut right? For the ball of Example 4.5, is the classmate's enclosed charge too large or too small, and why?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

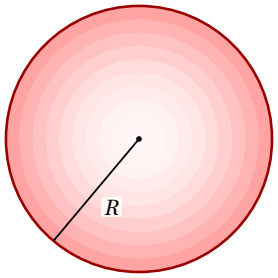


Figure 4.20 The ball of Exercise P4.5, in cross section. The shading darkens where the charge density is larger, drawn for $n = 2$.

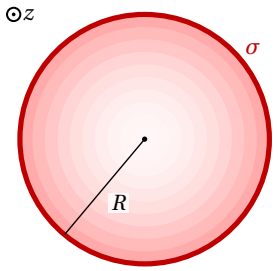


Figure 4.21 The cylinder of Exercise P4.6, end on. The shading darkens where ρ is larger; the heavy circle is the surface charge.

P4.5 A ball of radius R has the charge density $\rho(r) = \rho_0 (r/R)^n$, where ρ_0 is a constant and $n > -3$ (Figure 4.20).

(a) Use Gauss's law to find the electric field inside and outside the ball.

(b) Check that your answer agrees with Example 4.5 for $n = 2$, and that it is continuous at $r = R$.

(c) For a uniformly charged ball ($n = 0$), write the field inside and outside in terms of the total charge Q of the ball, and make a rough sketch of E versus r.

P4.6 An infinitely long cylinder of radius R, with its axis along the z axis, has the volume charge density $\rho = C_1 s^2$ inside it, and its surface carries the surface charge density $\sigma = C_2$, where C_1 and C_2 are constants (Figure 4.21). Use Gauss's law to find the electric field inside and outside the cylinder. Check that the field jumps at $s = R$ by the amount (4.6) predicts.

P4.7 In and around a ball of radius R, the electric field is

$$\mathbf{E} = \begin{cases} \dfrac{\rho_0 R}{\varepsilon_0} e^{-r/R} \hat{\mathbf{r}}, & r < R, \\ \dfrac{\rho_0 R^3}{e \varepsilon_0 r^2} \hat{\mathbf{r}}, & r > R, \end{cases}$$

where ρ_0 is a constant and e is the base of natural logarithms.

(a) What volume charge density produces this field, inside and outside the ball?

(b) Is there a surface charge on the surface of the ball? Justify your answer.

P4.8 A spherical shell of radius a carries a constant surface charge density σ_1 (Figure 4.22). Its center is a distance $d > a$ above an infinite plane that carries a constant surface charge density σ_2. Put the origin at the center of the shell, with the z axis perpendicular to the plane, so that the plane is $z = -d$. Find the electric field, in Cartesian components, at an arbitrary point (x, y, z) in each of three regions: inside the shell; outside the shell and above the plane; and below the plane.

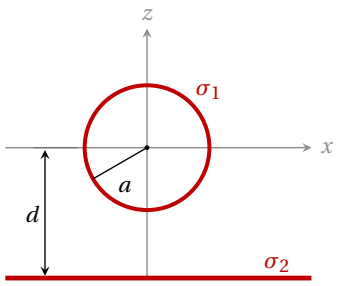


Figure 4.22 The shell and the plane of Exercise P4.8, in the xz plane.

P4.9 Gauss's law is true for every charge distribution, but it cannot be used to find the field of (a) a uniformly charged cube, (b) a uniformly charged line of finite length, or (c) a uniformly charged disk. For each, sketch the field lines you expect near the object, and use your sketch to explain which of the requirements of Section 4.3 fails, for any closed surface you might choose.

Exercises for 4.4 Boundary condition on $E^{\perp}$

Warm-up questions Read Section 4.4, then answer before class

W4.9 The pillbox argument for (4.6) needs no symmetry, while the Gauss's law method of Section 4.3 needs a great deal of it. Why can the pillbox argument do without symmetry? What does it tell you about the field, and what does it not tell you?

W4.10 Equation (4.6) depends on which side of the surface you call "above." Suppose you choose the other side instead, so that $\hat{\mathbf{n}}$ points the other way. What happens to each term of the equation? Does the jump change?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

Exercises for 4.5 Conductors and induced charges

Warm-up questions Read Section 4.5, then answer before class

W4.11 A neutral, solid conducting sphere is placed in a uniform external field. Which of these statements are true once the charges have stopped moving? Give a one-sentence reason for each. (a) The field inside the metal is zero. (b) There is no charge anywhere on the sphere. (c) The field just outside the sphere is perpendicular to its surface. (d) Just outside the sphere, the field is $(\sigma/\varepsilon_0)\,\hat{\mathbf{n}}$, where σ is the local surface charge density and $\hat{\mathbf{n}}$ the outward normal.

W4.12 The argument that the field inside a conductor is zero is that otherwise its free charges would move. Why does that argument not show, by itself, that the field is also zero inside an empty cavity in the conductor? What does Section 4.5.1 use instead?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P4.10 Section 4.5.2 states that the force per unit area on the surface of a conductor is $(\sigma^2/2\varepsilon_0)\,\hat{\mathbf{n}}$. This exercise derives it. Consider a small patch of a surface that carries the surface charge density σ (not necessarily on a conductor), with a unit normal $\hat{\mathbf{n}}$ pointing "above", and split the field near the patch into the field of the patch itself and the field $\mathbf{E}_{\text{other}}$ of all the other charges.

(a) Explain why the force on the patch comes from $\mathbf{E}_{\text{other}}$ alone, and why $\mathbf{E}_{\text{other}}$ is the same just above and just below the patch.

(b) Very close to the patch, much closer than its size, the patch looks like an infinite plane. Use (4.5) to write the fields $\mathbf{E}_{\text{above}}$ and $\mathbf{E}_{\text{below}}$ in terms of $\mathbf{E}_{\text{other}}$, σ, and $\hat{\mathbf{n}}$. Show that $\mathbf{E}_{\text{other}}$ is the average of $\mathbf{E}_{\text{above}}$ and $\mathbf{E}_{\text{below}}$, so that the force per unit area on any surface charge is $\sigma\,\frac{1}{2}(\mathbf{E}_{\text{above}}+\mathbf{E}_{\text{below}})$.

(c) For the surface of a conductor, with $\hat{\mathbf{n}}$ the outward normal, use the results of Section 4.5 to show that the force per unit area is $(\sigma^2/2\varepsilon_0)\,\hat{\mathbf{n}} = (\varepsilon_0/2)E^2\,\hat{\mathbf{n}}$, and that it points outward whatever the sign of σ.

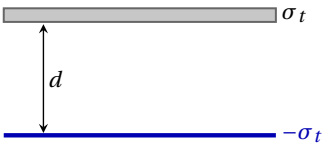


Figure 4.23 The conducting plate (gray) and the insulating sheet of part (b) of Exercise P4.11, edge on.

P4.11 A large, thin, isolated conducting plate carries a total charge per unit area σ_t (Figure 4.23). Ignore effects near its edges.

(a) Where does the charge go? Find the field on each side of the plate, and compare it with the field of an insulating sheet carrying σ_t.

(b) The plate is now placed a distance d above a large, parallel insulating sheet that carries the charge per unit area $-\sigma_t$. Find the charge per unit area on each face of the plate, and the field above the plate, between the plate and the sheet, and below the sheet.

(c) Check that, just outside each face of the plate, your field in part (b) is σ/ε_0, where σ is the charge per unit area on that face.



Figure 4.24 The two plates of Exercise P4.12, edge on.

P4.12 Two large, identical, parallel conducting plates (Figure 4.24), each of area A, carry charges Q_1 (the upper plate) and Q_2 (the lower plate). Ignore effects near the edges.

(a) Following Example 4.7, show that the two outer faces carry equal charges, $(Q_1 + Q_2)/2$ each, and the two inner faces carry equal and opposite charges.

(b) Find the charge on each face, and the field above, between, and below the plates, for $Q_1 = Q$, $Q_2 = 0$.

(c) Repeat for $Q_1 = Q$, $Q_2 = -Q$, and for $Q_1 = Q$, $Q_2 = -2Q$.

P4.13 A neutral conducting spherical shell has inner radius a and outer radius b (Figure 4.25). A point charge q is placed inside the cavity, a distance $a/2$ from the center.

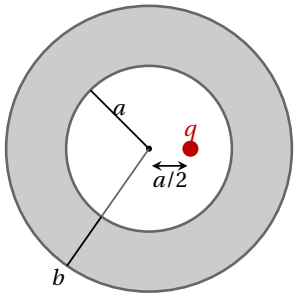


Figure 4.25 The conducting shell and the point charge of Exercise P4.13, in cross section.

(a) What is the total charge on the inner surface of the shell? On the outer surface?

(b) Is the surface charge density uniform on the inner surface? On the outer surface? Explain.

(c) Find the electric field outside the shell ($r > b$) and inside the metal ($a < r < b$).

(d) If the charge is moved to a different point in the cavity, which of your answers change?

P4.14 On a clear day, there is an electric field of about 100 V/m near the surface of the Earth, pointing down. Treat the Earth as a conducting sphere of radius 6.37×10^6 m.

(a) What is the surface charge density on the Earth's surface? What is the total charge of the Earth implied by this field?

(b) What is the outward stress (4.7) on the Earth's surface due to this charge? Compare it with atmospheric pressure.

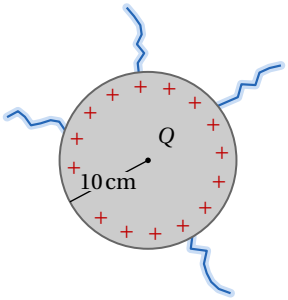


Figure 4.26 The charged conducting sphere of Exercise P4.15. When the field just outside it exceeds about 3×10^6 V/m, the air breaks down and sparks jump from the surface (part b).

P4.15 A conducting sphere of radius 10 cm carries a charge Q (Figure 4.26).

(a) How large would Q have to be for the outward electric stress on its surface to equal atmospheric pressure, 1.01×10^5 Pa? What would the field just outside the sphere be?

(b) Air breaks down, and sparks form, when the field exceeds about 3×10^6 V/m. What is the largest stress that a charged conductor in air can experience? Is electrostatic stress ever likely to damage a metal object in air?

Exercises for 4.6 Summary

P4.16 True or false? Explain your reasoning.

(a) If the electric field has no divergence anywhere inside a region of space, then there is no charge in that region.

(b) If the total charge inside a closed surface is zero, then the electric field is zero everywhere on the surface.

(c) A point charge sits at rest at (x_0, y_0, z_0), and there are no other charges anywhere. Then $\nabla \cdot \mathbf{E} = 0$ everywhere.

(d) The flux of a point charge's field through a sphere is the same whether the charge is just inside the sphere or just outside it.

(e) Gauss's law contains all of the physical information in Coulomb's law.

P4.17 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as the flux of a point charge's field through a closed surface of irregular shape, the divergence of a field calculated numerically from its values on a grid, field lines drawn so that the number leaving each charge is proportional to its charge, the field near the edges and corners of a uniformly charged cube, the field inside an off-center spherical hole in a uniformly charged ball, or the charge induced on a neutral conductor by a nearby point charge.

Chapter 5

Electric Potential and Potential Energy

Chapter 4 found the divergence of the electrostatic field from Coulomb's law. This chapter finds its curl. The curl turns out to be zero, and that one fact has a long list of consequences. The most important is that the field can be written as the negative of the gradient of a scalar function, the **electric potential**. Because the potential is a scalar, it is usually much easier to calculate than the field: there is one integral to do instead of three, and no unit vectors to keep track of. Once the potential is known, the field follows from it by taking derivatives. The potential is also the bridge between electrostatics and energy. It gives the work needed to move a charge from one place to another, and through that, the energy needed to assemble a distribution of charge, energy that turns out to be stored in the field itself. Along the way, the chapter completes the boundary conditions begun in Chapter 4, treats conductors and capacitors, and arrives at Poisson's equation, the starting point of Chapter 7.

5.1 Curl of Coulomb's law

Start again with a single point charge q, this time at the origin, so that its field is

$$\mathbf{E} = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}\hat{\mathbf{r}}.$$

Does this field have any curl? Its field lines point straight out from the charge, and nothing about them swirls, so the answer should be no: a small paddle wheel placed anywhere in this field (Section 1.4.3) would not turn. The formula for the curl in spherical coordinates agrees. Every one of its terms contains E_θ, E_ϕ, or a derivative of E_r with respect to θ or ϕ, and all of these are zero for a field that points along $\hat{\mathbf{r}}$ and depends only on r. So $\nabla \times \mathbf{E} = 0$ everywhere except, possibly, at the charge itself, where the field is infinite and the formula does not apply.

That "possibly" matters. The divergence of this same field (Section 4.1) is also zero everywhere except at the charge, and at the charge it is a delta function. As

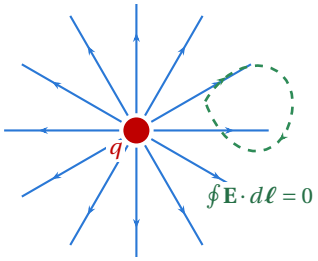


Figure 5.1 The field lines of a point charge q and a closed loop (dashed). Only the part of each step around the loop that moves toward or away from the charge contributes to $\mathbf{E}\cdot d\boldsymbol{\ell}$, and around the loop those contributions cancel.

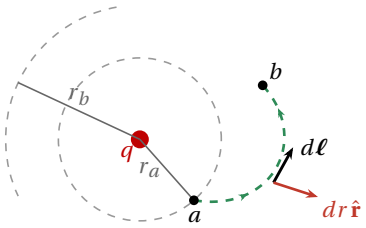


Figure 5.2 A path (dashed) from a to b near a point charge. Of each step $d\boldsymbol{\ell}$, only the radial part $dr\,\hat{\mathbf{r}}$ contributes to $\mathbf{E}\cdot d\boldsymbol{\ell}$, so the integral depends only on r_a and r_b.

Section 1.6.3 showed, a point where a field is infinite has to be checked separately, and for a curl the check is a circulation, a line integral around a closed loop. Take any path from a point **a** to a point **b**. With $d\boldsymbol{\ell}$ written in spherical coordinates, (1.12), each step along the path contributes

$$\mathbf{E}\cdot d\boldsymbol{\ell} = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}\hat{\mathbf{r}}\cdot\left(dr\,\hat{\mathbf{r}} + r\,d\theta\,\hat{\boldsymbol{\theta}} + r\sin\theta\,d\phi\,\hat{\boldsymbol{\phi}}\right) = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}\,dr.$$

Only the part of each step that moves directly away from the charge, or toward it, contributes (Figure 5.2). So

$$\int_{\mathbf{a}}^{\mathbf{b}} \mathbf{E}\cdot d\boldsymbol{\ell} = \frac{q}{4\pi\varepsilon_0}\int_{r_a}^{r_b}\frac{dr}{r^2} = \frac{q}{4\pi\varepsilon_0}\left(\frac{1}{r_a} - \frac{1}{r_b}\right), \tag{5.1}$$

where r_a and r_b are the distances of the two endpoints from the charge. The answer depends only on where the path starts and ends, not on the route between them. Around a closed loop the path ends where it starts, so $r_b = r_a$ and the integral is zero. That is true for every loop that does not pass through the charge itself, however closely it approaches the charge, so there is no delta function hiding in this curl. The curl of a point charge's field is zero everywhere.

Now use superposition. Any static distribution of charge can be divided into point charges, or into infinitesimal pieces dq, and its field is the sum of their fields. The curl is made of derivatives, so the curl of a sum is the sum of the curls:

$$\begin{aligned}\nabla\times\mathbf{E} &= \nabla\times(\mathbf{E}_1 + \mathbf{E}_2 + \cdots)\\ &= \nabla\times\mathbf{E}_1 + \nabla\times\mathbf{E}_2 + \cdots\\ &= 0.\end{aligned}$$

For a continuous distribution the same argument applies to the integral (3.13). The curl is taken with respect to the field point **r**, which is not a variable of integration, so it can be moved inside the integral, where it acts on the field of each piece. Therefore, for any static distribution of charge,

$$\nabla\times\mathbf{E} = 0. \qquad \text{(electrostatics)} \tag{5.2}$$

This is not obvious from a picture of the field. The field lines of two charges, for example, curve from one charge to the other (Figure 3.6c), and it is not easy to see by looking that the field has no swirl anywhere. But it does not, because the curl of each charge's field is zero.

Equation (5.2) is one of Maxwell's equations (Section 2.1), in the form it takes when nothing changes in time. The full equation is Faraday's law, (2.3), which says that a magnetic field that changes in time produces an electric field with a curl. Chapter 15 takes that up; until then, all charges are at rest, all fields are static, and (5.2) is exact. (The electric field produced by a changing magnetic field has curl but no divergence, so Gauss's law is unaffected.) Together, (5.2) and Gauss's law, $\nabla\cdot\mathbf{E} = \rho/\varepsilon_0$, give both the divergence and the curl of the electrostatic field. By the Helmholtz theorem (Section 2.3), that is enough to determine the field of

any charge distribution confined to a finite region. The two equations together therefore contain everything in Coulomb's law and superposition, which Gauss's law alone does not.

Like Gauss's law, (5.2) has an integral form. Integrate it over any surface $\mathscr{S}$, and use the curl theorem (1.37) to turn the surface integral of the curl into a line integral around the boundary of the surface:

$$\int_{\mathscr{S}} (\nabla \times \mathbf{E}) \cdot d\mathbf{a} = \oint \mathbf{E} \cdot d\boldsymbol{\ell} = 0.$$

Every closed loop is the boundary of some surface, so

$$\oint \mathbf{E} \cdot d\boldsymbol{\ell} = 0 \tag{5.3}$$

for every closed loop. Gauss's law in integral form is about the flux out of a closed surface; this is about the circulation around a closed loop. For a point charge it is exactly what (5.1) showed directly.

One consequence is easy to see. Along a field line, $\mathbf{E}$ points along the line, so if the line is followed in the direction of the field, $\mathbf{E} \cdot d\boldsymbol{\ell}$ is positive at every step. If a field line closed on itself, the line integral around it would therefore be positive, not zero. So the field lines of static charges never form closed loops. Each one begins on a positive charge or comes in from infinity, and ends on a negative charge or goes off to infinity. (The only exceptions are the few lines that run into or out of a null point, a point where the field is zero, as in Figure 3.6.)

5.2 Boundary condition on $E^{\parallel}$

Section 4.4 used Gauss's law and a short pillbox to show that the component of $\mathbf{E}$ perpendicular to a surface charge jumps by σ/ε_0. The curl of $\mathbf{E}$ does the same job for the components *parallel* to the surface. Draw a rectangular loop that straddles the surface, with two sides of length ℓ parallel to the surface, one just above it and one just below, and two short sides that cross it (Figure 5.3). This is the same sort of loop that we will use again with Ampère's law in Chapter 11, where it is called an "Amperian loop". By (5.3), the line integral of $\mathbf{E}$ around this loop, as around any closed loop, is zero. Now make the loop very short. The contributions of the two short sides go to zero (as long as the field is finite). Going around the loop counterclockwise, as in the figure, the lower side contributes $E^{\parallel}_{\text{below}}\,\ell$ and the upper side, which is traversed in the opposite direction, contributes $-E^{\parallel}_{\text{above}}\,\ell$, where $E^{\parallel}$ is the component of $\mathbf{E}$ along the lower side. So $E^{\parallel}_{\text{below}}\,\ell - E^{\parallel}_{\text{above}}\,\ell = 0$. The loop can be turned to lie along any direction in the surface, so every component of $\mathbf{E}$ parallel to the surface is continuous,

$$\mathbf{E}^{\parallel}_{\text{above}} = \mathbf{E}^{\parallel}_{\text{below}}. \tag{5.4}$$

With words, the parallel components of the field are the same on both sides of a surface, whatever charge the surface carries. A jump in them would be a curl

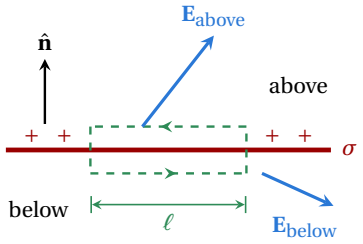


Figure 5.3 A short, squat loop straddling a surface charge σ. Only the components of the fields parallel to the long sides contribute to the line integral around it, and they must be equal above and below.

concentrated on the surface (Exercise P1.21), and the electrostatic field has no curl. Combined with (4.6), the two boundary conditions can be written as one vector equation,

$$\mathbf{E}_{\text{above}} - \mathbf{E}_{\text{below}} = \frac{\sigma}{\epsilon_0}\hat{\mathbf{n}}. \tag{5.5}$$

The potential, which the next section introduces, is continuous across the surface as well (Section 5.3.4). That completes the set of boundary conditions at a surface charge that Chapter 7 uses.

5.3 Scalar potential and its relation to electric field

A field with zero curl has several special properties, and the most useful of them is that it can be written as the gradient of a scalar function. This section collects those properties and uses them to define the electric potential.

5.3.1 Conservative fields

The following four statements about a vector field **E** are equivalent: if any one of them is true, all four are.[1]

1. $\nabla \times \mathbf{E} = 0$ everywhere.
2. $\oint \mathbf{E} \cdot d\boldsymbol{\ell} = 0$ around every closed loop.
3. $\int_{\mathbf{a}}^{\mathbf{b}} \mathbf{E} \cdot d\boldsymbol{\ell}$ is the same for every path from **a** to **b**.
4. $\mathbf{E} = -\nabla V$ for some scalar function V.

A field with these properties is called **conservative**, the name it has in mechanics, where a conservative force is one whose work does not depend on the path. It is also called **irrotational** or **curl-free**. Section 5.1 showed that every electrostatic field has property 1, so every electrostatic field has all four.

Each statement implies the next, and the last implies the first. Statement 1 implies statement 2 by the curl theorem, as in (5.3). Statement 2 implies statement 3. Take two different paths between the same two points, such as paths 1 and 2 from $\mathcal{O}$ to P in Figure 5.4, and make a closed loop by going out along path 1 and back along path 2. Reversing the direction of a path changes the sign of its line integral, so

$$0 = \oint \mathbf{E} \cdot d\boldsymbol{\ell} = \int_{\text{path 1}} \mathbf{E} \cdot d\boldsymbol{\ell} - \int_{\text{path 2}} \mathbf{E} \cdot d\boldsymbol{\ell},$$

and the two integrals are equal. Statement 3 implies statement 4. Pick any fixed

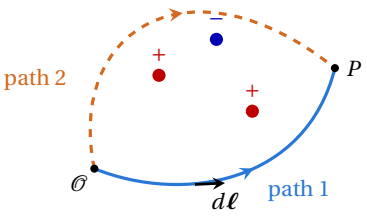


Figure 5.4 Two paths from the point $\mathcal{O}$ to the point P, in the field of a few charges. Because $\nabla \times \mathbf{E} = 0$, the line integral of **E** is the same along both. Going out along path 1 and back along path 2 makes a closed loop, around which the line integral is zero.

[1] The proof that statement 1 implies statement 2 needs every closed loop to be the boundary of a surface on which the curl is zero. That is automatic when the curl is zero everywhere in space, as it is for an electrostatic field, but it can fail when the curl is known to be zero only in a region with a hole through it. The field $\hat{\boldsymbol{\phi}}/s$ of (1.47) has zero curl everywhere except on the z axis, but its circulation around the axis is 2π. In the region outside the axis it satisfies statement 1 but not statement 2.

point $\mathcal{O}$, and define

$$V(\mathbf{r}) = -\int_{\mathcal{O}}^{\mathbf{r}} \mathbf{E}\cdot d\boldsymbol{\ell}.$$

Because of statement 3, this gives a single value at each point $\mathbf{r}$, with no need to say which path the integral follows. Extending the path by a small step $d\boldsymbol{\ell}$ beyond $\mathbf{r}$ adds one more term to the integral, so the change in V in that step is $dV = -\mathbf{E}\cdot d\boldsymbol{\ell}$. But by (1.19), the change in a function in a small step is $dV = \nabla V \cdot d\boldsymbol{\ell}$. The two agree for every direction of the step only if $\mathbf{E} = -\nabla V$. Finally, statement 4 implies statement 1, because the curl of a gradient field is always zero, (1.56).

So every electrostatic field can be written as

$$\mathbf{E} = -\nabla V, \tag{5.6}$$

where V is the **electric potential**. With words, the electric field points in the direction in which the potential decreases fastest, and its magnitude is the rate of that decrease. If you picture V as the height of a landscape, $\mathbf{E}$ points straight downhill, and it is strongest where the landscape is steepest.

The minus sign is a convention, chosen to match mechanics, where a conservative force is $\mathbf{F} = -\nabla U$ and pushes an object toward lower potential energy. With it, the field pushes a positive charge from high potential toward low potential, the way gravity pulls a ball downhill and the way heat flows from high temperature to low, along $-\nabla T$. Equation (5.6) holds only in electrostatics. When magnetic fields change in time, $\mathbf{E}$ has a curl and cannot be the gradient of anything, and Section 15.4 adds a second term to the equation.

The idea that a force is the gradient of a single function goes back to Lagrange, who used such a function for gravitation in the 1770s. The name came later. George Green, a largely self-taught miller from Nottingham, called it the "potential function" in *An Essay on the Application of Mathematical Analysis to the Theories of Electricity and Magnetism*, which he had printed privately in 1828. Gauss, apparently unaware of Green's work, called it simply "das Potential" in 1840.

The relation can be turned around. Integrate (5.6) along any path from $\mathbf{a}$ to $\mathbf{b}$, and use the gradient theorem (1.34):

$$V(\mathbf{b}) - V(\mathbf{a}) = -\int_{\mathbf{a}}^{\mathbf{b}} \mathbf{E}\cdot d\boldsymbol{\ell}. \tag{5.7}$$

With words, the potential difference between two points is the negative of the line integral of the field from one to the other, along any path. This equation can serve as the definition of the potential: any function V that satisfies it for every pair of points is a legitimate potential for the field $\mathbf{E}$. Path independence is remarkable when you think about what it says. In Figure 5.4, the two paths pass the charges on opposite sides, through regions where the field points in quite different directions. The two integrals add up entirely different contributions along the way, and yet they give the same total.



George Green (1793–1841, English) No portrait of Green is known. The picture shows the windmill in Sneinton, Nottingham, that his father, a baker, built in 1807; Green ran it after his father's death in 1829, and it was restored in the 1980s. He had four terms of school, at the ages of eight and nine, and taught himself mathematics. In 1828 he had his *Essay on the Application of Mathematical Analysis to the Theories of Electricity and Magnetism* printed for 51 subscribers. One of them, Sir Edward Bromhead, encouraged him to go to Cambridge, and he entered Caius College in 1833, at forty, and graduated in 1837. He died four years later, almost unknown. In 1845 William Thomson obtained copies of the *Essay* and had it reprinted; its integral theorem and the functions now named after Green are used throughout physics. (Wikipedia)

5.3.2 The reference point

Equation (5.7) gives only differences of potential. Adding a constant to V changes nothing physical, since the gradient of a constant is zero, just as in mechanics the zero of potential energy can be put wherever it is convenient. To give the potential a definite value at each point, choose a **reference point** $\mathcal{O}$ at which $V = 0$. Then (5.7) with $\mathbf{a} = \mathcal{O}$ gives

$$V(\mathbf{r}) = -\int_{\mathcal{O}}^{\mathbf{r}} \mathbf{E} \cdot d\boldsymbol{\ell}, \tag{5.8}$$

along any path from $\mathcal{O}$ to $\mathbf{r}$. Any convenient path will do; straight lines, and paths made of straight segments along coordinate directions, are usually the easiest. Notice what the variables mean here. The point $\mathbf{r}$ is the end of the path, where the potential is wanted. The variable of integration is a point that moves along the path from $\mathcal{O}$ to $\mathbf{r}$. It is not a source point, and to keep it distinct both from $\mathbf{r}$ and from the source points $\mathbf{r}'$ of Chapter 3, this book writes it with a bar, as in $\bar{r}$.

For a charge distribution confined to a finite region, the natural reference point is infinitely far away, where the distribution's field goes to zero: $V(\infty) = 0$. This is the convention used throughout the book unless something else is said. It fails, though, for idealized distributions that contain an infinite amount of charge, such as the infinite line and the infinite plane of Examples 4.3 and 4.4. For them the integral from infinity diverges, and the reference point has to be somewhere else.

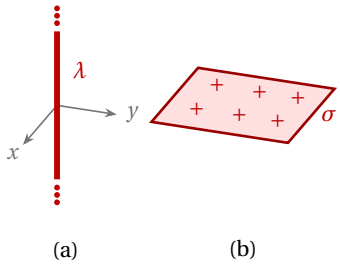
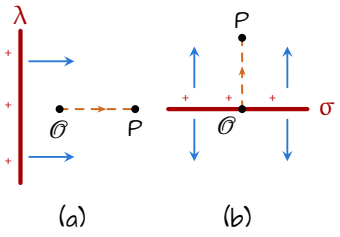


Figure 5.5 The charge distributions of Example 5.1: (a) the infinite line along the z axis and (b) the infinite plane.

Example 5.1

Find the potential of (a) an infinite line along the z axis with uniform line charge density λ, and (b) an infinite plane, the xy plane, with uniform surface charge density σ.

Solution:



We start with a sketch of each charge distribution, its field, and a path along which to integrate the field, from a reference point $\mathcal{O}$ to the field point P: (a) a radial line out from the line of charge, and (b) a line perpendicular to the plane. Where to put $\mathcal{O}$ is the question in both parts.

(a) From Example 4.3, $\mathbf{E} = (\lambda/2\pi\varepsilon_0 s)\,\hat{\mathbf{s}}$. Try putting the reference point at infinity, and integrate along a radial line, on which $d\boldsymbol{\ell} = d\bar{s}\,\hat{\mathbf{s}}$:

$$V(s) = -\int_{\infty}^{s} \frac{\lambda}{2\pi\varepsilon_0 \bar{s}}\, d\bar{s} = -\frac{\lambda}{2\pi\varepsilon_0}\Big[\ln \bar{s}\Big]_{\infty}^{s}.$$

The logarithm of infinity is infinite, so this fails. The field falls off as $1/s$, and that is not fast enough: it would take an infinite amount of work to bring a charge in from infinity.

Choose instead a reference point at some finite distance s_0 from the line (not on the line itself, where the potential is also infinite). Then

$$V(s) = -\int_{s_0}^{s} \frac{\lambda}{2\pi\varepsilon_0 \bar{s}}\, d\bar{s} = -\frac{\lambda}{2\pi\varepsilon_0} \ln\frac{s}{s_0}.$$

For $\lambda > 0$ the potential decreases as you move away from the line, as it should, since the field points away from the line. A different choice of s_0 only adds a constant to V, since $-\ln(s/s_0) = -\ln s + \ln s_0$, and it does not change $\mathbf{E}$.

(b) From (4.5), the field has magnitude $\sigma/2\varepsilon_0$ on both sides of the plane, pointing away from it. It does not fall off at all, and with the reference point at infinity the integral is again infinite. Put the reference point on the plane instead, and integrate along the z axis, with $d\boldsymbol{\ell} = d\bar{z}\hat{\mathbf{z}}$. (Why can the reference point be on the plane, when in (a) it could not be on the line? Answer in footnote.[2]) Above the plane, $E_z = +\sigma/2\varepsilon_0$, and

$$V(z) = -\int_0^z \frac{\sigma}{2\varepsilon_0}\, d\bar{z} = -\frac{\sigma z}{2\varepsilon_0}, \qquad z > 0.$$

Below it, $E_z = -\sigma/2\varepsilon_0$, and

$$V(z) = -\int_0^z \left(-\frac{\sigma}{2\varepsilon_0}\right) d\bar{z} = \frac{\sigma z}{2\varepsilon_0}, \qquad z < 0.$$

Both are $V = -\sigma|z|/2\varepsilon_0$. The potential falls off linearly on both sides of the plane, like a roof with its ridge along the plane.

Reality check: Taking $\mathbf{E} = -\nabla V$ gives back the fields we started from. In (a), $-\partial V/\partial s = \lambda/2\pi\varepsilon_0 s$. In (b), $-dV/dz = \sigma/2\varepsilon_0$ above the plane and $-\sigma/2\varepsilon_0$ below it.

The example shows what goes wrong. A reference point at infinity works only if the field falls off faster than 1/distance far away; otherwise the integral from infinity diverges, even though the field itself goes to zero. The field of any charge distribution confined to a finite region falls off as $1/r^2$ or faster, so for real charge distributions infinity is always a good reference point. The infinite line and plane are idealizations, and the trouble with them comes from the infinite amount of charge they contain.

The unit of potential is the **volt**, $1\,\mathrm{V} = 1\,\mathrm{J/C}$. By (5.7), potential is field times distance, and a field in newtons per coulomb times a distance in meters is an energy per charge, in joules per coulomb. For the same reason the unit of electric field, the newton per coulomb, is also the volt per meter, $1\,\mathrm{N/C} = 1\,\mathrm{V/m}$, which is how fields are usually quoted. The potential is also called the **voltage**, especially in circuits, where it is differences in potential that are measured.

[2] Close to the plane the field is still only $\sigma/2\varepsilon_0$, so the integral from the plane to any z is finite. Close to the line the field grows as $1/s$, and its integral, a logarithm, is infinite at $s = 0$.

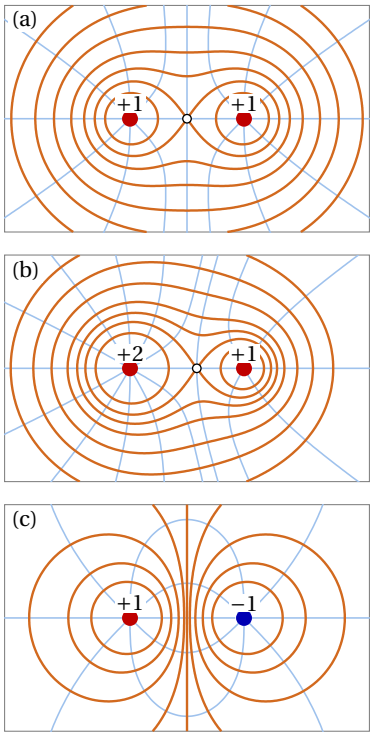


Figure 5.6 Equipotentials (orange) of the three pairs of charges of Figure 3.6, with the field lines from that figure drawn faintly (blue): (a) two equal positive charges, (b) charges +2 and +1, (c) charges +1 and −1. The field lines cross the equipotentials at right angles. At the null points (open circles), an equipotential crosses itself.

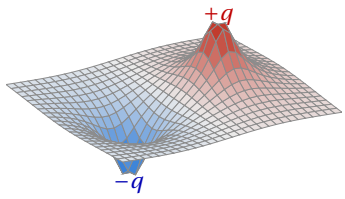


Figure 5.7 The potential of a positive and a negative point charge, in the plane that contains them, drawn as a landscape: a mountain at $+q$ and a pit at $-q$ (both cut off at the top and bottom, since V is infinite at the charges).

5.3.3 Why the potential is useful

Two things about the potential may seem surprising. The first is that a single scalar function can carry all of the information in the three components of **E**. That is possible only because the electrostatic field is a special kind of vector field. Its zero curl ties its three components together, through relations such as $\partial E_y/\partial x = \partial E_x/\partial y$ (the z component of (1.22)), so that they are not independent of each other. Most vector fields cannot be written as the gradient of anything.

The second is that the potential is often more useful than the field, for three reasons. It is a scalar, so there are no components or unit vectors to keep track of. Getting **E** from V takes only derivatives, and derivatives are almost always easier than integrals: a derivative may be messy, but it can always be done. The potential is also directly connected to energy (Section 5.4.1); potential differences, rather than fields, are what a voltmeter measures.

Take the landscape picture of V literally. A topographic map shows the height of the land with contour lines, each connecting points at the same height. The corresponding surfaces for the potential, on which V has a single value, are called **equipotential surfaces**, or **equipotentials**; in a two-dimensional drawing they appear as lines (Figure 5.6). The field is perpendicular to them everywhere. To see why, take a small step $d\boldsymbol{\ell}$ within an equipotential surface. The potential does not change along the step, so by (1.19) and (5.6),

$$\begin{aligned} dV &= \nabla V \cdot d\boldsymbol{\ell} \\ &= -\mathbf{E} \cdot d\boldsymbol{\ell} \\ &= 0, \end{aligned}$$

which says that **E** is perpendicular to every direction within the surface. The field lines therefore cross the equipotentials at right angles, pointing from higher potential toward lower. Where equipotentials drawn at equal intervals of V are crowded together, the field is strong, just as closely spaced contour lines on a map mean a steep slope. Near a positive charge the potential rises to a peak, like a mountain, and near a negative charge it falls into a pit (Figure 5.7). For a single point charge, the equipotentials are spheres centered on the charge. Can two different equipotentials cross? No, since V has only one value at each point. But a single equipotential can cross itself, at a null point, where $\mathbf{E} = 0$ and there is no field for it to be perpendicular to. That happens in Figure 5.6(a) and (b).

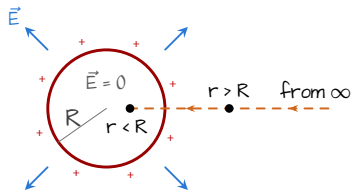
5.3.4 Continuity of the potential

Sections 4.4 and 5.2 found how the field behaves at a surface charge: its perpendicular component jumps by σ/ε_0, and its parallel components are continuous. The potential behaves more simply still, as the next example shows.

Example 5.2

A thin spherical shell of radius R carries a total charge Q, spread uniformly over its surface. Find the potential everywhere, with $V(\infty) = 0$.

Solution:



First, a picture of what we know: the shell, with no field inside and a radial field outside, and the path we will integrate along. The field was found in Example 4.2: it is zero inside the shell, and $Q/4\pi\varepsilon_0 r^2$ outside, pointing along $\hat{\mathbf{r}}$. Since the field is known, the quickest route is to integrate it with (5.8), along a radial path from the reference point at infinity, as drawn.

On that path $d\boldsymbol{\ell} = d\bar{r}\,\hat{\mathbf{r}}$, so $\mathbf{E}\cdot d\boldsymbol{\ell} = E(\bar{r})\,d\bar{r}$. Outside the shell ($r > R$),

$$\begin{aligned} V(r) &= -\int_\infty^r \frac{Q}{4\pi\varepsilon_0\,\bar{r}^2}\,d\bar{r} \\ &= \frac{Q}{4\pi\varepsilon_0}\left[\frac{1}{\bar{r}}\right]_\infty^r \\ &= \frac{Q}{4\pi\varepsilon_0\,r}. \end{aligned}$$

Inside the shell ($r < R$), the path from infinity crosses the shell, and the field has different forms on the two parts of the path, so the integral is split in two. It is the same integral as before; only the form of E changes along the way:

$$V(r) = -\int_\infty^R \frac{Q}{4\pi\varepsilon_0\,\bar{r}^2}\,d\bar{r} - \int_R^r 0\,d\bar{r} = \frac{Q}{4\pi\varepsilon_0\,R}.$$

So

$$V(r) = \begin{cases} \dfrac{Q}{4\pi\varepsilon_0\,R}, & r < R, \\[2ex] \dfrac{Q}{4\pi\varepsilon_0\,r}, & r > R. \end{cases}$$

Outside, the potential is that of a point charge Q at the center, as the field is. Inside, the potential is constant and equal to its value on the shell: there is no field inside, so it takes no work to move a charge around in there. But the potential inside is not zero, even though the field is. The potential at a point depends on the field all along the path from the reference point, so to find it inside the shell you need the field outside. That is typical. Fields are often found from the inside out, as with Gauss's law, where the Gaussian surface grows outward and encloses more and more charge. Potentials, with the reference point at infinity, are found from the outside in. The same reference point must be used for the whole problem: it would be wrong to start the inside integral at R as though $V(R)$ were zero.

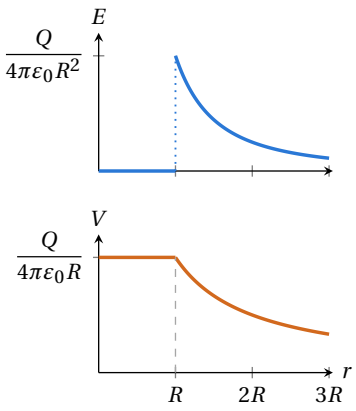


Figure 5.8 The field (top) and the potential (bottom) of a uniformly charged spherical shell of radius R and charge Q, from Example 5.2. The field jumps at the shell, by σ/ε_0. The potential is continuous there, but its slope changes abruptly.

Reality check: The field satisfies the boundary condition (4.6): at $r = R$ it jumps from 0 to $Q/4\pi\varepsilon_0 R^2$, which is σ/ε_0 with $\sigma = Q/4\pi R^2$. The potential does not jump (Figure 5.8).

The potential is continuous at every surface charge, not only this one. Take two points, one just above a surface and one just below it, a distance δ apart, and apply (5.7) along the short straight path between them:

$$V_{\text{above}} - V_{\text{below}} = -\int_{\text{below}}^{\text{above}} \mathbf{E}\cdot d\boldsymbol{\ell}.$$

As long as the field is finite, the integral is no larger than the largest value of $|\mathbf{E}|$ on the path times the length δ of the path, and that goes to zero as $\delta \to 0$. So

$$V_{\text{above}} = V_{\text{below}}. \tag{5.9}$$

With words, the potential has no jumps. If it did, its slope would be infinite at the jump, and by (5.6) so would the field, and the force on a charge there. What the potential can have is a kink. The perpendicular component of the field is the negative of the derivative of V along the normal, so a jump in $E^{\perp}$ at a surface charge is an abrupt change in the slope of V, as in Figure 5.8. (Chapter 7 writes this as (7.2).)

Is the jump in the field real? On the scale of atoms, a layer of surface charge has some thickness, and the field changes rapidly but continuously across it. The jump is an idealization, and an accurate one on any larger scale. One further idealization breaks (5.9). A **dipole layer** is a pair of sheets with charges $+\sigma$ and $-\sigma$ a distance δ apart, in the limit $\delta \to 0$ with the product $\sigma\delta$ held fixed. The field between the sheets, σ/ε_0, then becomes infinite, and the potential does jump across the layer (Exercise P5.4).

5.4 Potential of a point charge, relationship to potential energy

The potential of a point charge q at the origin follows from (5.8) with the reference point at infinity. Integrate along a radial line from infinity to a point a distance r from the charge. On that line $d\boldsymbol{\ell} = d\bar{r}\,\hat{\mathbf{r}}$, so

$$\begin{aligned} V(r) &= -\int_{\infty}^{r} \frac{q}{4\pi\varepsilon_0 \bar{r}^2}\,\hat{\mathbf{r}}\cdot(d\bar{r}\,\hat{\mathbf{r}}) \\ &= -\frac{q}{4\pi\varepsilon_0}\int_{\infty}^{r}\frac{d\bar{r}}{\bar{r}^2} \\ &= \frac{q}{4\pi\varepsilon_0}\left[\frac{1}{\bar{r}}\right]_{\infty}^{r} \\ &= \frac{1}{4\pi\varepsilon_0}\,\frac{q}{r}. \end{aligned}$$

The same result follows from (5.1) with $r_a = \infty$ and $r_b = r$, for any path. For a charge at $\mathbf{r}'$ instead of the origin, the distance r becomes $\mathscr{r} = |\mathbf{r} - \mathbf{r}'|$:

$$V(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\frac{q}{\mathscr{r}}. \tag{5.10}$$

The potential of a point charge falls off as $1/\mathscr{r}$, one power more slowly than its field. It is positive around a positive charge and negative around a negative one.

The path in this calculation runs inward, toward the charge, so it is tempting to write $d\boldsymbol{\ell} = -d\bar{r}\,\hat{\mathbf{r}}$. That gives the wrong sign. In a line integral the direction of travel can be specified either in $d\boldsymbol{\ell}$ or in the limits of integration, but not in both. The rule that avoids the trouble is to write $d\boldsymbol{\ell}$ in the form of (1.12) (or, in Cartesian coordinates, as $dx\hat{\mathbf{x}} + dy\hat{\mathbf{y}} + dz\hat{\mathbf{z}}$), no matter which way the path goes, and to let the limits of integration set the direction. Here the limits run from ∞ to r, so $d\bar{r}$ is negative at every step, and the integral takes care of the sign by itself. I often made this mistake myself before I understood where the extra minus sign came from. The minus sign in front of the integral is a different matter. It is part of the definition of V, and it is what makes the field point from high potential toward low; the limits take care of the direction of the path.

☞ *Caution*: Write $d\boldsymbol{\ell}$ in the form (1.12) whichever way a path goes, and let the limits of integration set the direction.

Potentials add, because fields do. If $\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2 + \cdots$, then the line integral of $\mathbf{E}$ is the sum of the line integrals of the separate fields, and $V = V_1 + V_2 + \cdots$, as long as the same reference point is used for all of them. For a collection of point charges,

$$V(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\sum_i \frac{q_i}{\mathscr{r}_i}, \tag{5.11}$$

where $\mathscr{r}_i$ is the distance from the ith charge to the field point (Figure 5.9). Compared with the field, (3.5), this is a sum of numbers rather than of vectors.

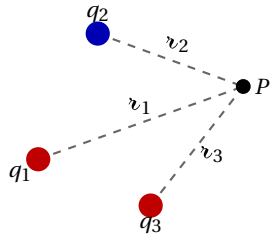


Figure 5.9 Point charges q_i and a field point P. Each charge contributes $q_i/4\pi\varepsilon_0\mathscr{r}_i$ to the potential at P, and only the distances $\mathscr{r}_i$ enter.

5.4.1 Work and potential energy

The potential is named for its connection with potential energy. A charge Q in an electrostatic field feels the force $\mathbf{F} = Q\mathbf{E}$. To move it slowly from $\mathbf{a}$ to $\mathbf{b}$, so that it gains no kinetic energy, you have to exert the opposite force, $-Q\mathbf{E}$, and the work you do is

$$\begin{aligned} W &= \int_{\mathbf{a}}^{\mathbf{b}} (-Q\mathbf{E})\cdot d\boldsymbol{\ell} \\ &= -Q\int_{\mathbf{a}}^{\mathbf{b}} \mathbf{E}\cdot d\boldsymbol{\ell} \\ &= Q\,[V(\mathbf{b}) - V(\mathbf{a})]\,, \end{aligned}$$

by (5.7). That is, $W = Q\Delta V$. The work does not depend on the path, and it is not lost: it is stored as potential energy, which the field gives back when the charge returns to $\mathbf{a}$, by any path. (Friction is the standard counterexample. The work done against friction depends on the path, a longer path costs more, and the work

is not recovered.) With the reference point at infinity, the potential energy of the charge Q at $\mathbf{r}$ is the work needed to bring it there from infinity,

$$U = Q\,V(\mathbf{r}). \tag{5.12}$$

With words, the potential is the potential energy per unit charge: potential is to potential energy as field is to force. The field and the potential describe what the charges that produce them would do to a charge placed nearby; the force and the potential energy are what they actually do to the particular charge Q. Here Q is a *test charge*, an extra charge that is not one of the charges producing V, and those charges are assumed to stay in place while Q moves.

Now let the field be that of a point charge q_1 at $\mathbf{r}_1$, and let the test charge be a second point charge q_2 at $\mathbf{r}_2$. By (5.10) and (5.12),

$$U = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r_{12}}, \tag{5.13}$$

where $r_{12} = |\mathbf{r}_2 - \mathbf{r}_1|$ is the distance between the charges. This is symmetric in the two charges: it is also the energy of q_1 in the field of q_2. So it is better thought of as the potential energy of the pair, the work needed to bring the two charges together from an infinite separation. For like charges, $U > 0$: it took work to push them together, and they fly apart if released, like the two ends of a compressed spring. For unlike charges, $U < 0$, and it takes work to pull them apart.

Potential energy is now being used in two ways, which are related but should be kept apart. The first, (5.12), is the energy of one charge in the field of other charges that are held fixed. The second, (5.13), is the energy of a configuration of charges, in which every charge both produces the field and feels it. Section 5.9 takes up the second kind in general: the energy needed to assemble a whole distribution of charge.

In atomic and nuclear physics the usual unit of energy is the **electron volt** (eV), the energy gained by a particle with charge e that falls through a potential difference of one volt:

$$1\,\text{eV} = \left(1.602176634 \times 10^{-19}\,\text{C}\right)(1\,\text{V}) = 1.602176634 \times 10^{-19}\,\text{J},$$

exactly, since e has had an exact value in the SI since 2019 (Section 17.4). For a particle with charge $\pm e$, a potential difference in volts is the same number as an energy in electron volts.

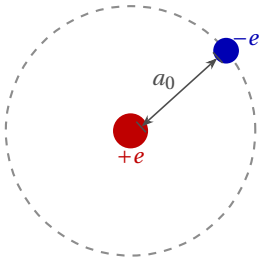
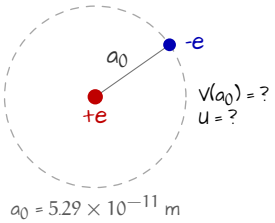


Figure 5.10 The simple picture of the hydrogen atom in Example 5.3: an electron in a circular orbit of radius a_0 around the proton.

Example 5.3

In the simplest picture of the hydrogen atom, the electron orbits the proton at a distance $a_0 = 5.29 \times 10^{-11}$ m, the Bohr radius. Find the potential of the proton at that distance, and the potential energy of the electron in electron volts.

Solution:



A labeled picture of the setup comes first: the proton, and the electron a distance a_0 from it. We want the proton's potential at that distance and the electron's potential energy there.

By (5.10),

$$\begin{aligned} V(a_0) &= \frac{1}{4\pi\varepsilon_0}\frac{e}{a_0} \\ &= \frac{\left(8.988\times10^9\,\mathrm{N\,m^2/C^2}\right)\left(1.602\times10^{-19}\,\mathrm{C}\right)}{5.29\times10^{-11}\,\mathrm{m}} \\ &= 27.2\,\mathrm{V}. \end{aligned}$$

The electron's charge is $-e$, so its potential energy is $U = -eV(a_0) = -27.2\,\mathrm{eV}$. That is exactly -1 hartree, the unit of energy in the atomic units described in Section 17.3.1; one hartree is $e^2/4\pi\varepsilon_0 a_0$. It takes only 13.6 eV, half as much, to ionize the atom. (Why does it take only half of 27.2 eV to ionize the atom? Answer in footnote.[3])

Reality check: The units work out: $(\mathrm{N\,m^2/C^2})(\mathrm{C})/\mathrm{m}$ is a newton meter per coulomb, a joule per coulomb, which is the volt.

5.5 Solving for V when given ρ: script-r method

A continuous distribution of charge can be divided into infinitesimal pieces dq, each of which acts like a point charge. By superposition, (5.11), the potential is the sum of the potentials of the pieces, and since there are infinitely many infinitesimal pieces, the sum becomes an integral. With dq from (3.8), the potential of a line, surface, or volume charge is

$$V(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\int_{\mathscr{L}}\frac{\lambda(\mathbf{r}')}{\mathscr{r}}\,d\ell', \tag{5.14}$$

$$V(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\int_{\mathscr{S}}\frac{\sigma(\mathbf{r}')}{\mathscr{r}}\,da', \tag{5.15}$$

$$V(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\int_{\mathscr{V}}\frac{\rho(\mathbf{r}')}{\mathscr{r}}\,d\tau', \tag{5.16}$$

where, as in Chapter 3, $\mathscr{r} = |\mathbf{r}-\mathbf{r}'|$ is the distance from the source point to the field point (Figure 5.11). Compare these with the field integrals (3.11)–(3.13). In

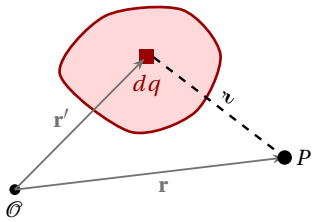


Figure 5.11 The script-r method for the potential. The piece dq of the charge distribution, at $\mathbf{r}'$, contributes $dq/4\pi\varepsilon_0\mathscr{r}$ at the field point P, and only the distance $\mathscr{r}$ enters.

[3] The electron is not at rest. Its kinetic energy, +13.6 eV, brings its total energy up to −13.6 eV, and ionizing the atom only has to raise the total to zero.

place of $\boldsymbol{\mathscr{r}}/\mathscr{r}^3$ there is $1/\mathscr{r}$, and there are no vectors at all: there is one integral to do instead of three, and no variable unit vectors for Colton's First Commandment to apply to. The last of the three is also exactly the scalar function U of the Helmholtz theorem, (2.5), with $D = \rho/\varepsilon_0$, and the theorem's construction of the field is $\mathbf{E} = -\nabla V$.

The contribution of each piece, $dq/4\pi\varepsilon_0\mathscr{r}$, goes to zero far away, so these integrals put the reference point at infinity. They work for any charge distribution confined to a finite region. For a distribution with an infinite amount of charge, such as the infinite line, they give an infinite potential everywhere, and the potential has to be found some other way, as in Example 5.1 (see also Exercise P5.7).

Like their counterparts for the field, these integrals are blueprints, and the steps for constructing a particular integral are those of Section 3.5, without the vector parts:

1. Draw a picture, mark an arbitrary piece of the charge distribution, and draw $\mathbf{r}$, $\mathbf{r}'$, and $\boldsymbol{\mathscr{r}}$.

2. Write down $\mathbf{r}$, $\mathbf{r}'$, $\boldsymbol{\mathscr{r}} = \mathbf{r} - \mathbf{r}'$, and $\mathscr{r}$, each on a line of its own, and the line, area, or volume element. Only the distance $\mathscr{r}$ goes into the integral, but writing the vector $\boldsymbol{\mathscr{r}}$ first is the surest way to get it right. The element is the size of one piece (a piece of a ring of radius R has length $R\,d\phi'$, not $d\phi'$), and $\lambda\,d\ell'$, $\sigma\,da'$, or $\rho\,d\tau'$ is the charge on it.

3. Write the charge density in terms of the primed variables, using the density, not the total charge. Then set the limits: integrate over the source, counting each piece of it once.

4. Put the pieces into the blueprint and do the integral.

5. Check the result. Far from a distribution with net charge Q, the potential must approach that of a point charge, $Q/4\pi\varepsilon_0 r$.

If the field is wanted as well, it comes from $\mathbf{E} = -\nabla V$. There are two cautions here. First, to take a gradient you need V as a function of all three coordinates near the point of interest, not only along a line. If V is known only on the z axis, then $-dV/dz$ gives the z component of the field on the axis, but the potential along the axis says nothing about E_x or E_y. When symmetry says that the field on the axis points along $\hat{\mathbf{z}}$, that is enough; when it does not, it is not enough. Second, $V = 0$ at a point does not mean that $\mathbf{E} = 0$ there, and $\mathbf{E} = 0$ does not mean that $V = 0$, as the inside of the shell of Example 5.2 showed. The value of a function says nothing about its slope. The examples show both cautions at work.

5.5.1 Worked examples

Example 5.4

A ring of radius R lies in the xy plane, centered on the origin, and carries a uniform line charge density λ. Find the potential and the field at a point on the z axis.

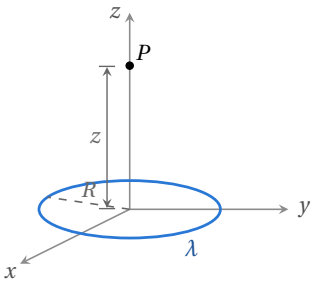
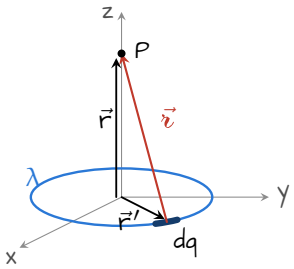


Figure 5.12 A ring of radius R with uniform line charge density λ, and the field point P on the z axis.

Solution:



We copy the ring from Figure 5.12 and add a piece of it, $dq = \lambda R\, d\phi'$, the vectors $\mathbf{r}$ (to P) and $\mathbf{r}'$ (to the piece), and the separation vector from the piece to P.

For the piece of the ring at angle ϕ',

$$\begin{aligned}
\mathbf{r} &= z\hat{\mathbf{z}} \\
\mathbf{r}' &= R\hat{\mathbf{s}}' = R\left(\cos\phi'\,\hat{\mathbf{x}} + \sin\phi'\,\hat{\mathbf{y}}\right) \\
\boldsymbol{\ell} &= \mathbf{r} - \mathbf{r}' = -R\cos\phi'\,\hat{\mathbf{x}} - R\sin\phi'\,\hat{\mathbf{y}} + z\hat{\mathbf{z}} \\
\ell &= \sqrt{R^2 + z^2},
\end{aligned}$$

the same distance for every piece, and $d\ell' = R\, d\phi'$. By (5.14),

$$\begin{aligned}
V(z) &= \frac{1}{4\pi\varepsilon_0}\int_0^{2\pi} \frac{\lambda R\, d\phi'}{\sqrt{R^2+z^2}} \\
&= \frac{1}{4\pi\varepsilon_0}\,\frac{2\pi R\lambda}{\sqrt{R^2+z^2}} \\
&= \frac{1}{4\pi\varepsilon_0}\,\frac{Q}{\sqrt{R^2+z^2}},
\end{aligned}$$

where $Q = 2\pi R\lambda$ is the total charge. Since every piece of the ring is the same distance from P, the ring's potential there is the same as if all of its charge were at a single point that distance away.

By symmetry, the field on the axis points along the axis. (Why does symmetry rule out a horizontal component on the axis? Answer in footnote.[4]) So only $\partial V/\partial z$ is needed, and the potential on the axis supplies it:

$$\mathbf{E} = -\frac{\partial V}{\partial z}\,\hat{\mathbf{z}} = \frac{1}{4\pi\varepsilon_0}\,\frac{Qz}{\left(R^2+z^2\right)^{3/2}}\,\hat{\mathbf{z}}.$$

[4] Each piece of the ring has a partner on the opposite side, the same distance from P, whose field has the opposite horizontal component.

With Coulomb's law the same result takes a vector integral, in which the horizontal components have to be shown to cancel.

Reality check: Far from the ring, $|z| \gg R$, the potential approaches $Q/4\pi\varepsilon_0|z|$ and the field approaches $Q/4\pi\varepsilon_0 z^2$, pointing away from the ring: the potential and field of a point charge, as they should. At the center, $z = 0$, the field is zero, as it must be, since there the fields of the pieces cancel in pairs.

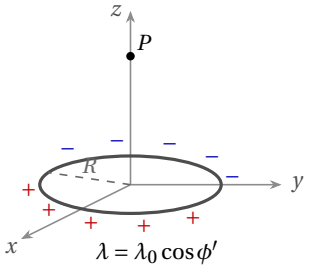
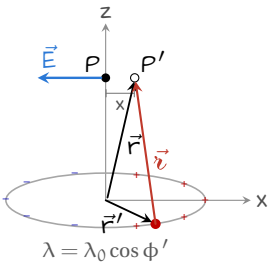


Figure 5.13 The ring of Example 5.5, with its charge positive on the side toward $+x$ and negative on the side toward $-x$, and a field point P on the z axis.

Example 5.5

The ring of Example 5.4 now carries the line charge density $\lambda = \lambda_0 \cos\phi'$, positive on the side toward $+x$ and negative on the side toward $-x$, as in Situation 3 of Tutorial 3.1. Find the potential on the z axis, and use it to find the field there.

Solution:



A sketch shows what to expect: the ring, with its positive half toward $+x$ and its negative half toward $-x$; a field point P on the axis, where the field must point from the positive half toward the negative half, along $-\hat{\mathbf{x}}$; and a second point P' a small distance x off the axis, which we will need below, with the vectors $\mathbf{r}$, $\mathbf{r}'$ and $\boldsymbol{\imath}$ for a piece of the ring.

We will need the point P', at $(x, 0, z)$, as well as P, which is the same point with $x = 0$, so we write the vectors for P':

$$\mathbf{r} = x\hat{\mathbf{x}} + z\hat{\mathbf{z}}$$
$$\mathbf{r}' = R\hat{\mathbf{s}}' = R\left(\cos\phi'\hat{\mathbf{x}} + \sin\phi'\hat{\mathbf{y}}\right)$$
$$\boldsymbol{\imath} = \mathbf{r} - \mathbf{r}' = \left(x - R\cos\phi'\right)\hat{\mathbf{x}} - R\sin\phi'\hat{\mathbf{y}} + z\hat{\mathbf{z}}$$
$$\imath = \sqrt{\left(x - R\cos\phi'\right)^2 + R^2\sin^2\phi' + z^2} = \sqrt{R^2 + z^2 - 2xR\cos\phi' + x^2}.$$

On the axis, where $x = 0$, every piece of the ring is the distance $\sqrt{R^2 + z^2}$ from P, so

$$V(0,0,z) = \frac{1}{4\pi\varepsilon_0}\int_0^{2\pi} \frac{\lambda_0 \cos\phi' R\, d\phi'}{\sqrt{R^2 + z^2}} = 0,$$

since the positive and negative halves cancel. The potential is zero everywhere on the axis. Does that mean that the field is zero there? No. The positive charge on one side and the negative charge on the other produce a field on the axis that points from the positive side toward the negative side, along $-\hat{\mathbf{x}}$. The potential on the axis is zero, but its slope in the x direction is not, and the slope is what gives the field. It is like finding the

slope of $f(x) = x^2$ at $x = 1$. If you set $x = 1$ first, you get $f = 1$, whose derivative is zero, which is wrong; the variable has to be kept until after the derivative is taken.

So find the potential at P', a small distance x off the axis. To first order in x, the binomial expansion of $1/\mathscr{r}$ from the last line above gives

$$\frac{1}{\mathscr{r}} \approx \frac{1}{\sqrt{R^2+z^2}}\left(1 + \frac{xR\cos\phi'}{R^2+z^2}\right).$$

The potential is then

$$V(x,0,z) \approx \frac{\lambda_0 R}{4\pi\varepsilon_0\sqrt{R^2+z^2}}\int_0^{2\pi}\cos\phi'\left(1+\frac{xR\cos\phi'}{R^2+z^2}\right)d\phi' = \frac{\lambda_0 R^2\, x}{4\varepsilon_0\left(R^2+z^2\right)^{3/2}},$$

using $\int_0^{2\pi}\cos^2\phi'\,d\phi' = \pi$. The potential rises linearly with x across the axis, so on the axis

$$E_x = -\frac{\partial V}{\partial x} = -\frac{\lambda_0 R^2}{4\varepsilon_0\left(R^2+z^2\right)^{3/2}},$$

pointing along $-\hat{\mathbf{x}}$, as expected. The same steps with a small y give no first-order term, so $E_y = 0$ on the axis, and $E_z = -\partial V/\partial z = 0$ there too, since V vanishes at every point of the axis. You can check that this agrees with the field you found from Coulomb's law in the tutorial.

Reality check: Reflecting the ring in the yz plane, $x \to -x$, swaps its positive and negative halves, which turns the charge into its negative. The potential must then change sign too, $V(-x,0,z) = -V(x,0,z)$, and the result, proportional to x, does.

Example 5.6

A square loop with sides of length a lies in the xy plane, centered on the origin, and carries a uniform line charge density λ. Find the potential and the field at a point on the z axis.

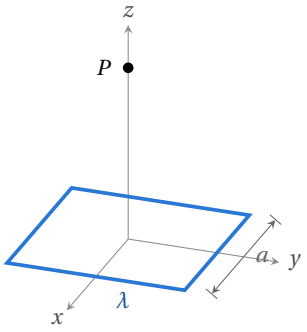
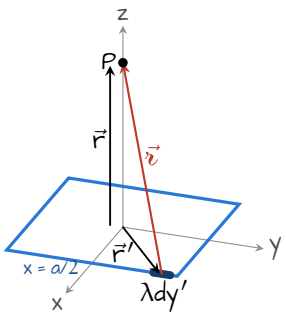


Figure 5.14 A square loop of side a with uniform line charge density λ, and the field point P on the z axis.

Solution:



We start from the loop of Figure 5.14 and add one piece of the side at $x = a/2$, the vectors $\mathbf{r}$ (to the field point P) and $\mathbf{r}'$ (to the piece), and the separation vector from the piece to P. The loop has symmetry, but not one of the three kinds that Gauss's law needs, so the field cannot be found that way. The potential, though, is not hard. Start with the side at $x = a/2$.

For its piece at y',

$$\begin{aligned}
\mathbf{r} &= z\hat{\mathbf{z}} \\
\mathbf{r}' &= \frac{a}{2}\hat{\mathbf{x}} + y'\hat{\mathbf{y}} \\
\boldsymbol{\mathscr{r}} &= \mathbf{r} - \mathbf{r}' = -\frac{a}{2}\hat{\mathbf{x}} - y'\hat{\mathbf{y}} + z\hat{\mathbf{z}} \\
\mathscr{r} &= \sqrt{a^2/4 + y'^2 + z^2},
\end{aligned}$$

and $d\ell' = dy'$. With the standard integral $\int dy/\sqrt{y^2+c^2} = \ln\left(y + \sqrt{y^2+c^2}\right)$, the potential of this side is

$$V_{\text{side}} = \frac{\lambda}{4\pi\varepsilon_0}\int_{-a/2}^{a/2} \frac{dy'}{\sqrt{y'^2 + a^2/4 + z^2}} = \frac{\lambda}{4\pi\varepsilon_0}\ln\frac{\sqrt{a^2/2+z^2}+a/2}{\sqrt{a^2/2+z^2}-a/2}.$$

The other three sides are each the same distance from P, piece for piece, so they contribute the same potential. Because the potential is a scalar, the four contributions simply add, and

$$V(z) = 4V_{\text{side}} = \frac{\lambda}{\pi\varepsilon_0}\ln\frac{\sqrt{a^2/2+z^2}+a/2}{\sqrt{a^2/2+z^2}-a/2}.$$

For the field, by contrast, each side's contribution is a vector pointing in a different direction. By symmetry, the field on the axis points along $\hat{\mathbf{z}}$: the loop looks the same after a quarter turn about the axis, so a horizontal component would have to point in four directions at once. So again only $\partial V/\partial z$ is needed. Writing $S = \sqrt{a^2/2+z^2}$, so that $dS/dz = z/S$,

$$\begin{aligned}
E_z &= -\frac{\partial V}{\partial z} \\
&= -\frac{\lambda}{\pi\varepsilon_0}\frac{z}{S}\left(\frac{1}{S+a/2} - \frac{1}{S-a/2}\right) \\
&= \frac{1}{4\pi\varepsilon_0}\frac{4\lambda a z}{\left(z^2+a^2/4\right)\sqrt{z^2+a^2/2}},
\end{aligned}$$

using $S^2 - a^2/4 = z^2 + a^2/4$.

Reality check: Far from the loop ($z \gg a$), S is nearly z, and to first order in a/z,

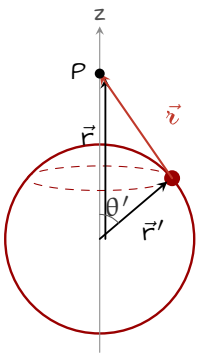
$$\begin{aligned} V &\approx \frac{\lambda}{\pi\varepsilon_0}\ln\frac{z+a/2}{z-a/2} \\ &\approx \frac{\lambda}{\pi\varepsilon_0}\frac{a}{z} \\ &= \frac{1}{4\pi\varepsilon_0}\frac{4\lambda a}{z}, \end{aligned}$$

the potential of a point charge equal to the loop's total charge, $4\lambda a$. The field likewise approaches $(1/4\pi\varepsilon_0)(4\lambda a/z^2)$.

Example 5.7

Find the potential of the uniformly charged spherical shell of Example 5.2 directly from its charge, with (5.15).

Solution:



The sketch comes first, and symmetry tells us how to draw it: the potential depends only on the distance r from the center, so we put the field point on the z axis, a distance r from the center. Then we add a patch of the shell at polar angle θ', the vectors $\mathbf{r}$ (to the field point) and $\mathbf{r}'$ (to the patch), and the separation vector from the patch to the field point. (The sketch shows a field point outside the shell; for one inside, the same triangle has $r < R$.)

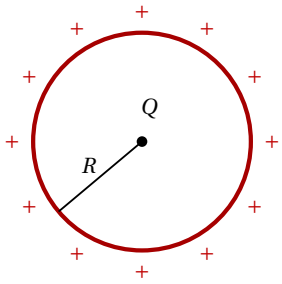


Figure 5.15 The uniformly charged spherical shell of Example 5.7, radius R and total charge Q, in cross section.

For the patch of the shell at polar angle θ' and azimuth ϕ',

$$\begin{aligned} \mathbf{r} &= r\,\hat{\mathbf{z}} \\ \mathbf{r}' &= R\,\hat{\mathbf{r}}' = R\left(\sin\theta'\cos\phi'\,\hat{\mathbf{x}} + \sin\theta'\sin\phi'\,\hat{\mathbf{y}} + \cos\theta'\,\hat{\mathbf{z}}\right) \\ \boldsymbol{\mathscr{r}} &= \mathbf{r} - \mathbf{r}' = -R\sin\theta'\cos\phi'\,\hat{\mathbf{x}} - R\sin\theta'\sin\phi'\,\hat{\mathbf{y}} + \left(r - R\cos\theta'\right)\hat{\mathbf{z}} \\ \mathscr{r} &= \sqrt{R^2\sin^2\theta' + (r - R\cos\theta')^2} = \sqrt{r^2 + R^2 - 2rR\cos\theta'}, \end{aligned}$$

the law of cosines for the triangle in the sketch, and $da' = R^2\sin\theta'\,d\theta'\,d\phi'$. With $\sigma = Q/4\pi R^2$,

$$V(r) = \frac{1}{4\pi\varepsilon_0}\int_0^{2\pi}\int_0^{\pi}\frac{\sigma R^2\sin\theta'\,d\theta'\,d\phi'}{\sqrt{r^2+R^2-2rR\cos\theta'}} = \frac{\sigma R^2}{2\varepsilon_0}\int_0^{\pi}\frac{\sin\theta'\,d\theta'}{\sqrt{r^2+R^2-2rR\cos\theta'}}.$$

The numerator is, apart from a constant factor, the derivative of the expression under the square root, so

$$\int_0^{\pi}\frac{\sin\theta'\,d\theta'}{\sqrt{r^2+R^2-2rR\cos\theta'}} = \frac{1}{rR}\left[\sqrt{r^2+R^2-2rR\cos\theta'}\,\right]_0^{\pi} = \frac{1}{rR}\Big[(r+R) - |r-R|\Big].$$

The absolute value is the one subtle step: $\sqrt{(r-R)^2}$ is $r-R$ outside the shell but $R-r$ inside it. Outside, the bracket is $2R$, and inside, it is $2r$, so the integral is $2/r$ outside and $2/R$ inside. Therefore

$$V(r) = \frac{\sigma R^2}{\epsilon_0 r} = \frac{Q}{4\pi\epsilon_0 r} \quad (r > R), \qquad V(r) = \frac{\sigma R^2}{\epsilon_0 R} = \frac{Q}{4\pi\epsilon_0 R} \quad (r < R).$$

Compare this with the script-r calculation of the field of the same shell in Section 3.5.3: without vector components to keep track of, the integral takes only a few lines.

Reality check: These are the results of Example 5.2, which found the potential from the field instead.

Example 5.8

Find the potential inside and outside the ball of Example 4.5, which has radius R and charge density $\rho = \rho_0 r^2/R^2$, in two ways: from its field, and directly from its charge.

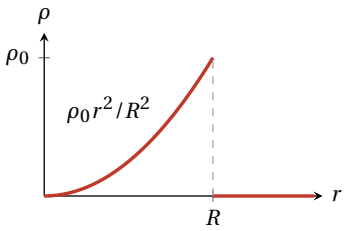
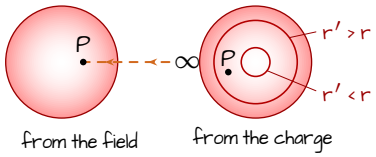


Figure 5.16 The charge density of the ball of Example 5.8, $\rho_0 r^2/R^2$ inside and zero outside.

Solution:



We sketch the ball, darker toward its surface where the density is larger, and the two routes to the potential: from the field, along a radial path in from the reference point at infinity, and from the charge, by building the ball out of thin shells, some inside the field point P and some outside it.

From the field. Example 4.5 found that the field is $\rho_0 r^3/5\epsilon_0 R^2$ inside the ball and $\rho_0 R^3/5\epsilon_0 r^2$ outside it, along $\hat{\mathbf{r}}$. Working from the outside in, along a radial path from infinity, outside the ball

$$V(r) = -\int_\infty^r \frac{\rho_0 R^3}{5\epsilon_0 \bar{r}^2}\, d\bar{r} = \frac{\rho_0 R^3}{5\epsilon_0 r}, \qquad r > R,$$

which is $Q/4\pi\epsilon_0 r$ for the ball's total charge, $Q = 4\pi\rho_0 R^3/5$. Inside, the path comes in from infinity through the outside region first:

$$\begin{aligned} V(r) &= -\int_\infty^R \frac{\rho_0 R^3}{5\epsilon_0 \bar{r}^2}\, d\bar{r} - \int_R^r \frac{\rho_0 \bar{r}^3}{5\epsilon_0 R^2}\, d\bar{r} = \frac{\rho_0 R^2}{5\epsilon_0} - \frac{\rho_0}{20\epsilon_0 R^2}\left(r^4 - R^4\right) \\ &= \frac{\rho_0 R^2}{4\epsilon_0} - \frac{\rho_0 r^4}{20\epsilon_0 R^2}, \qquad r < R. \end{aligned}$$

Reality check: At $r = R$ both expressions give $\rho_0 R^2/5\epsilon_0$, so the potential is continuous, as it must be.

From the charge. The ball can be built from thin concentric shells. The shell of radius r' and thickness dr' carries the charge $dq = \rho(r')\,4\pi r'^2\,dr'$, and by Example 5.2 its potential is $dq/4\pi\varepsilon_0 r$ at a field point outside it ($r' < r$), and $dq/4\pi\varepsilon_0 r'$ at a field point inside it ($r' > r$). A field point inside the ball is outside the shells with $r' < r$ and inside the shells with $r' > r$, so

$$\begin{aligned} V(r) &= \frac{1}{4\pi\varepsilon_0}\left[\frac{1}{r}\int_0^r \rho(r')\,4\pi r'^2\,dr' + \int_r^R \frac{\rho(r')\,4\pi r'^2}{r'}\,dr'\right] \\ &= \frac{\rho_0}{\varepsilon_0 R^2}\left[\frac{1}{r}\int_0^r r'^4\,dr' + \int_r^R r'^3\,dr'\right] = \frac{\rho_0}{\varepsilon_0 R^2}\left[\frac{r^4}{5} + \frac{R^4 - r^4}{4}\right] \\ &= \frac{\rho_0 R^2}{4\varepsilon_0} - \frac{\rho_0\, r^4}{20\varepsilon_0 R^2}, \end{aligned}$$

the same as before. (Outside the ball, the field point is outside every shell, and the first term alone, with its upper limit changed to R, gives $\rho_0 R^3/5\varepsilon_0 r$.) The potential is largest at the center, where it is $\rho_0 R^2/4\varepsilon_0$, even though the field there is zero.

5.5.2 Choosing a route

There are now two ways to find the potential of a given charge distribution. One is to find the field first, usually with Gauss's law, and integrate it with (5.8). The other is to integrate the charge directly with (5.14)–(5.16), and then, if the field is wanted, to take $-\nabla V$. Which is easier depends on the symmetry of the problem. If the symmetry is good enough for Gauss's law, find the field first and integrate it: Gauss's law gives the field in a few lines, and the remaining integral is one-dimensional. If it is not, as for the ring and the square loop, find the potential first: one scalar integral is easier than the three components of a vector integral, and the derivatives at the end are easy. Usually one route is easy and the other is hard.

5.6 Numerical method: direct integration

When the integrals (5.14)–(5.16) cannot be done by hand, the chunk-and-sum method of Section 3.6 does them numerically, with one simplification: each chunk now contributes a number instead of a vector. Divide the charge distribution into N small chunks, treat chunk i as a point charge Δq_i at its center $\mathbf{r}'_i$, and add up their potentials with (5.11):

$$V(\mathbf{r}) \approx \frac{1}{4\pi\varepsilon_0}\sum_{i=1}^{N}\frac{\Delta q_i}{\mathscr{r}_i}, \qquad \mathscr{r}_i = |\mathbf{r} - \mathbf{r}'_i|. \tag{5.17}$$

Then increase N until the answer stops changing at the accuracy you need. As before, that last step is what tells you that the answer is right.

As an example, take a ring of radius $R = 5.0\,\text{cm}$, in the xy plane and centered on the origin, carrying a charge $Q = 1.0\,\text{nC}$ spread uniformly around it

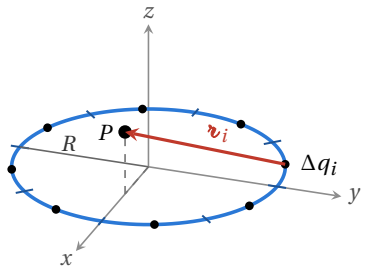


Figure 5.17 A ring of charge divided into $N = 8$ equal arcs (the dots are their centers), with the field point P and the separation vector $\boldsymbol{\mathscr{r}}_i$ from the center of one arc to P. To scale: $R = 5.0\,\text{cm}$, and P is 2.5 cm from the axis and 2.5 cm above the plane of the ring.

(Figure 5.17). Find the potential at the point P that is $s = 2.5\,\text{cm}$ from the axis and $z = 2.5\,\text{cm}$ above the plane of the ring. Example 5.4 found the potential on the axis; off the axis, the integral cannot be done in terms of elementary functions. Divide the ring into N equal arcs, with centers at the angles $\phi_i = (i - \frac{1}{2})\,2\pi/N$, each carrying $\Delta q = Q/N$, and put P at $\phi = 0$, so that $\mathbf{r} = s\hat{\mathbf{x}} + z\hat{\mathbf{z}}$. Table 5.1 shows the results of (5.17).

N	V (V)
1	113.7
2	146.8
4	161.4
8	163.33
16	163.363
32	163.363

Table 5.1 The potential at P from (5.17), with the ring divided into N equal arcs.

With a single chunk, the whole ring is treated as a point charge at $\phi = \pi$, on the far side of the ring from P, and the answer is far too low. By $N = 16$ the result has converged to six significant figures, $V = 163.363\,\text{V}$. It converges this quickly because the integrand is periodic in ϕ, and, as for the cylinder of Section 3.6, the midpoint rule converges very quickly for periodic integrands. For a point on the axis, by contrast, a single chunk already gives the exact answer (at $z = 2.5\,\text{cm}$ it is $160.8\,\text{V}$): every piece of the ring is the same distance from a point on the axis, so the chunks all contribute equally.

To get the field from the potential, take the derivatives numerically. The simplest accurate way is the **central difference**,

$$E_x \approx -\frac{V(x+h,y,z) - V(x-h,y,z)}{2h}, \tag{5.18}$$

and similarly for E_y and E_z, with a small step h. The error of a central difference is proportional to h^2. For the ring, with V calculated by the chunk sum at $N = 64$, central differences at P give the values in Table 5.2, compared with $E_s = -144.421\,\text{V/m}$ and $E_z = 1705.836\,\text{V/m}$ from the chunk sum for the field itself, (3.22). When h is made ten times smaller, from 1 mm to 0.1 mm, the error becomes a hundred times smaller, as an h^2 error should. The step cannot be made arbitrarily small, though. When h is very small, $V(x+h)$ and $V(x-h)$ agree in nearly all of their digits, and rounding errors take over their difference. The field at P points mostly up, away from the plane of the ring, and slightly in toward the axis.

h (mm)	E_s	E_z
5	−136.1	1693.7
1	−144.09	1705.36
0.1	−144.418	1705.832

Table 5.2 Central-difference values of the field components at P, in V/m, for three steps h.

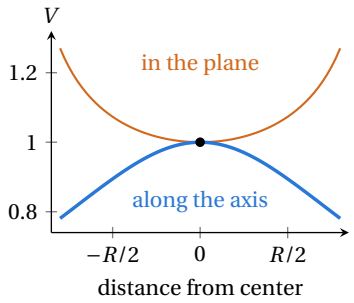


Figure 5.18 The potential near the center of a uniformly charged ring of radius R, in units of $Q/4\pi\varepsilon_0 R$. Along the axis (blue) it falls on both sides of the center; in the plane of the ring (orange) it rises. The center is a saddle point of V.

The ring has one more lesson. At its center the field is zero by symmetry, so a charge placed there would be in equilibrium. Is the equilibrium stable? Figure 5.18 shows the potential near the center, along the axis from Example 5.4 and in the plane of the ring from (5.17). Along the axis the potential falls on both sides of the center, and in the plane it rises. The center is a **saddle point**, like the middle of a horse's saddle or a pass between two mountains. A positive charge at the center is pushed back if it is displaced in the plane of the ring, but pushed farther away if it is displaced along the axis, so its equilibrium is unstable. Second central differences (or fitting parabolas to the curves near the center) give the curvatures at the center,

$$\frac{\partial^2 V}{\partial x^2} = \frac{\partial^2 V}{\partial y^2} = +\frac{1}{2}\frac{Q}{4\pi\varepsilon_0 R^3}, \qquad \frac{\partial^2 V}{\partial z^2} = -\frac{Q}{4\pi\varepsilon_0 R^3},$$

and they add up to zero. That is no accident. Section 5.10 shows that in a region with no charge the three second derivatives of V always add up to zero, so V

cannot curve upward in all three directions at once. In fact V has no maximum or minimum at all in a region with no charge (Section 7.1.2), so no arrangement of static charges can hold a charge in stable equilibrium in empty space. That is *Earnshaw's theorem*. Devices that trap charged particles, such as the ion traps used in atomic clocks and in some quantum computers, get around it with fields that change in time, or with magnetic fields.

Tips for Mathematica and Python.

- The chunk sum (5.17) takes only a line or two, with `numpy` arrays in Python or with `Table` and `Total` in Mathematica, as for the field.
- The built-in integrators (`NIntegrate` in Mathematica; `quad`, `dblquad`, and `tplquad` in Python's `scipy.integrate`) handle (5.14)–(5.16) with one integration where the field needs three, one for each component.
- Check a numerical method against a case with a known answer before trusting it. For the ring, the potential off the axis can be written in terms of a special function, the complete elliptic integral of the first kind,

$$V(s,z) = \frac{Q}{4\pi\varepsilon_0}\,\frac{2}{\pi}\,\frac{K(m)}{\sqrt{(R+s)^2+z^2}}, \qquad m = \frac{4Rs}{(R+s)^2+z^2},$$

  which is `EllipticK[m]` in Mathematica and `scipy.special.ellipk(m)` in Python. (Both take the *parameter* m; some books write $K(k)$ with $m = k^2$.) It agrees with the converged value in the table.
- As with the field, a field point on the charge distribution itself needs care, since the integrand is infinite there. On a surface or volume charge the potential is still finite (on a line charge it is not), but the integral should be split at the singular point, or Mathematica's `Exclusions` option used.

5.7 Conductors as equipotentials

Section 4.5 showed that in electrostatics the field inside a conductor is zero. It follows that *a conductor is an equipotential*: every point of it, inside and on its surface, is at the same potential. To see this, take any two points a and b of a conductor and join them by a path that runs through the metal (Figure 5.19). The field is zero at every point of the path, so by (5.7)

$$V(b) - V(a) = -\int_a^b \mathbf{E}\cdot d\boldsymbol{\ell} = 0.$$

Any path from a to b would give the same result; the path through the metal is the one that makes the answer obvious. (Just outside the surface, the field is not zero.)

This gives a second explanation of a fact from Section 4.5: the field just outside a conductor is perpendicular to its surface, because the surface is an

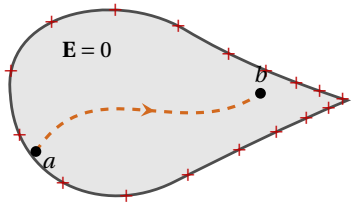


Figure 5.19 A charged conductor of irregular shape. Any two of its points, such as a on the surface and b inside, can be joined by a path through the metal, where $\mathbf{E} = 0$, so they are at the same potential. The charge, all of it on the surface, tends to crowd where the surface curves most sharply.

equipotential, and field lines cross equipotentials at right angles. It also means that a conductor has a single potential, which can be called *the potential of the conductor*. A conducting sphere of radius R with charge Q, for example, carries all of its charge on its surface, spread uniformly, so its field and potential are those of the shell of Example 5.2. The potential of the sphere is $Q/4\pi\varepsilon_0 R$, the same inside it as on its surface.

Being an equipotential does not mean having a uniform surface charge. On an isolated sphere the charge is spread uniformly, but on any other shape it is not. It arranges itself in whatever way makes the potential the same everywhere on the conductor, and there is only one such way (Section 7.1.4). The next example shows where it goes.

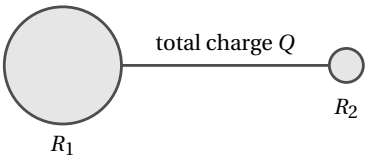


Figure 5.20 Two distant conducting spheres, of radii R_1 and R_2, joined by a thin conducting wire and carrying a total charge Q.

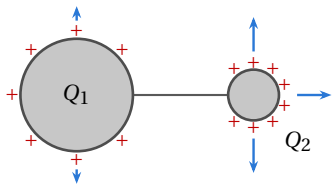
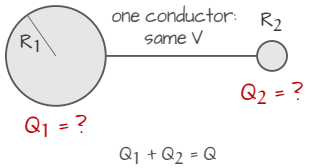


Figure 5.21 The answer to Example 5.9: the larger sphere has more of the charge, but the smaller one has the larger surface charge density (signs drawn closer together) and the stronger field just outside it. Schematic. Animation of one sphere shrinking and growing back: the two potentials stay equal, while the shrinking sphere's surface charge density and field grow.

Example 5.9

Two conducting spheres, of radii R_1 and R_2, are far apart compared with their sizes and are joined by a long, thin conducting wire. A total charge Q is placed on them. How is it shared between the spheres, and which sphere has the larger surface charge density?

Solution:



We redraw the spheres and the wire and label what we want, the charges Q_1 and Q_2, which add up to Q. The two spheres and the wire form a single conductor, so the spheres are at the same potential.

Because they are far apart, the potential of each is very nearly that of an isolated sphere: the other sphere's contribution is smaller by a factor of roughly the ratio of a sphere's radius to the distance between them, and the thin wire carries very little charge. So

$$\frac{Q_1}{4\pi\varepsilon_0 R_1} = \frac{Q_2}{4\pi\varepsilon_0 R_2}, \qquad \text{that is,} \qquad \frac{Q_1}{R_1} = \frac{Q_2}{R_2},$$

with $Q_1 + Q_2 = Q$. Solving these together,

$$Q_1 = \frac{R_1}{R_1 + R_2} Q, \qquad Q_2 = \frac{R_2}{R_1 + R_2} Q.$$

The larger sphere gets more of the charge, in proportion to its radius. The surface charge densities are $\sigma = Q/4\pi R^2$, so

$$\frac{\sigma_1}{\sigma_2} = \frac{Q_1/R_1^2}{Q_2/R_2^2} = \frac{R_2}{R_1}.$$

The smaller sphere has the larger surface charge density, in inverse proportion to its radius, and the field just outside it, σ/ε_0, is larger by the same factor (Figure 5.21).

Reality check: For two equal spheres, $R_1 = R_2$, each gets half of the charge and the two surface charge densities are equal, as symmetry requires.

A sharply curved part of a conductor behaves much like a very small sphere: the surface charge density, and with it the field just outside, tends to be largest where the surface curves most sharply, and it becomes very large near a sharp point. That has practical consequences. Air breaks down when the field in it exceeds about 3×10^6 V/m: the few free electrons in the air are accelerated enough between collisions to knock electrons off the molecules they hit, and the air becomes a conductor. Near a sharp point on a charged conductor this happens first, and charge leaks away into the air in a faint glow called a **corona discharge**. That is why high-voltage equipment is built with smooth, rounded surfaces, and why the terminal of a Van de Graaff generator is a large, smooth sphere. (Exercise P5.12 shows how such a generator keeps adding charge to its terminal.) The same crowding of charge makes the tips of tall, pointed objects, such as lightning rods, the places where the field under a thundercloud is strongest.

5.8 Capacitance

Take two conductors and connect them to the terminals of a battery (Figure 5.22). The battery moves charge from one conductor to the other. It does not create charge; it acts like a pump, and it keeps pumping until the potential difference between the conductors equals its own voltage. The conductors are then left with equal and opposite charges, $+Q$ and $-Q$. Such a pair of conductors is a **capacitor**.

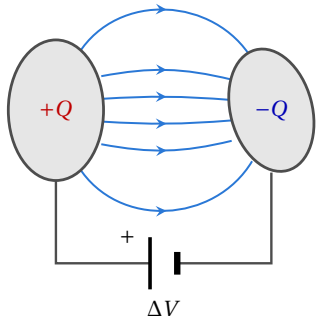


Figure 5.22 A capacitor: two conductors with charges $+Q$ and $-Q$. A battery connected between them moves charge from one to the other until their potential difference equals its voltage, ΔV. The field lines run from the positive conductor to the negative one.

How much charge does a given voltage put on the conductors? Whatever their shapes, Q is proportional to the potential difference. If the charge at every point of the two conductors is doubled, each conductor is still an equipotential, so (by uniqueness, as above) that is how the doubled charge arranges itself. Doubling the surface charge density at every point doubles the field everywhere, which doubles the potential difference. The ratio of the two is the **capacitance**,

$$C = \frac{Q}{\Delta V}, \tag{5.19}$$

where ΔV is the potential of the positive conductor minus that of the negative one,

$$\Delta V = V_+ - V_- = -\int_-^+ \mathbf{E}\cdot d\boldsymbol{\ell}, \tag{5.20}$$

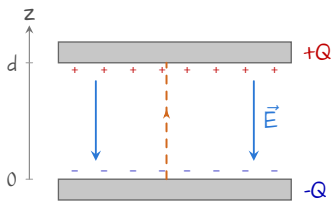
with the line integral taken along any path from the negative conductor to the positive one. With this convention, C is positive. Integrating in the other direction gives $-\Delta V$; sign errors of this kind are common in capacitor calculations, and the simplest way to avoid them is always to integrate from − to +. The capacitance depends only on the geometry: the sizes and shapes of the conductors and how they are arranged. (Chapter 9 adds the effect of an insulating material between them.) The unit of capacitance is the **farad**, 1 F = 1 C/V, which is a large capacitance; most capacitors in electronic circuits are measured in microfarads, nanofarads, or picofarads.

To find a capacitance, put $+Q$ on one conductor and $-Q$ on the other, find the field between them, integrate it from the negative conductor to the positive one, and divide Q by the result. The charge Q should cancel, which is a useful check.

Example 5.10

Find the capacitance of two parallel conducting plates, each of area A, separated by a distance d that is small compared with the size of the plates.

Solution:



The plates, drawn edge-on, come first: $+Q$ on the upper plate and $-Q$ on the lower one, with the z axis pointing up and the lower plate at $z = 0$. The dashed line is the path of integration, from the negative plate up to the positive one.

Ignoring effects near the edges, the charges sit on the inner faces of the plates (Exercise P4.12) with surface charge densities $\pm\sigma = \pm Q/A$, and by Example 4.6 the field between the plates is uniform, $\mathbf{E} = -(\sigma/\varepsilon_0)\,\hat{\mathbf{z}}$, pointing from the positive plate down to the negative one. (Why is the field between the plates σ/ε_0, when a single sheet of charge gives $\sigma/2\varepsilon_0$? Answer in footnote.[5]) Integrating from the negative plate at $z = 0$ up to the positive plate at $z = d$, with $d\boldsymbol{\ell} = dz\,\hat{\mathbf{z}}$,

$$\begin{aligned}\Delta V &= -\int_0^d \left(-\frac{\sigma}{\varepsilon_0}\hat{\mathbf{z}}\right)\cdot(dz\,\hat{\mathbf{z}}) \\ &= \frac{\sigma d}{\varepsilon_0} \\ &= \frac{Qd}{\varepsilon_0 A}.\end{aligned}$$

So

$$C = \frac{Q}{\Delta V} = \frac{\varepsilon_0 A}{d}.$$

The capacitance grows with the area of the plates and shrinks as they are moved apart. For plates 10 cm on a side and 1.0 mm apart, it is

$$C = \frac{\left(8.85\times10^{-12}\,\mathrm{F/m}\right)\left(0.010\,\mathrm{m}^2\right)}{1.0\times10^{-3}\,\mathrm{m}} = 88.5\,\mathrm{pF}.$$

Reality check: The charge canceled, as it should, and the units are right: ε_0 is in farads per meter, so $\varepsilon_0 A/d$ is in farads.

[5] Each plate's charge alone gives $\sigma/2\varepsilon_0$ on both sides of it. Between the plates the two fields point the same way and add; outside, they point opposite ways and cancel.

Example 5.11

(a) A conducting sphere of radius a is surrounded by a concentric conducting shell of inner radius b. Find the capacitance. (b) Find the capacitance of an isolated conducting sphere of radius a.

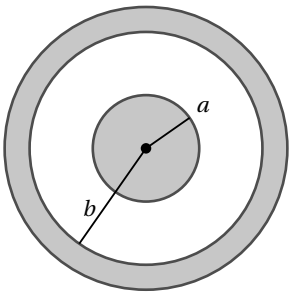
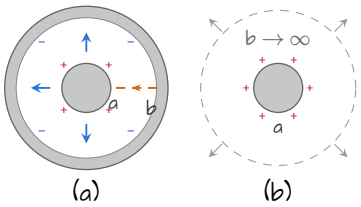


Figure 5.23 The conducting sphere and shell of Example 5.11, in cross section.

Solution:



We draw both conductors in cross section: (a) the sphere and the shell, with the field between them and the path of integration from the shell in to the sphere, and (b) the isolated sphere, whose partner conductor is far away. In (a), put $+Q$ on the inner sphere and $-Q$ on the shell.

(a) By Gauss's law, the field between them is that of a point charge Q at the center, $Q/4\pi\varepsilon_0 r^2$ along $\hat{\mathbf{r}}$. (Why does the charge $-Q$ on the shell not change the field between the conductors? Answer in footnote.[6]) Integrating from the negative conductor at $r = b$ inward to the positive one at $r = a$, with $d\boldsymbol{\ell} = d\bar{r}\,\hat{\mathbf{r}}$ and the limits setting the direction,

$$\Delta V = -\int_b^a \frac{Q}{4\pi\varepsilon_0 \bar{r}^2}\, d\bar{r} = \frac{Q}{4\pi\varepsilon_0}\left(\frac{1}{a} - \frac{1}{b}\right),$$

which is positive, as it should be. So

$$\begin{aligned} C &= \frac{Q}{\Delta V} \\ &= \frac{4\pi\varepsilon_0}{1/a - 1/b} \\ &= 4\pi\varepsilon_0 \frac{ab}{b-a}. \end{aligned}$$

Reality check: If the gap is narrow, with $b = a + d$ and $d \ll a$, this becomes $4\pi\varepsilon_0 a^2/d = \varepsilon_0 A/d$, where $A = 4\pi a^2$ is the area of the sphere: up close, the two surfaces look like parallel plates.

(b) An isolated sphere seems to have only one conductor, but the second one is far away: let the shell recede to infinity, $b \to \infty$. Then

$$C = 4\pi\varepsilon_0\, a.$$

A sphere of radius 10 cm has $C = 11\,\text{pF}$. Even the Earth, with $a = 6.37 \times 10^6\,\text{m}$, has a capacitance of only $7.1 \times 10^{-4}\,\text{F}$, which shows how large a farad is.

Reality check: The capacitance $4\pi\varepsilon_0 a$ is also the ratio of the sphere's charge to its potential, $Q/(Q/4\pi\varepsilon_0 a)$, with the reference point at infinity.

[6] It is spread uniformly over the shell, and a uniformly charged spherical shell produces no field inside itself (Example 4.2).

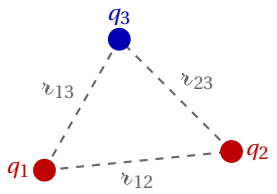


Figure 5.24 Three point charges brought in one at a time. Each pair contributes $q_i q_j/4\pi\varepsilon_0 r_{ij}$ to the energy of the configuration.

5.9 Energy stored in the field

How much work does it take to assemble a distribution of charge? This is the second kind of potential energy of Section 5.4.1, the energy of a configuration, in which all of the charges both produce the field and feel it.

5.9.1 The energy of a collection of point charges

Start with point charges, and bring them in from infinity one at a time (Figure 5.24). The first charge, q_1, costs nothing, since there is no field yet to push against. Bringing in the second costs q_2 times the potential of the first at its location, which is the pair energy (5.13), $q_1 q_2/4\pi\varepsilon_0 r_{12}$. The third costs q_3 times the potential of the first two at its location, $q_3\,(q_1/r_{13} + q_2/r_{23})/4\pi\varepsilon_0$, and so on. Each pair of charges contributes one term, and the total is

$$U = \frac{1}{4\pi\varepsilon_0}\sum_{\text{pairs}}\frac{q_i q_j}{r_{ij}} = \frac{1}{2}\,\frac{1}{4\pi\varepsilon_0}\sum_i\sum_{j\neq i}\frac{q_i q_j}{r_{ij}}.$$

Three charges have three pairs, five charges have ten, and ten charges have 45. The double sum on the right counts each pair twice, as (i, j) and as (j, i), and the factor of $\frac{1}{2}$ corrects for that; it is written this way because it leads to a simpler form. The terms with $j = i$ are left out, since a point charge does no work on itself (and those terms would be infinite). For each i, the inner sum is the potential at the location of charge i produced by all of the other charges, so

$$U = \frac{1}{2}\sum_i q_i\, V(\mathbf{r}_i), \tag{5.21}$$

where $V(\mathbf{r}_i)$ is the potential at $\mathbf{r}_i$ of every charge *except* q_i. This is the work needed to assemble the charges, starting from infinite separation, and it does not depend on the order in which they are brought in. It is also the energy that is released, as kinetic energy, if they are allowed to fly apart.

5.9.2 Continuous charge distributions

For a continuous distribution, the sum becomes an integral,

$$U = \frac{1}{2}\int \rho\, V\, d\tau, \tag{5.22}$$

over the region where there is charge. (There is no need to leave out the potential of the charge at the point itself here: the potential produced by an infinitesimal piece of a continuous distribution at its own location is negligible.) There is no need for primes, either. The answer is a number, an energy, not a function of position.

Gauss's law turns (5.22) into a statement about the field alone. Replace ρ by $\varepsilon_0 \nabla\cdot\mathbf{E}$:

$$U = \frac{\varepsilon_0}{2}\int (\nabla\cdot\mathbf{E})\, V\, d\tau.$$

The product rule (1.52), $\nabla\cdot(V\mathbf{E}) = V(\nabla\cdot\mathbf{E}) + \mathbf{E}\cdot\nabla V$, moves the derivative from $\mathbf{E}$ onto V, the way integration by parts does in one dimension. With $\nabla V = -\mathbf{E}$,

$$(\nabla\cdot\mathbf{E})\,V = \nabla\cdot(V\mathbf{E}) - \mathbf{E}\cdot\nabla V = \nabla\cdot(V\mathbf{E}) + E^2.$$

So, using the divergence theorem on the first term (this is (1.59)),

$$U = \frac{\varepsilon_0}{2}\left[\oint_{\mathcal{S}} V\mathbf{E}\cdot d\mathbf{a} + \int_{\mathcal{V}} E^2\,d\tau\right],$$

where $\mathcal{V}$ is the volume of integration and $\mathcal{S}$ is its surface. The original integral (5.22) was over the region containing the charge, but any larger volume gives the same result, since $\rho = 0$ in the added parts. So take the volume to be a sphere of radius b, and let $b\to\infty$ (Figure 5.25). Far from the charges, V falls off at least as fast as $1/b$ and E at least as fast as $1/b^2$, while the area of the sphere grows as b^2, so the surface integral falls off at least as fast as $1/b$ and vanishes. That leaves

$$U = \frac{\varepsilon_0}{2}\int_{\text{all space}} E^2\,d\tau. \tag{5.23}$$

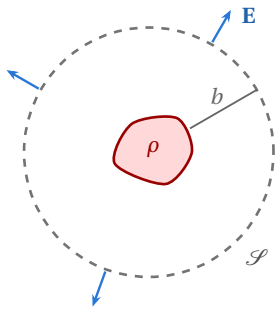


Figure 5.25 Any volume that contains all of the charge gives the same energy. For a sphere of radius b, the surface term falls off at least as fast as $1/b$, so it vanishes as $b\to\infty$.

Here $E^2 = \mathbf{E}\cdot\mathbf{E}$, which in any coordinate system is the sum of the squares of the three components, since the unit vectors are perpendicular to each other.

The two forms (5.22) and (5.23) give the same number, but they look very different, and they integrate over different regions: (5.22) only where there is charge, and (5.23) everywhere there is field. They suggest two ways of thinking about where the energy is. In one, it is the work that was needed to assemble the charges, and it belongs to the charges. In the other, it is stored in the field, with an **energy density**

$$u = \frac{\varepsilon_0}{2}E^2, \tag{5.24}$$

in joules per cubic meter, wherever the field is. In electrostatics no experiment can tell the two pictures apart: they are the same number with two names. Once fields change in time, the field picture is the one that works. Electromagnetic waves carry energy across empty space, far from any charge (Section 20.2). The energy density $\frac{1}{2}\varepsilon_0 E^2$ is also the same expression as the electrostatic stress on the surface of a conductor, (4.7), and Exercise P5.21 shows why.

Example 5.12

A ball of radius R carries a total charge Q, spread uniformly through its volume. Find its energy with (5.22) and with (5.23).

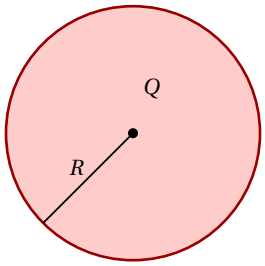


Figure 5.26 The uniformly charged ball of Example 5.12, in cross section.

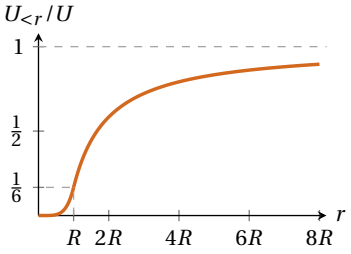
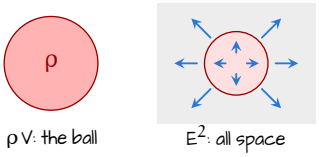


Figure 5.27 The fraction of the field energy of a uniformly charged ball that lies within a distance r of its center. Only one sixth of it is inside the ball; the other five sixths are outside, where there is no charge.

Solution:



Before computing, we sketch where each integral runs. The integral (5.22) covers only the ball, where the charge is; (5.23) covers all space, inside the ball and out, where the field is. The charge density is $\rho = 3Q/4\pi R^3$.

By Gauss's law, as in Exercise P4.5, the field is $Qr/4\pi\varepsilon_0 R^3$ inside the ball and $Q/4\pi\varepsilon_0 r^2$ outside it, along $\hat{\mathbf{r}}$. Integrating from the outside in, as in Example 5.8, the potential inside the ball is

$$V(r) = \frac{Q}{4\pi\varepsilon_0 R} - \int_R^r \frac{Q\bar{r}}{4\pi\varepsilon_0 R^3}\,d\bar{r} = \frac{Q}{8\pi\varepsilon_0 R}\left(3 - \frac{r^2}{R^2}\right).$$

(Why does the potential inside begin with the term $Q/4\pi\varepsilon_0 R$? Answer in footnote.[7])

From the charge. The integral (5.22) runs over the ball only, where the charge is:

$$U = \frac{1}{2}\int_0^R \frac{3Q}{4\pi R^3}\,\frac{Q}{8\pi\varepsilon_0 R}\left(3 - \frac{r^2}{R^2}\right)4\pi r^2\,dr = \frac{3Q^2}{16\pi\varepsilon_0 R^4}\int_0^R \left(3r^2 - \frac{r^4}{R^2}\right)dr$$
$$= \frac{3Q^2}{16\pi\varepsilon_0 R^4}\cdot\frac{4R^3}{5} = \frac{3}{5}\,\frac{Q^2}{4\pi\varepsilon_0 R}.$$

From the field. The integral (5.23) runs over all space, in two parts:

$$U = \frac{\varepsilon_0}{2}\left[\int_0^R \left(\frac{Qr}{4\pi\varepsilon_0 R^3}\right)^2 4\pi r^2\,dr + \int_R^\infty \left(\frac{Q}{4\pi\varepsilon_0 r^2}\right)^2 4\pi r^2\,dr\right]$$
$$= \frac{Q^2}{8\pi\varepsilon_0}\left[\frac{1}{R^6}\int_0^R r^4\,dr + \int_R^\infty \frac{dr}{r^2}\right]$$
$$= \frac{Q^2}{8\pi\varepsilon_0}\left[\frac{1}{5R} + \frac{1}{R}\right]$$
$$= \frac{3}{5}\,\frac{Q^2}{4\pi\varepsilon_0 R},$$

the same, $U = 3Q^2/20\pi\varepsilon_0 R$. The bracket shows where the field energy is: $1/5R$ of it comes from inside the ball and $1/R$ from outside, so only one sixth of the energy is in the ball, and five sixths are outside, where there is no charge at all (Figure 5.27).

Reality check: The example also checks the surface term. If the field integral is stopped at a sphere of radius $b > R$ instead of at infinity, it misses the field energy beyond b, which is $Q^2/8\pi\varepsilon_0 b$. On that sphere, $V = Q/4\pi\varepsilon_0 b$ and $E = Q/4\pi\varepsilon_0 b^2$, so the surface term is

$$\frac{\varepsilon_0}{2}\oint V\mathbf{E}\cdot d\mathbf{a} = \frac{\varepsilon_0}{2}\,\frac{Q}{4\pi\varepsilon_0 b}\,\frac{Q}{4\pi\varepsilon_0 b^2}\,4\pi b^2 = \frac{Q^2}{8\pi\varepsilon_0 b},$$

exactly the missing amount. As $b \to \infty$ it goes to zero, as the argument above said it must.

[7] It is the potential at the surface, $V(R)$, which the outer part of the path, from infinity to R, contributes. Out there the field is that of a point charge Q.

5.9.3 Self-energy

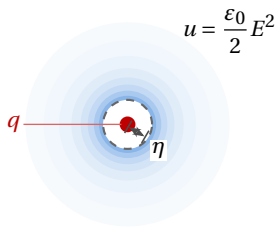


Figure 5.28 A point charge q and a small sphere of radius η around it. The energy density, $\varepsilon_0 E^2/2$, grows as $1/r^4$ toward the charge (shading), so the energy outside the sphere grows without limit as $\eta \to 0$.

Something goes wrong when (5.23) is applied to a point charge. The energy in the field of a point charge q, outside a small sphere of radius η around it (Figure 5.28), is

$$\frac{\varepsilon_0}{2}\int_\eta^\infty \left(\frac{q}{4\pi\varepsilon_0 r^2}\right)^2 4\pi r^2\,dr = \frac{q^2}{8\pi\varepsilon_0\eta},$$

and it grows without limit as $\eta \to 0$. The field energy of a point charge is infinite. In terms of work, it is the work it would take to squeeze a finite charge into a point, against its own repulsion. The point-charge formula (5.21) does not have this problem, because it leaves out the terms with $j = i$, which are exactly these **self-energies**. The continuous forms (5.22) and (5.23) include them. For a smooth, continuous distribution the self-energy of each infinitesimal piece is negligible, and the two kinds of formula agree; for point charges they do not, and (5.21) is the one to use. Whether a real electron has a size that keeps its self-energy finite is a question that classical electrodynamics cannot answer (Exercise P5.19).

A related point is that energy does not add up the way fields do. If $\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2$, then

$$E^2 = E_1^2 + E_2^2 + 2\,\mathbf{E}_1 \cdot \mathbf{E}_2,$$

so the energy of two charge distributions together is the sum of their separate energies plus a cross term, $\varepsilon_0 \int \mathbf{E}_1 \cdot \mathbf{E}_2\,d\tau$, which is their interaction energy. For two point charges the cross term is exactly the pair energy (5.13), while the other two terms are their infinite self-energies.

5.9.4 The energy of a capacitor

Charging a capacitor means moving charge from one conductor to the other. When the conductors carry $+q$ and $-q$, the potential difference between them is q/C, and moving a further charge dq from the negative conductor to the positive one takes the work $dW = (q/C)\,dq$. But the potential difference grows as the charging proceeds, so the work has to be integrated. Charging the capacitor from zero to Q takes

$$\begin{aligned} W &= \int_0^Q \frac{q}{C}\,dq \\ &= \frac{Q^2}{2C} \\ &= \frac{1}{2}C\Delta V^2 \\ &= \frac{1}{2}Q\Delta V. \end{aligned} \tag{5.25}$$

The last form also follows from (5.22): the charge $+Q$ sits at the potential V_+ and the charge $-Q$ at V_-, so

$$\frac{1}{2}\int \rho V\,d\tau = \frac{1}{2}(QV_+ - QV_-) = \frac{1}{2}Q\Delta V.$$

The field energy agrees, too. In a parallel-plate capacitor the field $E = Q/\varepsilon_0 A$ fills the volume Ad between the plates, so

$$\begin{aligned}\frac{\varepsilon_0}{2} E^2 \, Ad &= \frac{\varepsilon_0}{2} \frac{Q^2}{\varepsilon_0^2 A^2} \, Ad \\ &= \frac{Q^2 d}{2\varepsilon_0 A} \\ &= \frac{Q^2}{2C},\end{aligned}$$

with $C = \varepsilon_0 A/d$. The energy of a capacitor can be thought of as stored in the field between its plates.

5.10 Laplace's and Poisson's equations



Siméon Denis Poisson (1781–1840, French) Poisson entered the École Polytechnique in 1798 at the top of the entrance list, and Laplace became his mentor. He taught at the École from 1802, and in 1809 he became the first professor of mechanics in the new Faculty of Sciences in Paris. In 1812 he used the potential function to work out how charge distributes itself over conducting spheres, and in 1813 he showed that inside matter Laplace's equation must be replaced by the equation now named after him. Poisson's ratio in elasticity and the Poisson distribution in probability are named after him too; the distribution appears in his 1837 study of the probability of verdicts in criminal and civil trials. As a judge in the French Academy's 1818 competition on diffraction, he pointed out that Fresnel's wave theory predicted a bright spot at the center of a circular shadow, an apparently absurd result; Arago looked, and found the spot. According to Arago, Poisson used to say that life is good for only two things, to study mathematics and to teach it. (Wikipedia)

There is one more relation to find among ρ, $\mathbf{E}$, and V: the one that gives the charge density directly from the potential. Gauss's law gives ρ from $\mathbf{E}$, and $\mathbf{E} = -\nabla V$ gives $\mathbf{E}$ from V. Together,

$$\rho = \varepsilon_0 \nabla \cdot (-\nabla V) = -\varepsilon_0 \nabla^2 V,$$

or, with the minus sign moved to the other side, as it is usually written,

$$\nabla^2 V = -\frac{\rho}{\varepsilon_0}. \qquad \text{(Poisson's equation)} \quad (5.26)$$

In a region where there is no charge, it becomes

$$\nabla^2 V = 0. \qquad \text{(Laplace's equation)} \quad (5.27)$$

Poisson's equation is a single second-order differential equation for a single scalar function, and it contains both of the differential equations of electrostatics: the potential guarantees that $\nabla \times \mathbf{E} = 0$, and the equation itself is Gauss's law.

Read from left to right, (5.26) gives the charge density from the potential by differentiation. Read the other way, it is a differential equation to be solved for V, and the script-r integral (5.16) is the solution that goes to zero far away. To check, take the Laplacian of (5.16). The Laplacian acts on the field point $\mathbf{r}$, so it passes through the integral to act on $1/\mathscr{r}$, and by (1.46),

$$\begin{aligned}\nabla^2 V &= \frac{1}{4\pi\varepsilon_0} \int \rho(\mathbf{r}') \, \nabla^2 \left(\frac{1}{\mathscr{r}}\right) d\tau' \\ &= \frac{1}{4\pi\varepsilon_0} \int \rho(\mathbf{r}') \left[-4\pi\,\delta^3(\mathbf{r}-\mathbf{r}')\right] d\tau' \\ &= -\frac{\rho(\mathbf{r})}{\varepsilon_0}.\end{aligned}$$

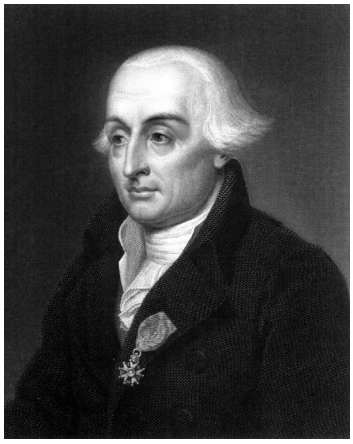
For a particular case, take the ball of Example 5.8. Inside it, $V = \rho_0 R^2/4\varepsilon_0 - \rho_0 r^4/20\varepsilon_0 R^2$, and with the spherical Laplacian (1.32),

$$\begin{aligned}\nabla^2 V &= \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial V}{\partial r}\right) + \cancel{\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial V}{\partial\theta}\right)} + \cancel{\frac{1}{r^2\sin^2\theta}\frac{\partial^2 V}{\partial\phi^2}} \\ &= \frac{1}{r^2}\frac{d}{dr}\left(-\frac{\rho_0 r^5}{5\varepsilon_0 R^2}\right) \\ &= -\frac{\rho_0 r^2}{\varepsilon_0 R^2} \\ &= -\frac{\rho}{\varepsilon_0}.\end{aligned}$$

Outside it, V is proportional to $1/r$, whose Laplacian is zero everywhere except at $r = 0$, so **Laplace's equation** holds there, as it should.

Laplace's equation is more useful than it might seem. It does not describe a world with no charges: the charges are usually there, on conductors at the edges of the region where the potential is wanted. Given the potential on those edges, Laplace's equation determines the potential everywhere inside, and Chapter 7 is about solving it. One property of its solutions has already appeared. Written out in Cartesian coordinates, Laplace's equation says that the three second derivatives of V add up to zero, which is what the saddle point at the center of the ring in Section 5.6 illustrated.

Laplace's equation is older than the electric potential. Lagrange introduced a potential function for gravitation in the 1770s, and Laplace showed in the 1780s that outside the attracting bodies it satisfies the equation now named after him, writing it first in spherical coordinates and then in Cartesian ones. In 1813 Laplace's protégé Poisson pointed out that inside the attracting matter the right side is not zero but proportional to the density.



Joseph-Louis Lagrange (1736–1813, Italian-French) Lagrange was born Giuseppe Lodovico Lagrangia in Turin, and at nineteen he was appointed professor at the royal artillery school there. The analytic method for the calculus of variations that he sent to Euler in 1755 made his name, and in 1766 he succeeded Euler as director of mathematics at the Berlin Academy. He moved to Paris in 1787. His *Mécanique analytique* (1788) reduced mechanics to general formulas, without a single diagram, and its equations of motion are the basis of the Lagrangian mechanics taught today. He served on the commission that created the metric system, taught analysis at the new École polytechnique, and is buried in the Panthéon. (Wikipedia)

5.11 Summary: Triangle diagram of ρ, E, V

The electrostatic field has no curl. That one fact makes the field the negative of the gradient of a scalar potential, and together with Gauss's law it makes the charge density, the field, and the potential three ways of describing the same thing. Each can be found from either of the others. The relations among them fit on a diagram that I call the **triangle diagram**:

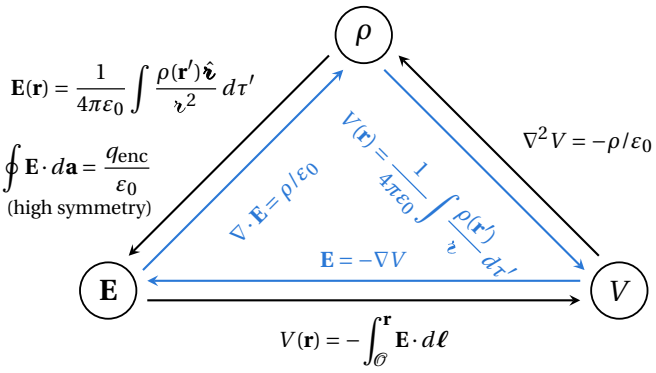


Figure 5.29 The triangle diagram. Each of ρ, $\mathbf{E}$, and V can be found from either of the others. The black arrows go counterclockwise around the outside of the triangle, and the blue arrows go clockwise around the inside.

If you know any one of the three, you can find the other two.

$\nabla\times\mathbf{E} = 0, \quad \oint \mathbf{E}\cdot d\boldsymbol{\ell} = 0$	electrostatic fields have no curl and no circulation
$\mathbf{E} = -\nabla V$	the field points downhill, from high potential to low
$V(\mathbf{r}) = -\int_{\mathscr{O}}^{\mathbf{r}} \mathbf{E}\cdot d\boldsymbol{\ell}$	the potential from the field; the reference point $\mathscr{O}$ is usually at infinity
$V(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\int \frac{\rho(\mathbf{r}')}{\imath}\,d\tau'$	the potential from the charge, with the reference point at infinity
$\nabla^2 V = -\frac{\rho}{\varepsilon_0}$	Poisson's equation; Laplace's equation, $\nabla^2 V = 0$, where there is no charge
$\mathbf{E}^{\parallel}_{\text{above}} = \mathbf{E}^{\parallel}_{\text{below}}, \quad V_{\text{above}} = V_{\text{below}}$	at a surface charge, the parallel field and the potential are continuous
$W = Q\Delta V, \quad U = QV$	the work to move a test charge, and its potential energy
$U = \frac{1}{2}\int \rho V\,d\tau = \frac{\varepsilon_0}{2}\int E^2\,d\tau$	the energy of a charge distribution; the energy density is $\frac{1}{2}\varepsilon_0 E^2$
$C = \frac{Q}{\Delta V}, \quad W = \frac{Q^2}{2C} = \frac{1}{2}C\Delta V^2$	capacitance, and the energy stored in a capacitor

Choosing a route. If the symmetry allows Gauss's law, find $\mathbf{E}$ first and integrate it to get V, from the outside in. If not, find V first with the script-r integral, which has no vectors in it, and take $-\nabla V$ if the field is wanted. When the integrals cannot be done by hand, the chunk-and-sum method works for V as it does for $\mathbf{E}$, and central differences give the field from the potential. Write $d\boldsymbol{\ell}$ in the form (1.12) and let the limits of integration set the direction of a path. To find a field from a potential, you need the potential near the point, not only at it: $V = 0$ at a point does not mean $\mathbf{E} = 0$ there.

Conductors and capacitors. A conductor in electrostatics is an equipotential. Its charge tends to crowd onto the most sharply curved parts of its surface. A capacitor is two conductors with charges $\pm Q$; its capacitance depends only on the geometry, and is best found by integrating the field from the negative conductor to the positive one.

Energy. The energy of a set of point charges is $\frac{1}{2}\sum q_i V(\mathbf{r}_i)$, with each V from the other charges only. For a continuous distribution it can be found from the charge, $\frac{1}{2}\int \rho V\, d\tau$, or from the field, $\frac{1}{2}\varepsilon_0 \int E^2\, d\tau$ over all space. The field energy of a point charge is infinite.

Results worth remembering. A point charge: $V = q/4\pi\varepsilon_0 r$. A uniformly charged spherical shell: $Q/4\pi\varepsilon_0 r$ outside, and the constant $Q/4\pi\varepsilon_0 R$ inside. An infinite line: $-(\lambda/2\pi\varepsilon_0)\ln(s/s_0)$, with the reference point at a distance s_0. Parallel plates: $C = \varepsilon_0 A/d$. An isolated sphere: $C = 4\pi\varepsilon_0 a$.

Exercises

Exercises for 5.1 Curl of Coulomb's law

Warm-up questions Read Section 5.1, then answer before class

W5.1 A classmate argues that a small paddle wheel placed near a point charge would turn after all: the field is stronger on the side of the wheel nearer the charge, so the blades on that side are pushed harder. What is wrong with the argument? Sketch the wheel, with its axle perpendicular to the line from the charge, and the push on each of four blades.

W5.2 The argument of this section used two facts about the field of a point charge: it points along $\hat{\mathbf{r}}$, and its strength depends only on r. (a) Suppose the field fell off as $1/r^3$ instead of $1/r^2$, so that the flux through a sphere around the charge would depend on the sphere's size (Section 4.2). Would $\int_{\mathbf{a}}^{\mathbf{b}} \mathbf{E}\cdot d\boldsymbol{\ell}$ still depend only on the endpoints? Would the curl still be zero? (b) Suppose instead that the field pointed along $\hat{\mathbf{r}}$ but its strength depended on θ as well as on r. Answer the same two questions.

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P5.1 Two vector fields, with k a constant, are

$$\mathbf{E}_a = k\left[(y+z)\,\hat{\mathbf{x}} + (x+z)\,\hat{\mathbf{y}} + (x+y)\,\hat{\mathbf{z}}\right],$$
$$\mathbf{E}_b = k\left[(y-z)\,\hat{\mathbf{x}} + (z-x)\,\hat{\mathbf{y}} + (x-y)\,\hat{\mathbf{z}}\right].$$

(a) Which of them could be an electrostatic field? Explain.

(b) For that field, find the potential $V(x,y,z)$, with the reference point at the origin. Check your answer by calculating $-\int \mathbf{E}\cdot d\boldsymbol{\ell}$ from the origin to the point $(1,2,3)$ along two paths: the straight line, and the curve $\mathbf{r}(t) = t^2\,\hat{\mathbf{x}} + 2t^3\,\hat{\mathbf{y}} + 3t\,\hat{\mathbf{z}}$, with t running from 0 to 1.

(c) What charge density produces this field? Sketch its equipotentials in the xy plane.

(d) For the other field, calculate $\int \mathbf{E}\cdot d\boldsymbol{\ell}$ from the origin to $(1,2,3)$ along the same two paths. Then calculate $\oint \mathbf{E}\cdot d\boldsymbol{\ell}$ around the unit circle in the xy plane, centered on the origin, going counterclockwise as seen from above, and check your answer with the curl theorem.

Exercises for 5.2 Boundary condition on $E^{\parallel}$

Warm-up questions Read Section 5.2, then answer before class

W5.3 Section 4.5 argued that the field just outside a conductor is perpendicular to its surface, since a parallel component would push the surface charges along the surface. Give a second argument, from the boundary condition of this section. A wire that carries a current has a field inside it, along the wire. What does the same boundary condition say about the field just outside such a wire?

W5.4 Each pair gives the field just below and just above a small patch of a surface, in kV/m, with $\hat{\mathbf{n}} = \hat{\mathbf{z}}$ pointing from below to above: (a) $3\hat{\mathbf{x}} + 4\hat{\mathbf{z}}$ below and $3\hat{\mathbf{x}} - 2\hat{\mathbf{z}}$ above; (b) $3\hat{\mathbf{x}} + 4\hat{\mathbf{z}}$ below and $3\hat{\mathbf{y}} + 4\hat{\mathbf{z}}$ above; (c) zero below and $5\hat{\mathbf{z}}$ above. Which of these could occur in electrostatics? For each one that could, find the surface charge density on the patch.

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

Exercises for 5.3 Scalar potential and its relation to electric field

Warm-up questions Read Section 5.3, then answer before class

W5.5 Suppose some field that is not electrostatic had $\oint \mathbf{E} \cdot d\boldsymbol{\ell} = 5\,\text{V}$ around a certain loop, traversed counterclockwise. (Chapter 15 meets such fields.) Try to define a potential for it anyway with (5.8), taking the reference point $\mathcal{O}$ on the loop. What value of V would you assign to $\mathcal{O}$ itself, along a path that goes once around the loop? Twice? Which of the four statements of Section 5.3.1 fail for this field?

W5.6 Outside the shell of Example 5.2, where are the equipotentials on which V is 1, $\frac{3}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$ times $Q/4\pi\varepsilon_0 R$? Sketch them. Are they equally spaced, and what does their spacing tell you about the field? Which points of space are at the first of these values?

W5.7 Two infinite lines of charge, parallel to the z axis, cross the xy plane at $(-a, 0)$, with the charge $+\lambda$ per unit length, and at $(a, 0)$, with $-\lambda$. Use Example 5.1 and superposition to find the potential at a point a distance s_+ from the positive line and s_- from the negative one, with the reference point the same distance s_0 from both lines. Does s_0 appear in your answer? Could the reference point now be at infinity, and why does the difficulty of a single line not arise?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

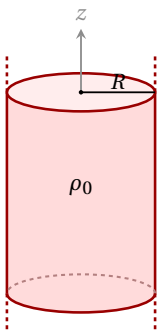


Figure 5.30 A piece of the cylinder of Exercise P5.2.

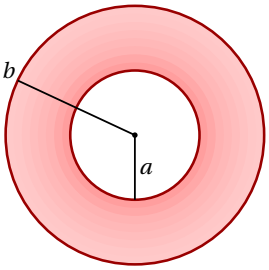


Figure 5.31 The shell of Exercise P5.3, in cross section, drawn for $b = 2a$. The shading darkens where the charge density is larger.

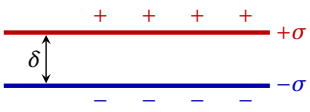


Figure 5.32 The dipole layer of part (b) of Exercise P5.4, edge on.

P5.2 An infinitely long cylinder of radius R, with its axis along the z axis, carries a uniform charge density ρ_0 (Figure 5.30).

(a) Use Gauss's law to find the electric field inside and outside the cylinder.

(b) Explain why the reference point for the potential cannot be put at infinity.

(c) Taking $V = 0$ on the axis, find the potential inside and outside the cylinder.

(d) Check that V and dV/ds are both continuous at $s = R$. Why should dV/ds be continuous there? Sketch $V(s)$.

P5.3 A thick spherical shell, with inner radius a and outer radius b, has the charge density $\rho = \kappa/r$ for $a < r < b$, where κ is a constant (Figure 5.31). There is no charge anywhere else.

(a) Find the electric field in the three regions $r < a$, $a < r < b$, and $r > b$.

(b) Find the potential in the three regions, with $V(\infty) = 0$. Work from the outside in.

(c) Plot E and V against r for $b = 2a$. Are they continuous? Should they be?

P5.4 This exercise looks at the one situation in which the potential is not continuous.

(a) Use the line integral of $\mathbf{E}$ across a surface that carries a surface charge density σ to show that V is continuous across it. Then use (4.6) to show that the derivative of V along the normal to the surface jumps by $-\sigma/\varepsilon_0$.

(b) A *dipole layer* is a pair of parallel sheets, with surface charge densities $+\sigma$ and $-\sigma$, a distance δ apart (Figure 5.32). Let $\delta \to 0$ while the product $\tau = \sigma\delta$ stays fixed. Show that the potential jumps by τ/ε_0 across the layer, and say which side is at the higher potential.

(c) Why does the argument of part (a) fail for the dipole layer?

Exercises for 5.4 Potential of a point charge, relationship to potential energy

Warm-up questions Read Section 5.4, then answer before class

W5.8 A classmate says that a charge released from rest always moves toward lower potential, like a ball rolling downhill. Is that right for a proton? For an electron? Does the electron's potential energy rise or fall?

W5.9 A classmate finds the potential of a positive point charge q by integrating from infinity with $d\boldsymbol{\ell} = -d\bar{r}\,\hat{\mathbf{r}}$, and gets $V = -q/4\pi\varepsilon_0 r$. Without looking for the slip in the calculation, give two separate reasons why this result cannot be right.

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P5.5 In nuclear physics it is convenient to measure distances in femtometers ($1\,\text{fm} = 10^{-15}\,\text{m}$) and energies in MeV. In those units,

$$\frac{e^2}{4\pi\varepsilon_0} = 1.440\,\text{MeV fm}.$$

Treat nuclei as point charges as long as they do not touch, and take the radius of a nucleus with A protons and neutrons to be $(1.2\,\text{fm})\,A^{1/3}$.

(a) Check the value of $e^2/4\pi\varepsilon_0$ from the values of e and ε_0.

(b) An alpha particle (charge $2e$, $A = 4$) with a kinetic energy of 7.7 MeV heads straight for a gold nucleus (charge $79e$, $A = 197$), which you may treat as fixed. How close does the alpha particle come before it stops and turns back? Does it touch the nucleus?

(c) A uranium-238 nucleus decays by emitting an alpha particle, leaving behind a thorium-234 nucleus (charge $90e$). Find the electrostatic potential energy of the alpha particle and the thorium nucleus when they are just touching.

(d) The alpha particles from uranium-238 come out with a kinetic energy of only about 4.2 MeV. Outside the nucleus, the only force on an alpha particle is the electrostatic repulsion of the thorium nucleus. At what distance from the thorium nucleus must the alpha particle have been at rest to end up with 4.2 MeV? Why is that a puzzle for classical physics? (Quantum mechanics resolves it: the alpha particle *tunnels* through the region where its kinetic energy would be negative.)

P5.6 Common salt, NaCl, is a crystal of Na^+ and Cl^- ions on alternating sites of a cubic lattice (Figure 5.33). Put an ion at every point $(i, j, k)\,r_0$,

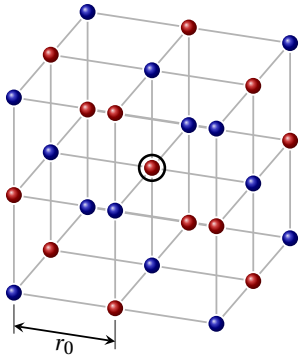


Figure 5.33 The cube of ions around one sodium ion (circled) in Exercise P5.6, for $n = 1$: Na^+ red, Cl^- blue.

where i, j, and k are integers and $r_0 = 0.282\,\text{nm}$ is the distance between neighboring ions, with charge $+e$ where $i+j+k$ is even and $-e$ where it is odd. The potential at a Na^+ site due to all of the other ions can be written $V = -Me/4\pi\varepsilon_0 r_0$, where the number M is called the *Madelung constant*. No closed form for M is known, so it has to be calculated numerically.

(a) Write a program that adds up the potentials (5.11) of the ions in a cube of $(2n+1)^3$ sites centered on a Na^+ ion (leaving out the ion itself), and plot the resulting value of M against n for $n = 1$ to 20. Does the sum settle down?

(b) The trouble is that each cube has a net charge on its surface layers. Repeat the calculation with the charges of the ions on the faces of the cube multiplied by $\frac{1}{2}$, those on its edges by $\frac{1}{4}$, and those at its corners by $\frac{1}{8}$, which makes each cube neutral. (This is *Evjen's method*.) How large must n be for M to be correct to four significant figures?

(c) The *lattice energy* of a crystal is the energy needed to take it apart into isolated ions. Use (5.21) (Section 5.9.1) to show that, treating the ions as point charges, the lattice energy is $Me^2/4\pi\varepsilon_0 r_0$ per pair of ions, and evaluate it in electron volts. Look up the measured lattice energy of NaCl and compare. What might account for the difference?

Exercises for 5.5 Solving for V when given ρ: script-r method

Warm-up questions Read Section 5.5, then answer before class

W5.10 Remove one side of the square loop of Example 5.6, leaving three sides with the same λ. What is the potential at a point on the z axis? (No new integral is needed.) A classmate then finds the field on the axis as $-(\partial V/\partial z)\hat{\mathbf{z}}$. What has the classmate left out, and which way does it point?

W5.11 Would you find the field first or the potential first in each of these cases? Give a one-sentence reason for each. (a) The potential inside and outside an infinite slab, $-a < z < a$, with uniform charge density ρ. (b) The potential at the center of a thin hemispherical shell of radius R with uniform surface charge density σ. (c) The potential at points on the axis of a thin disk of radius R with uniform σ.

W5.12 In Example 5.7, suppose you had written $\sqrt{(r-R)^2} = r - R$ inside the shell as well. What potential would you have found inside? Without using the example's answer, give two reasons why it cannot be right. Would checking that V is continuous at $r = R$ have caught the mistake?

Also: without looking back, summarize this section in a few sentences, and say what

you found hardest to understand.

P5.7 A straight line segment along the z axis, from z_1 to z_2, carries a uniform line charge density λ (Figure 5.34).

(a) Use (5.14) to find the potential at a point in the xy plane a distance s from the z axis.

(b) Now take $z_1 = -L$ and $z_2 = L$. Show that the potential at any fixed s grows without limit as $L \to \infty$, but that the difference $V(s_1) - V(s_2)$ between two points at distances s_1 and s_2 from the axis approaches a finite limit. Compare that limit with Example 5.1.

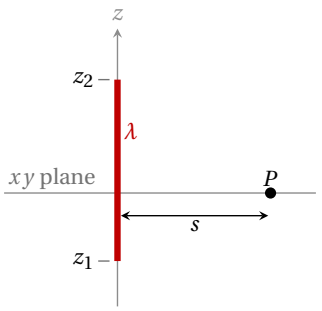


Figure 5.34 The line segment of Exercise P5.7 and the point P, seen from the side.

P5.8 A line segment along the x axis, from $x = 0$ to $x = L$, carries a uniform line charge density λ (Figure 5.35).

(a) Without doing any calculation, write down an approximate expression for the potential at the point x_0 on the x axis, for $x_0 \gg L$.

(b) Find the potential at x_0 exactly, for any $x_0 > L$.

(c) Expand your answer to part (b) in powers of L/x_0. In the point-charge formula $\lambda L/4\pi\varepsilon_0 D$, which choice of the distance D — x_0, $x_0 - L/2$, or $x_0 - L$ — gives the best approximation, and how large is its fractional error?



Figure 5.35 The line segment of Exercise P5.8 and the distant point x_0.

P5.9 A square of side a lies in the xy plane, centered on the origin, with its sides parallel to the axes (Figure 5.36). Point charges $+q$ sit at its two corners with $y = a/2$, and point charges $-q$ at its two corners with $y = -a/2$. The point P is on the x axis, at $(d, 0, 0)$.

(a) Find the electric field at P.

(b) Find the potential at P.

(c) Your answers should give $V = 0$ but $\mathbf{E} \neq 0$ at P. How can that be consistent with $\mathbf{E} = -\nabla V$?

(d) Find the potential at the point $(d, y, 0)$, and use it to find the field at P as $-\partial V/\partial y$. Check that it agrees with part (a).

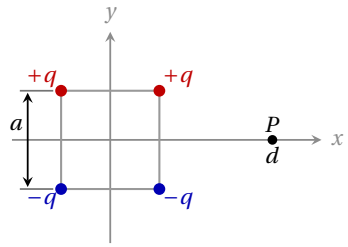


Figure 5.36 The charges of Exercise P5.9 at the corners of the square, and the point P.

Exercises for 5.6 Numerical method: direct integration

Warm-up questions Read Section 5.6, then answer before class

W5.13 Taking $E_s = -144.421\,\text{V/m}$ as the right answer, find the error of the central-difference value at $h = 1\,\text{mm}$ in Table 5.2. If the error is proportional to h^2, what should it be at $h = 5\,\text{mm}$? Does the table agree? Why can h not be made as small as you like?

W5.14 This section's last tip says that V is finite at a point on a surface or volume charge but infinite on a line charge. To see why, find V at such a point due only to the charge within a small distance η of it, for uniform volume, surface, and line charges (treat the surface as flat and the line as straight, that close). Which go to zero as $\eta \to 0$?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P5.10 In Exercise P3.7 you wrote a program that calculates the field of a model of a DNA molecule by adding the fields of its phosphate charges. Extend it to calculate the potential, with $V = 0$ at infinity, by adding the potentials (5.11) of the charges.

(a) For $N = 200$ base pairs, calculate V at the 72 points of part (b) of that exercise, at $s = 1.2\,\text{nm}$, and plot V against ϕ. By what fraction does V vary around the molecule? Compare with the variation of E_s that you found there, and explain the difference.

(b) At a fixed ϕ, calculate V at $s = 2\,\text{nm}$ and at $s = 5\,\text{nm}$ for $N = 100$, 200, and 400. Show that V itself depends strongly on N, but the difference $V(5\,\text{nm}) - V(2\,\text{nm})$ hardly depends on N at all, and compare that difference with the one for an infinite uniform line charge with $\lambda = -2e/h$ (Example 5.1). Explain both observations.

(c) At the point of part (a) where V is most negative, calculate E_s by a central difference (5.18), for several values of the step (call it δ here, since h is the rise per base pair), and compare with the field your program from Exercise P3.7 gives directly.

(d) How much work would it take to move a sodium ion (charge $+e$) from the point of part (c) straight out to $s = 5\,\text{nm}$? Give your answer in electron volts, and compare it with the typical thermal energy of a molecule at room temperature, about 0.025 eV. (In water the answer is much smaller, partly because of the water's own response to the field, which Chapter 9 describes, and partly because of the ions of the salt solution around the molecule.)

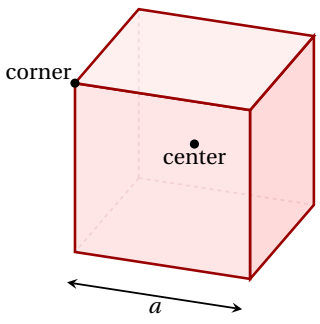


Figure 5.37 The charged cube of Exercise P5.11, with its center and one corner marked.

P5.11 A cube of side a is filled with charge of uniform density ρ (Figure 5.37).

(a) Use the chunk-and-sum method (5.17) to calculate the potential at the center of the cube and at one of its corners, in units of $\rho a^2/\varepsilon_0$. For the center, use an even number of divisions along each edge. Why?

(b) The potential at a corner of a uniformly charged cube depends on the size L of the cube as L^2. Explain why, and use this to show that the potential at the center of a cube is exactly twice the potential at its corner. (Hint: a cube of side a can be cut into eight cubes of side $a/2$, each of which has a corner at the center of the big cube.) Check your numerical results against this. How does the error in your results depend on the number of divisions?

(c) Compare your results with the potential at the center and on the surface of a ball of the same volume and charge density.

Exercises for 5.7 Conductors as equipotentials

Warm-up questions Read Section 5.7, then answer before class

W5.15 A conducting cube is given a charge Q. A classmate says the charge must spread uniformly over the six faces, since by symmetry no face is different from any other. What is wrong with the argument? If the charge were spread uniformly, would the potential at the center of a face be higher or lower than at a corner? (Compare how much of the charge lies close to each point.) Which way would the charge then move?

W5.16 Two distant conducting spheres, of radii R and $3R$, each carry the charge $Q/2$. They are then joined by a long, thin wire. A classmate says that no charge will flow, since the two charges are already equal. Is that right? If charge flows, which way does it go, and how much ends up on each sphere?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P5.12 A conducting sphere of radius a, carrying a charge $+q$, sits at the center of a thin conducting spherical shell of radius b, which carries a charge $-2q$ (Figure 5.38).

(a) Find the charge on the surface of the sphere and on the inner and outer surfaces of the shell, and find the potential everywhere, with $V(\infty) = 0$. Which conductor is at the higher potential?

(b) The sphere is now connected to the shell by a thin wire. Where does the charge go? Find the potential everywhere, and the change in the electrostatic energy of the system (Section 5.9).

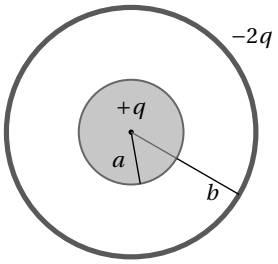


Figure 5.38 The sphere and the shell of Exercise P5.12, in cross section.

(c) In a Van de Graaff generator, a moving belt carries charge into the terminal, a large, hollow, conducting sphere, where a contact on the inside transfers the charge to the sphere. Use your answer to part (b) to explain why charge can be added to the terminal in this way no matter how much charge it already has.

Exercises for 5.8 Capacitance

Warm-up questions Read Section 5.8, then answer before class

W5.17 A classmate reads $C = Q/\Delta V$ and concludes that a capacitor carrying twice as much charge has twice the capacitance. What is wrong? Use the doubling argument of this section to explain why Q is proportional to ΔV for any pair of conductors, whatever their shapes.

W5.18 A classmate says that a battery fills a capacitor with charge, the way a pump fills a tire with air. What is the total charge on the two conductors before and after the capacitor is charged? What, then, is the Q in $C = Q/\Delta V$, and in what sense is a charged capacitor "charged"?

W5.19 A thin, uncharged metal sheet of thickness t, as large as the plates, is slid into the gap of the capacitor of Example 5.10, parallel to the plates and touching neither. Sketch the charges on the sheet and the field in the gap. Does the capacitance increase, decrease, or stay the same? Does it matter where in the gap the sheet is placed?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

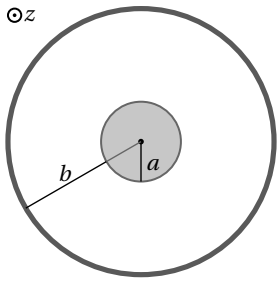


Figure 5.39 The coaxial cable of Exercise P5.13, end on.

P5.13 A coaxial cable consists of a long, straight inner conductor of radius a, surrounded by a thin cylindrical outer conductor of radius b, with air between them (Figure 5.39).

(a) Find the capacitance per unit length.

(b) Evaluate it for $a = 0.50\,\text{mm}$ and $b = 2.0\,\text{mm}$.

(c) With a potential difference of 100 V between the conductors, find the energy stored per meter of cable (Section 5.9), and the field at the surface of the inner conductor.

(d) What ratio b/a gives a capacitance of $100\,\text{pF/m}$?

(e) Suppose that the outer radius and the potential difference are fixed. Show that the field at the surface of the inner conductor is smallest when $b/a = e$.

P5.14 The ionosphere is a layer of the atmosphere, about 100 km above the ground, that conducts electricity because ultraviolet light from the Sun has ionized its atoms. The ground and the ionosphere can be modeled as two concentric conducting spheres, of radii $r_1 = 6400\,\text{km}$ (the ground) and $r_2 = 6500\,\text{km}$ (the ionosphere), with a potential difference of 300 kV between them. The ground is negatively charged and the ionosphere positively charged, with equal and opposite charges.

(a) What is the capacitance of the system?

(b) What is the charge Q of the ionosphere, and the surface charge density of the ground?

(c) How much energy is stored in the field between the two spheres (Section 5.9)?

(d) Find the field at the ground in this model, and compare it with the fair-weather field of about 100 V/m of Exercise P4.14. What does the comparison suggest about the model?

(e) Find the capacitance from the parallel-plate formula, with $A = 4\pi r_1^2$ and $d = r_2 - r_1$, and compare. What area would make the parallel-plate formula exact?

P5.15 Two flat plates, each of width w and length L, face each other, but they are not quite parallel: the gap between them is d_1 along one of the edges of length w and d_2 along the opposite edge, and in between it varies linearly (Figure 5.40). Both gaps are small compared with w and L.

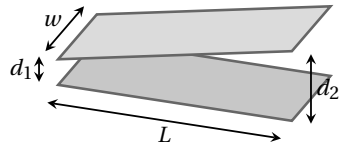


Figure 5.40 The plates of Exercise P5.15, with the gaps much exaggerated.

(a) Divide the plates into thin strips parallel to the edges of length w. Each pair of facing strips acts like a small parallel-plate capacitor. Are these small capacitors in series or in parallel? (Capacitors in parallel have the same potential difference, and their charges add.) Find the capacitance of the plates.

(b) Check your answer in the limit $d_2 \to d_1$.

(c) For $wL = 100\,\text{cm}^2$, $d_1 = 1.0\,\text{mm}$, and $d_2 = 2.0\,\text{mm}$, compare your answer with the parallel-plate formula using the average gap.

Exercises for 5.9 Energy stored in the field

Warm-up questions Read Section 5.9, then answer before class

W5.20 A test charge Q at a point where the potential is V has the potential energy QV (Section 5.4.1), but the energy of a distribution of charge is $\frac{1}{2}\int \rho V \, d\tau$. Where does the $\frac{1}{2}$ come from, and why does QV not need it? Use two point charges to explain.

W5.21 When a charge $+q$ and a charge $-q$ are brought in from far apart to a distance d apart, the energy stored in their field changes by the pair energy (5.13). Does it increase or decrease? How can it decrease, when the energy density $\frac{1}{2}\varepsilon_0 E^2$ is never negative? Where does the field become stronger as the charges come together, and where weaker?

W5.22 A battery of voltage ΔV charges an uncharged capacitor. It moves the whole charge Q from one conductor to the other through its full voltage, so it does the work $Q\Delta V$, but (5.25) says that the capacitor stores only $\frac{1}{2}Q\Delta V$. Where does the other half go? Would it be different with wires of smaller resistance? (Does the resistance appear in either the battery's work or the stored energy?)

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

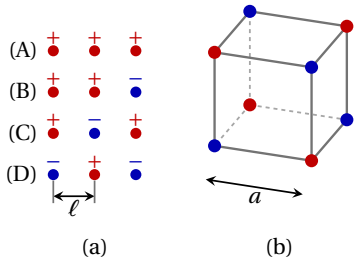


Figure 5.41 The arrangements of Exercise P5.16: (a) the four rows of part (a); (b) the cube of part (b). Red is $+q$ and blue is $-q$.

P5.16 This exercise compares the work needed to assemble a few arrangements of point charges, each brought in from infinity.

(a) Three point charges are placed in a row, a distance ℓ apart, in each of four arrangements (Figure 5.41a): (A) $+q$, $+q$, $+q$; (B) $+q$, $+q$, $-q$; (C) $+q$, $-q$, $+q$; and (D) $-q$, $+q$, $-q$. Without a calculator, rank the four arrangements by the work needed to assemble them, from least to greatest.

(b) Eight point charges, alternately $+q$ and $-q$, sit at the corners of a cube of side a (Figure 5.41b), so that the three nearest neighbors of each charge all have the opposite sign, as in a crystal of salt. Find the work needed to assemble them, and compare it with the work needed to assemble eight charges $+q$ at the same corners.

P5.17 A thin spherical shell of radius R carries a total charge Q, spread uniformly over its surface. Find its energy in three ways:

(a) with $\frac{1}{2}\int \sigma V \, da$;

(b) with $\frac{1}{2}\varepsilon_0 \int E^2 \, d\tau$;

(c) by building it up: find the work needed to bring a small charge dq in from infinity and spread it over the shell when the shell already carries a charge q, and integrate.

(d) Compare your answer with the energy of a uniformly charged ball of the same radius and charge (Example 5.12). Why is the shell's energy smaller?

P5.18 A ball of radius R has the charge density $\rho = \rho_0 (r/R)^n$, as in Exercise P4.5. Take $n \geq 0$ in parts (a)–(c).

(a) Use the field you found in that exercise to find the potential inside and outside the ball.

(b) Find the energy of the ball with (5.22) and with (5.23), and check that they agree.

(c) Write the energy in terms of the total charge Q of the ball, and check it against Example 5.12 for $n = 0$. How does the energy change as n increases, for fixed Q and R? What does it approach as $n \to \infty$, and why?

(d) The field of Exercise P4.5 is valid for any $n > -3$. For which of these values of n is the energy finite?

P5.19 The field energy of a point charge is infinite, which has led to the suggestion that the electron is not a point.

(a) Model the electron as a thin spherical shell of radius R, with its charge $-e$ spread uniformly over it. Find R by setting the energy stored in its field equal to the electron's rest energy, mc^2. Give both a symbolic and a numerical answer.

(b) Experiments that scatter electrons from other particles at high energies have found no sign of any size at all, down to distances far smaller than your answer to part (a). What does that suggest about the model?

P5.20 A nucleus with A protons and neutrons can be modeled as a uniformly charged ball of radius $R = (1.2\,\text{fm})\, A^{1/3}$ (Figure 5.42). Use the value of $e^2/4\pi\varepsilon_0$ in MeV fm from Exercise P5.5.

(a) Use Example 5.12 to find the electrostatic energy of a uranium-238 nucleus ($Z = 92$, $A = 238$), in MeV.

(b) Suppose the nucleus splits into two equal fragments, each with $Z = 46$ and $A = 119$. Find the electrostatic energy of each fragment, and the electrostatic energy released by the split once the fragments are far apart.

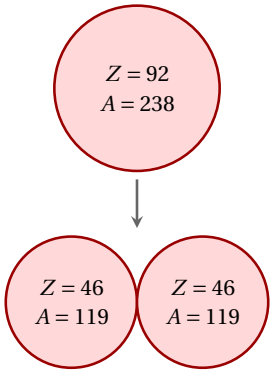


Figure 5.42 The uranium-238 nucleus of Exercise P5.20 as a uniformly charged ball, and the two fragments of parts (b) and (c), just touching. Drawn to scale.

(c) Find the electrostatic interaction energy of the two fragments when they are just touching, treating each one as a point charge at its center.

(d) Which of your energies would you expect to end up as the kinetic energy of the two fragments as they fly apart? Explain.

P5.21 A parallel-plate capacitor has plates of area A with air between them. The plates are pulled slowly apart, from a separation d_1 to a separation d_2.

(a) The plates are connected to a battery that keeps their potential difference at ΔV. Find the change in the energy stored in the capacitor. Does it increase or decrease?

(b) Charge flows through the battery during the process. Find the work done on the battery, and use conservation of energy to find the work that you do. Is it positive?

(c) Now the battery is disconnected before the plates are pulled apart, so that the charge Q on the plates stays fixed. Find the change in the stored energy and the work that you do, and from these, the force between the plates.

(d) Show that the force per unit area on each plate in part (c) is equal both to the energy density $\frac{1}{2}\epsilon_0 E^2$ between the plates and to the electrostatic stress (4.7). Explain, in terms of field energy, why the force per unit area equals the energy density.

Exercises for 5.10 Laplace's and Poisson's equations

Warm-up questions Read Section 5.10, then answer before class

W5.23 A classmate's program for the potential of a few point charges reports, at a point in empty space, $\partial^2 V/\partial x^2 = 2.0$, $\partial^2 V/\partial y^2 = 3.0$, and $\partial^2 V/\partial z^2 = 1.0$, all in $\mathrm{V/m^2}$. What do you conclude? What would you conclude if the point were instead inside a region with a continuous charge density?

W5.24 An electron sits at the center of a uniformly charged positive ball, like the ball of Example 5.12. Is it in stable equilibrium? (Look at how the potential curves near the center.) Why does this not contradict Earnshaw's theorem (Section 5.6)?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P5.22 For the ball of Exercise P5.18, with $\rho = \rho_0(r/R)^n$ and $n \geq 0$, find the potential by solving Poisson's equation directly.

(a) Inside the ball, write Poisson's equation in spherical coordinates for a potential that depends only on r, and integrate it twice.

(b) Do the same outside the ball, where it becomes Laplace's equation.

(c) There are four constants of integration. Find them from the conditions that V is finite at the center, that $V \to 0$ at infinity, and that V and dV/dr are both continuous at $r = R$. Why must dV/dr be continuous there? Compare your result with Exercise P5.18.

P5.23 Two concentric conducting spheres, of radii a and $b > a$, are held at potentials V_a and V_b (Figure 5.43).

(a) Solve Laplace's equation for the potential between the spheres, assuming that it depends only on r.

(b) Find the charge on the inner sphere from the field at its surface, and use it to find the capacitance, $C = Q/(V_a - V_b)$. Compare with Example 5.11.

(c) For $V_a = 100\,\text{V}$, $V_b = 0$, $a = 1.0\,\text{cm}$, and $b = 2.0\,\text{cm}$, find the potential halfway between the spheres, at $r = 1.5\,\text{cm}$. Is it the average of V_a and V_b? Explain why or why not.

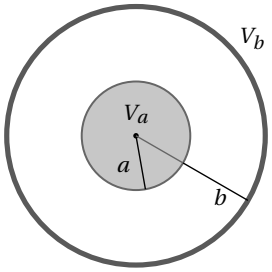


Figure 5.43 The two spheres of Exercise P5.23, in cross section.

Exercises for 5.11 Summary: Triangle diagram of ρ, $\mathbf{E}$, V

P5.24 Everywhere in space, the potential is $V = V_0\, e^{-r^2/a^2}$, where V_0 and a are constants. This exercise goes around the triangle diagram.

(a) Find $\mathbf{E}$ from V.

(b) Find ρ from $\mathbf{E}$ with Gauss's law, and again from V with Poisson's equation. Where is ρ positive, and where is it negative?

(c) Find the charge enclosed by a sphere of radius r centered on the origin, from the integral form of Gauss's law. What is the total charge?

(d) Integrate $\mathbf{E}$ from infinity to get V back.

(e) Use the script-r integral (5.16) with your ρ to find the potential at the origin, and check that it is V_0.

P5.25 True or false? Explain your reasoning.

(a) If $\mathbf{E} = 0$ throughout a region, then $V = 0$ there.

(b) If $V = 0$ throughout a region, then $\mathbf{E} = 0$ there.

(c) For a single point charge at rest, $\nabla \times \mathbf{E} = 0$ everywhere, even at the charge.

(d) On a clear day the electric field in the open air just above flat ground is nearly uniform, about 100 V/m pointing down (Exercise P4.14). So a point 2 m above the ground is at a potential about 200 V higher than the ground.

(e) To find V from $\mathbf{E}$ with a line integral, the path must start at the origin.

(f) Field lines are perpendicular to equipotential surfaces wherever $\mathbf{E} \neq 0$.

(g) Inside a positively charged conducting sphere, $V = 0$.

(h) A thin spherical shell carries surface charge that is positive on some parts of it and negative on others, with a total charge of zero. Then $V = 0$ at the center of the shell.

(i) For two point charges, $\frac{1}{2}\varepsilon_0 \int E^2 \, d\tau$ over all space is $q_1 q_2 / 4\pi\varepsilon_0 r_{12}$.

(j) An arrangement of static charges can hold a positive point charge in stable equilibrium at a point in empty space.

P5.26 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as the equipotentials and field lines of a group of point charges, the paths of alpha particles scattered by a gold nucleus, the potential of a charged disk at points off its axis, the crowding of charge onto the sharp points of a conductor, the capacitance of two parallel plates of finite size, or the energy needed to assemble a small crystal of ions.

Chapter 6

Mathematics Review 2: Orthogonal Functions

Separating variables in the partial differential equations of electromagnetism produces the same few families of functions again and again. Laplace's equation gives sines and cosines in Cartesian coordinates and Legendre polynomials in spherical coordinates. The wave equation in cylindrical coordinates, which you will meet in the study of waveguides (Section 23.5), gives Bessel functions. These look like three unrelated pieces of mathematics, but all three share one property, and that property is what makes them useful.

The property is **orthogonality**. Loosely, two functions are orthogonal if multiplying them together and integrating gives zero. The word is borrowed deliberately from vectors, where two vectors are orthogonal if their dot product vanishes, and the analogy is closer than it first appears. In both cases, orthogonality is what lets you pick out one component of something built from many pieces: you calculate the *projection* of the whole onto one piece, and every other piece contributes nothing to that projection (Figure 6.1).

That procedure, which works the same way for all three families, is the subject of this chapter.

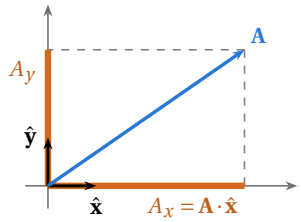


Figure 6.1 A vector built from orthogonal pieces, $\mathbf{A} = A_x\hat{\mathbf{x}} + A_y\hat{\mathbf{y}}$. The dot product with $\hat{\mathbf{x}}$ picks out A_x alone, because $\hat{\mathbf{x}}\cdot\hat{\mathbf{y}} = 0$. Fourier's trick (Section 6.1.3) does the same thing with functions.

6.1 Sines and cosines: Fourier series

6.1.1 The series

Any reasonably well-behaved periodic function with period T can be written as a sum of sines and cosines,

$$f(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos\left(\frac{2\pi n t}{T}\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{2\pi n t}{T}\right). \tag{6.1}$$

The numbers a_n and b_n are the **Fourier coefficients**. Each one says how much of that particular sine or cosine is present in the original function.

It is usually tidier to work in terms of the **fundamental frequency**

$$\omega_0 = \frac{2\pi}{T}, \tag{6.2}$$

because then the series makes its most important structural feature obvious:

$$f(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos(n\omega_0 t) + \sum_{n=1}^{\infty} b_n \sin(n\omega_0 t). \tag{6.3}$$

The frequency of every term in the series is a whole-number multiple of the lowest frequency. A periodic function does not contain arbitrary frequencies — it contains ω_0, $2\omega_0$, $3\omega_0$, and so on, and nothing in between.

The coefficients are given by

$$a_0 = \frac{1}{T}\int_0^T f(t)\,dt, \tag{6.4}$$

$$a_n = \frac{2}{T}\int_0^T f(t)\cos(n\omega_0 t)\,dt, \tag{6.5}$$

$$b_n = \frac{2}{T}\int_0^T f(t)\sin(n\omega_0 t)\,dt, \tag{6.6}$$

with a_0 being simply the average value of the function over one period. The integrals may be taken over any full period — from $-T/2$ to $T/2$, or anywhere else convenient — since a periodic function integrated over any complete period gives the same answer.

☞ In this book t and T are used because Fourier series are most familiar in time. Everything here works identically in space: replace t by x, the period T by a spatial period L, and ω (radians per second) by k (radians per meter). That spatial version is the one you will actually use when we come to boundary-value problems.

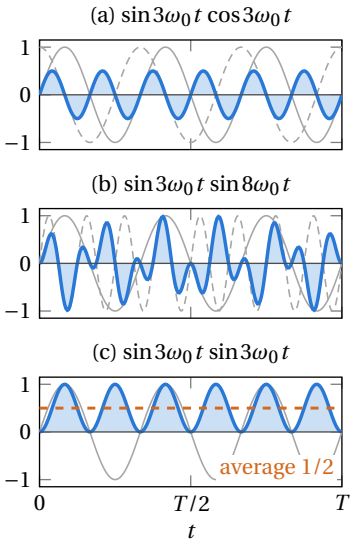


Figure 6.2 Why the orthogonality relations hold. Each panel shows two of the functions (thin gray curves, the second one dashed) and their product (blue, shaded to the axis) over one period. (a) A sine and a cosine of the same frequency: the product, $\frac{1}{2}\sin 6\omega_0 t$, has as much area below the axis as above it, so its integral is zero. (b) Two sines of different frequencies: again the areas cancel. (c) A sine times itself: the product is never negative, so nothing cancels. Its average value is $\frac{1}{2}$, so its integral over a period is $T/2$.

6.1.2 Orthogonality

Where do (6.4)–(6.6) come from? From the orthogonality of sines and cosines, which is worth stating precisely because everything else follows from it.

Sines and cosines whose frequencies are multiples of ω_0 obey three relations, for positive integers m and n. A sine and a cosine are *always* orthogonal, whatever the two integers:

$$\int_0^T \sin(n\omega_0 t)\cos(m\omega_0 t)\,dt = 0 \qquad \text{for all } m, n. \tag{6.7}$$

Two sines, or two cosines, are orthogonal provided the integers differ — and when they do not differ, the integral gives $T/2$:

$$\int_0^T \sin(n\omega_0 t)\sin(m\omega_0 t)\,dt = \begin{cases} 0 & m \neq n \\ T/2 & m = n \end{cases}, \tag{6.8}$$

$$\int_0^T \cos(n\omega_0 t)\cos(m\omega_0 t)\,dt = \begin{cases} 0 & m \neq n \\ T/2 & m = n \end{cases}. \tag{6.9}$$

You can check any of these by direct integration, and it is worth doing at least one. But the reason they are true is easier to see in a picture than in an antiderivative (Figure 6.2): the product of two different such functions encloses as much area above the axis as below it, so the positive and negative areas cancel. The single exception is a function multiplied by itself, where the product is never negative and no cancellation can occur.

6.1.3 Fourier's trick

With the orthogonality relations in hand, the coefficient formulas can be derived rather than merely asserted. The method is sometimes called **Fourier's trick**,[1] and it is the single most important technique in this chapter — you will use it again for Legendre polynomials, and again for Bessel functions, with only the details changed.



Joseph Fourier (1768–1830, French) Fourier introduced the series that bear his name in a memoir on the conduction of heat, read to the Institute in Paris in 1807. Lagrange and Laplace doubted that an arbitrary function could be expanded in sines and cosines, but a revised version won the Institute's prize competition of 1811, and his *Théorie analytique de la chaleur* appeared in 1822. The son of a tailor in Auxerre, he was orphaned at ten and educated at the town's military school. During the Revolution he was arrested in 1794 and released after the fall of Robespierre; he then taught at the new École polytechnique, succeeding Lagrange in 1797. He went with Napoleon to Egypt in 1798, as secretary of the Institute in Cairo, and did his work on heat while prefect of the department of Isère, at Grenoble. In 1822 he became perpetual secretary of the Académie des sciences. (Wikipedia)

Deriving the coefficient formulas

We will do the b_n coefficients; the a_n derivation runs the same way with cosines.

Start with the series itself:

$$f(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos(n\omega_0 t) + \sum_{n=1}^{\infty} b_n \sin(n\omega_0 t).$$

Multiply both sides by $\sin(m\omega_0 t)$, for some particular integer m, and integrate over a full period. The summation and the integral can be interchanged, since one acts on n and the other on t:

$$\begin{aligned}\int_0^T f(t)\sin(m\omega_0 t)\,dt = &\int_0^T a_0 \sin(m\omega_0 t)\,dt \\ &+ \sum_{n=1}^{\infty} a_n \int_0^T \cos(n\omega_0 t)\sin(m\omega_0 t)\,dt \\ &+ \sum_{n=1}^{\infty} b_n \int_0^T \sin(n\omega_0 t)\sin(m\omega_0 t)\,dt.\end{aligned}$$

Now take the three terms on the right one at a time.

The first term vanishes. The constant a_0 comes out of the integral, and a sine integrated over a whole number of its own periods gives zero — equal positive and negative area.

The second term vanishes, for every n and m, because sines and cosines are always orthogonal, by (6.7).

In the third term, only one piece survives. It is an infinite sum, and by (6.8) every term in it is zero except the single one where n equals m. That term contributes $b_m T/2$.

[1] The term was possibly coined by Richard Feynman, in his lectures published in 1963: R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Vol. I, Section 50-4. https://feynmanlectures.caltech.edu/I_50.html.

This is the critical step, and it is the one worth making sure you understand. An infinite sum has collapsed to a single term, and it did so because of orthogonality alone. The rest of the chapter applies this same step to other families of functions.

What remains is

$$\int_0^T f(t)\sin(m\omega_0 t)\,dt = b_m\frac{T}{2},$$

and renaming the dummy index m back to n gives (6.6).

Notice what the trick accomplished. You began with a function containing infinitely many unknown coefficients, and you extracted one of them exactly, by choosing what to multiply by. That is the purpose of an orthogonal family: multiplying by one member and integrating projects out that member's coefficient, and no other.

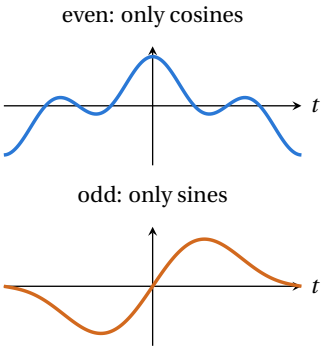


Figure 6.3 An even function, $f(-t) = f(t)$, and an odd one, $f(-t) = -f(t)$, over one period. The first contains only cosine terms, the second only sine terms.

6.1.4 Symmetry shortcuts

Before computing any integrals, look at the symmetry of your function (Figure 6.3). It will often halve the work:

- If $f(t)$ is **even**, only cosine terms appear. Every b_n is zero.
- If $f(t)$ is **odd**, only sine terms appear. Every a_n is zero, and so is a_0.

This is not a separate fact to memorize — it follows directly from the integrals. An odd function times a cosine is odd, and an odd function integrated over a symmetric interval gives zero.

6.1.5 Worked example: the square wave

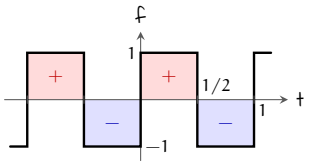
Example 6.1

Find the Fourier series of the square wave

$$f(t) = \begin{cases} -1, & -\frac{1}{2} < t < 0 \\ +1, & 0 < t < \frac{1}{2} \end{cases}$$

repeated with period $T = 1$.

Solution:



A graph of the function over two periods is the place to start. Since $T = 1$, the fundamental frequency is $\omega_0 = 2\pi$, and every term in the series has a frequency that is a multiple of 2π. Two observations save most of the labor. The average value of the function is zero, so $a_0 = 0$. And the function is odd, so every a_n vanishes and only sine terms survive.

That leaves only the b_n integral:

$$b_n = \frac{2}{T}\int_0^T f(t)\sin(n\omega_0 t)\,dt = 2\int_{-1/2}^{1/2} f(t)\sin(2\pi n t)\,dt$$
$$= 2\left(\int_{-1/2}^{0}(-1)\sin(2\pi n t)\,dt + \int_0^{1/2}(+1)\sin(2\pi n t)\,dt\right)$$
$$= \frac{2\,(1-\cos(\pi n))}{\pi n}.$$

(Why may the integral run from $-T/2$ to $T/2$ instead of from 0 to T? Answer in footnote.[2]) Now read off what that expression does. For even n, $\cos(\pi n) = 1$ and the coefficient is zero. For odd n, $\cos(\pi n) = -1$ and the coefficient is $4/\pi n$. So only the odd harmonics are present:

$$f(t) = \frac{4}{\pi}\sin(2\pi t) + \frac{4}{3\pi}\sin(6\pi t) + \frac{4}{5\pi}\sin(10\pi t) + \frac{4}{7\pi}\sin(14\pi t) + \cdots.$$

A square wave is built entirely from odd multiples of its fundamental, with amplitudes falling off as $1/n$.

Reality check: At $t = \frac{1}{4}$ every sine in the series is $+1$ or -1, so the series gives $(4/\pi)(1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots)$. The sum in parentheses is $\pi/4$ (it is the Taylor series of $\arctan x$ at $x = 1$), so the series gives 1, the value of the square wave there.

Figure 6.4 shows the series at work: the sum of the first few terms, and the coefficients themselves, which say at a glance which frequencies are present and how much of each.

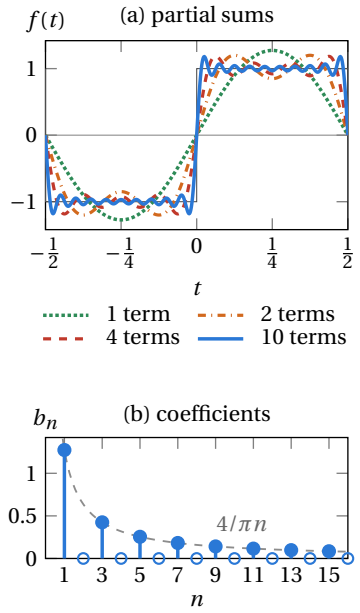


Figure 6.4 (a) The square wave of Example 6.1 (thin black) and the sums of the first 1, 2, 4 and 10 nonzero terms of its Fourier series. Each added term brings the sum closer to the square wave, except right next to the jumps: there the sum overshoots, and as more terms are added the overshoot gets narrower but not smaller, approaching about 9% of the jump. This is called the Gibbs phenomenon. (b) The coefficients b_n: zero for even n, and $4/\pi n$ for odd n. Animation

6.1.6 Other notations you will meet

Fourier series are written several different ways, and the differences are pure convention. Three you should recognize:

[2] The integrand is periodic, with period T, so its integral over any full period has the same value (Section 6.1.1). The square wave is given on $-\frac{1}{2} < t < \frac{1}{2}$, so that is the convenient period to use.

- $a_0/2$ **instead of** a_0**.** Many sources write the constant term as $a_0/2$, which makes their a_0 twice the size of the one defined here. The reason is tidiness: with that choice, the general a_n formula (6.5) also works for $n = 0$, so there is one formula instead of two.
- T **versus** ω_0**.** Writing $2\pi nt/T$ and writing $n\omega_0 t$ are the same thing. The second is preferred here because it makes the harmonic structure visible.
- **Space instead of time.** With x for t, a spatial period L for T, and k for ω, everything carries over unchanged.

None of these choices changes any physics as long as you are consistent within a convention, so just be sure to know what convention is being used by a given reference.



Adrien-Marie Legendre (1752–1833, French) Legendre introduced the polynomials that bear his name in 1782, as the coefficients in an expansion of the gravitational potential, in a study of the attraction of spheroids. He was the first to publish the method of least squares, though Gauss had found it earlier; he conjectured the law of quadratic reciprocity, later proved by Gauss, and, in 1798, the prime number theorem, proved in 1896; and he spent decades on the theory of elliptic integrals. His *Éléments de géométrie* (1794) was the leading elementary geometry text for about a century. The engraving shown here is from Georges Barral's *Le Panthéon scientifique de la tour Eiffel* (1892), long after Legendre's death; the only other known portrait, and the only authenticated one, is an 1820 watercolor caricature by Julien-Léopold Boilly. For two centuries, until 2005, books printed a profile of the politician Louis Legendre as his portrait. (Wikipedia)

6.2 Legendre polynomials

The second family arrives when Laplace's equation is separated in spherical coordinates (Section 7.4.3). For problems with no ϕ dependence, the polar part of the solution is a **Legendre polynomial**, $P_\ell(\cos\theta)$.

6.2.1 The polynomials

The Legendre polynomials $P_\ell(x)$ are a series of polynomials, one for each nonnegative integer ℓ, in which the highest power of x is x^ℓ. Like sines and cosines, they come up often in partial differential equations, they have useful properties, and they have been studied for centuries. Mathematica knows them as `LegendreP[`ℓ`,x]`, just as it knows `Sin[x]`. The first six are

$$\begin{aligned} P_0(x) &= 1, & P_3(x) &= \tfrac{1}{2}\left(5x^3 - 3x\right), \\ P_1(x) &= x, & P_4(x) &= \tfrac{1}{8}\left(35x^4 - 30x^2 + 3\right), \\ P_2(x) &= \tfrac{1}{2}\left(3x^2 - 1\right), & P_5(x) &= \tfrac{1}{8}\left(63x^5 - 70x^3 + 15x\right). \end{aligned} \tag{6.10}$$

In physics problems the argument is usually $x = \cos\theta$, so that, for example,

$$P_2(\cos\theta) = \tfrac{3}{2}\cos^2\theta - \tfrac{1}{2}.$$

Several properties can be read off (6.10) and Figure 6.5.

- P_ℓ contains only even powers of x when ℓ is even, and only odd powers when ℓ is odd, so the polynomials alternate between even and odd functions.
- At the ends of the interval every one of them equals ± 1:

$$P_\ell(1) = 1, \qquad P_\ell(-1) = (-1)^\ell.$$

- P_ℓ has ℓ zeros between -1 and 1.

All of the polynomials can be generated from **Rodrigues' formula**,

$$P_\ell(x) = \frac{1}{2^\ell\,\ell!}\left(\frac{d}{dx}\right)^\ell \left(x^2-1\right)^\ell. \tag{6.11}$$

For $\ell = 2$, for example,

$$\begin{aligned} P_2(x) &= \frac{1}{2^2\,2!}\frac{d^2}{dx^2}\left(x^4-2x^2+1\right) \\ &= \frac{1}{8}\left(12x^2-4\right) \\ &= \tfrac{1}{2}\left(3x^2-1\right). \end{aligned}$$

The Legendre polynomials are solutions of **Legendre's differential equation**,

$$\frac{d}{dx}\left[\left(1-x^2\right)\frac{df}{dx}\right] + \ell(\ell+1)\,f = 0, \tag{6.12}$$

which, written in terms of θ with $x = \cos\theta$, becomes

$$\frac{d}{d\theta}\left(\sin\theta\,\frac{df}{d\theta}\right) = -\ell(\ell+1)\sin\theta\,f. \tag{6.13}$$

This second form is the one that arises from Laplace's equation in spherical coordinates. Like any second-order equation, Legendre's equation has a second solution for each ℓ, written $Q_\ell(x)$ and called the **Legendre function of the second kind**. The Q_ℓ diverge at $x = \pm 1$ — that is, on the z axis, where $\theta = 0$ or π — so for any problem whose region includes the axis they are discarded.

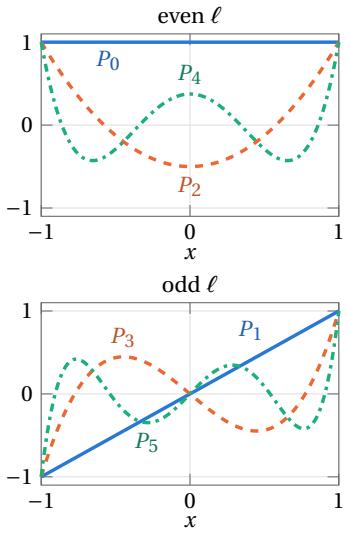


Figure 6.5 The Legendre polynomials P_0 through P_5 on the interval $-1 \le x \le 1$, with the even and odd orders plotted separately.

6.2.2 Orthogonality

Legendre polynomials are orthogonal too, so the same trick applies. They are orthogonal on the interval $-1 \le x \le 1$,

$$\int_{-1}^{1} P_\ell(x)P_m(x)\,dx = \begin{cases} 0 & \ell \neq m \\ \dfrac{2}{2\ell+1} & \ell = m \end{cases}, \tag{6.14}$$

and in a physics problem the substitution $x = \cos\theta$, with $dx = -\sin\theta\,d\theta$, turns that interval into the full range of polar angles. Using the minus sign to swap the limits of integration gives

$$\int_0^{\pi} P_\ell(\cos\theta)P_m(\cos\theta)\sin\theta\,d\theta = \begin{cases} 0 & \ell \neq m \\ \dfrac{2}{2\ell+1} & \ell = m \end{cases}. \tag{6.15}$$

Figure 6.6 shows two examples. The product P_2P_3 is the product of an even and an odd function, so it is odd, and its integral from -1 to 1 is clearly zero. The

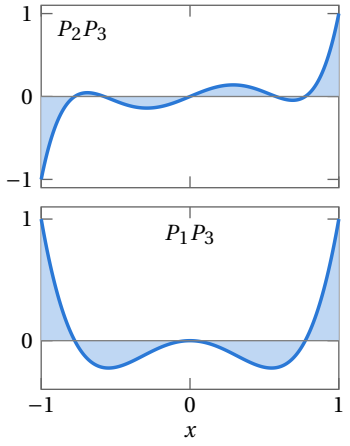


Figure 6.6 Two products of Legendre polynomials, with the area under each shaded. Top: P_2P_3 is odd, so it integrates to zero by symmetry. Bottom: P_1P_3 is even, but its positive and negative areas still cancel.

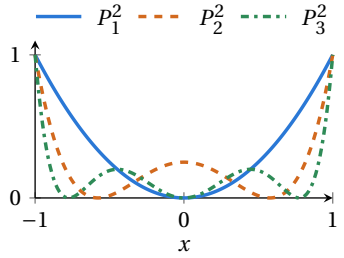


Figure 6.7 P_1^2, P_2^2, and P_3^2. The areas under them are 2/3, 2/5, and 2/7: the normalization $2/(2\ell+1)$ shrinks as ℓ grows.

product P_1P_3 is even, so it is not so obvious that it integrates to zero; but the positive and negative areas still cancel. The integral of $P_\ell P_m$ from -1 to 1 is zero whenever $\ell \neq m$.

Compare (6.14) with (6.8): the structure is identical, and only the normalization differs. Where the sine integral gave $T/2$ regardless of n, the Legendre integral gives $2/(2\ell+1)$, which depends on ℓ (Figure 6.7) — a small bookkeeping difference that is easy to forget and produces answers wrong by a factor.

So if a potential on a spherical boundary is given as some function of θ, you expand it as $\sum_\ell c_\ell P_\ell(\cos\theta)$ and extract the coefficients by Fourier's trick, exactly as before,

$$c_\ell = \frac{2\ell+1}{2}\int_{-1}^{1} f(x)P_\ell(x)\,dx. \tag{6.16}$$

6.2.3 Comparison with sines and cosines

The parallels between the two families are close enough to be worth setting out point by point, as in Table 6.1.

	SINES AND COSINES	LEGENDRE POLYNOMIALS
1.	Two oscillatory functions, $\sin x$ and $\cos x$. Sometimes one of them is not used, because of the symmetry of the problem.	Two functions for each ℓ, $P_\ell(x)$ and $Q_\ell(x)$. Typically Q_ℓ is not used, because it is infinite at $x=\pm1$.
2.	The value of $\sin x$ or $\cos x$ for arbitrary x comes from a calculator or a computer program.	The value of $P_\ell(x)$ for arbitrary x comes from a calculator or a computer program.
3.	$\sin x = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!}$, $\cos x = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{(2k)!}$.	$P_\ell(x) = \frac{1}{2^\ell \ell!}\left(\frac{d}{dx}\right)^\ell \left(x^2-1\right)^\ell$, or $P_\ell(x) = 2^\ell \sum_{k=0}^{\ell} \binom{\ell}{k}\binom{(\ell+k-1)/2}{\ell} x^k$.[3]
4.	$\sin(m\pi x)$ is odd for all m; $\cos(m\pi x)$ is even for all m.	$P_\ell(x)$ is odd for odd ℓ and even for even ℓ.
5.	At $x=1$, $\sin(m\pi x)=0$ and $\cos(m\pi x)=\pm1$ for all m.	At $x=1$, $P_\ell(x)=1$ for all ℓ; at $x=-1$, $P_\ell(x)=\pm1$.
6.	$f=\sin(m\pi x)$ satisfies $f''+(m\pi)^2 f=0$.	$f=P_\ell(x)$ satisfies $\frac{d}{dx}\left[\left(1-x^2\right)\frac{df}{dx}\right]+\ell(\ell+1)f=0$.
7.	$\sin(n\pi x)$ is orthogonal to $\sin(m\pi x)$ on $(0,1)$: $\int_0^1 \sin(n\pi x)\sin(m\pi x)\,dx = \begin{cases} 0 & n\neq m \\ \frac{1}{2} & n=m \end{cases}$.	$P_\ell(x)$ is orthogonal to $P_m(x)$ on $(-1,1)$, as in (6.14); $P_\ell(\cos\theta)$ is orthogonal to $P_m(\cos\theta)$ on $(0,\pi)$, with a weighting of $\sin\theta$, as in (6.15).

Table 6.1 Sines and cosines compared with Legendre polynomials, point by point.

[3] The upper entry of the second binomial coefficient need not be an integer. For any number a, $\binom{a}{\ell} = a(a-1)\cdots(a-\ell+1)/\ell!$.

6.2.4 Colton's trick

There is a useful shortcut. If the boundary function is *already* a polynomial in $\cos\theta$ — and in textbook problems it very often is — then you do not need to evaluate (6.16) at all. You can write the function as a combination of Legendre polynomials by inspection, matching powers of $\cos\theta$ until both sides agree.

For instance, a boundary potential proportional to $\cos^2\theta$ can be matched against $P_2(\cos\theta) = \frac{3}{2}\cos^2\theta - \frac{1}{2}$ and $P_0(\cos\theta) = 1$. Matching the $\cos^2\theta$ term takes $\frac{2}{3}P_2$, which leaves behind a constant of $-\frac{1}{3}$, and adding $\frac{1}{3}P_0$ cancels it:

$$\cos^2\theta = \tfrac{2}{3}P_2(\cos\theta) + \tfrac{1}{3}P_0(\cos\theta). \qquad \text{(found by inspection)} \quad (6.17)$$

Figure 6.8 shows the two terms adding up. Because the expansion is unique, a combination found by inspection *is* the answer, and no integrals need to be evaluated. Chapter 7 uses this repeatedly: once both sides of an equation are written as sums of the same orthogonal functions (here, Legendre polynomials), the coefficients of each function must agree, a rule which I will shamelessly call **Colton's trick**, by analogy to Fourier's trick, even though this idea is centuries old.

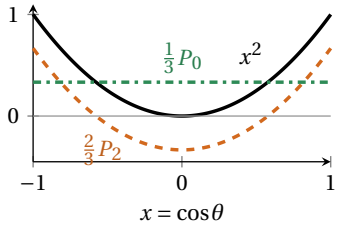


Figure 6.8 The expansion (6.17), plotted against $x = \cos\theta$: the two terms (dashed and dash-dotted) add up to $x^2 = \cos^2\theta$ (solid).

Colton's trick works for sines and cosines too, with trigonometric identities doing the job that matching powers of $\cos\theta$ did above. Take the function $f(t) = \sin^3(\omega_0 t)$, and write $u = \omega_0 t$ for short. The identity $\sin^2 u = \frac{1}{2}(1 - \cos 2u)$ gives $\sin^3 u = \frac{1}{2}\sin u - \frac{1}{2}\sin u\cos 2u$, and the identity $\sin u\cos 2u = \frac{1}{2}(\sin 3u - \sin u)$ turns the last term into sines of multiples of u:

$$\sin^3(\omega_0 t) = \tfrac{3}{4}\sin(\omega_0 t) - \tfrac{1}{4}\sin(3\omega_0 t). \qquad \text{(from two identities)} \quad (6.18)$$

The right side is already a Fourier series, so Colton's trick reads off every coefficient at once: $b_1 = \frac{3}{4}$, $b_3 = -\frac{1}{4}$, and all the others are zero.

Fourier's trick gives the same coefficients, but with much more work. The a_n all vanish because $\sin^3$ is odd (Section 6.1.4), but the b_n must be found one integral at a time. By (6.6),

$$\begin{aligned} b_1 &= \frac{2}{T}\int_0^T \sin^4(\omega_0 t)\,dt \\ &= \frac{2}{T}\cdot\frac{3T}{8} \\ &= \frac{3}{4}, \end{aligned}$$

and that integral by itself takes two more identities, or a table of integrals. Then b_3 needs $\int_0^T \sin^3(\omega_0 t)\sin(3\omega_0 t)\,dt$, which is harder still, and every other b_n needs an integral of its own just to show that it is zero.

6.2.5 A connection to come

Legendre polynomials also appear somewhere that seems to have nothing to do with Laplace's equation: in the expansion of $1/\mathscr{r}$ for a distant field point. That

expansion — the generating function — is derived in Section 8.3.2, where it is needed, rather than here, because it makes more sense alongside the physics it produces. It is worth knowing in advance that the same functions arrive by two quite different routes.

6.3 Bessel functions

The third family comes from separating the wave equation in *cylindrical* coordinates, which is where it arises in this book: in the study of waveguides (Section 23.5). The radial part of the solution satisfies Bessel's equation,

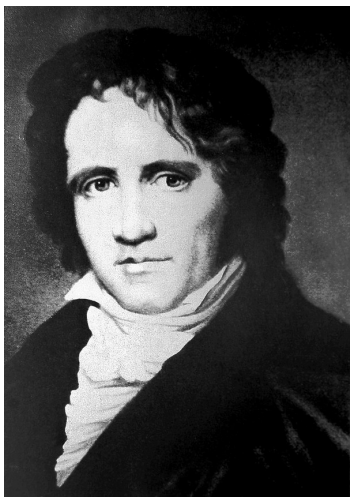
$$x^2\frac{d^2f}{dx^2}+x\frac{df}{dx}+\left(x^2-\alpha^2\right)f=0, \tag{6.19}$$

and the solution that stays finite at $x=0$ is the **Bessel function** $J_\alpha(x)$. For our purposes α is a non-negative integer. (Laplace's equation in cylindrical coordinates, for the problems with no z dependence that are solved in Section 7.4.4, does not give Bessel functions. It gives powers of s, and $\ln s$, together with sines and cosines of ϕ.)

Bessel functions are not elementary — they cannot be written in closed form with sines, exponentials and powers — but they are well understood, they have been tabulated since the nineteenth century, and Mathematica knows them as `BesselJ[α,x]` just as it knows `Sin[x]`. Not being elementary is a statement about notation, not about difficulty.

Qualitatively, $J_\alpha(x)$ oscillates and decays, rather like a sine wave whose amplitude slowly dies away. Unlike a sine wave, its zeros are not evenly spaced. The zeros matter enough to have their own notation: $u_{\alpha m}$ is the mth zero of J_α. The first few are

$$\begin{aligned} J_0:&\quad u_{01}=2.405,\quad u_{02}=5.520,\quad u_{03}=8.654,\\ J_1:&\quad u_{11}=3.832,\quad u_{12}=7.016,\quad u_{13}=10.173. \end{aligned}$$



Friedrich Wilhelm Bessel (1784–1846, German) Bessel left school at 14 to become an apprentice in a trading firm in Bremen, where an interest in finding a ship's position at sea led him to astronomy. In 1804 his calculation of the orbit of Halley's comet impressed the astronomer Heinrich Olbers; two years later Bessel became an assistant at the observatory in Lilienthal, and in 1810, without a university degree, he was made director of the new observatory at Königsberg, which he ran until his death. In 1838 he published the parallax of the star 61 Cygni, 0.314 arcsecond, one of the first reliable measurements of the distance to a star. The functions named for him came from his work on planetary orbits; he treated them systematically in 1824, though Daniel Bernoulli and Euler had met special cases decades earlier. (Wikipedia)

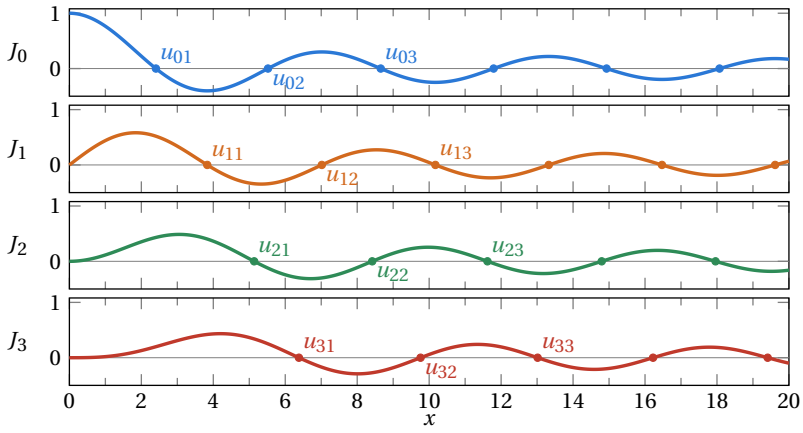


Figure 6.9 The Bessel functions J_0 through J_3, with their zeros marked by dots and the first three of each labeled $u_{\alpha m}$. Only J_0 is nonzero at $x = 0$; the others start from zero, more slowly the higher the order, and their first zeros move outward as the order increases. The zeros are not evenly spaced, although far from the origin the spacing approaches π.



Olinde Rodrigues (1795–1851, French) Benjamin Olinde Rodrigues was born in Bordeaux and studied in Paris. His thesis of 1816, on the attraction of spheroids, contains the formula now named for him, (6.11), which James Ivory and Carl Jacobi later found independently. He made his career as a banker and financier and was one of the leaders of the Saint-Simonians, the followers of the social reformer Henri de Saint-Simon. In a paper of 1840 on rotations he gave the rule for combining two rotations, which amounts to the multiplication of quaternions, three years before William Rowan Hamilton discovered them. (Wikipedia)

Two further facts are worth recording now. Bessel's equation is second order, so it has a second family of solutions, written $Y_\alpha(x)$ (or sometimes N_α) and called Bessel functions of the second kind. They diverge at $x = 0$ (Figure 6.10), so for any problem whose region includes the axis they are discarded on physical grounds — the same reasoning that discards the $1/r$ solution inside a sphere. And the derivative of J_0 is $-J_1$, which means the maxima and minima of J_0 sit at the zeros of J_1.

6.3.1 Orthogonality, with a weight

Bessel functions are orthogonal too, and this is why they belong in this chapter — but with one complication the other two families do not have. In the physical setup the substitution $x = u_{\alpha m} s$ is made, where s is the cylindrical radial coordinate scaled so that the boundary sits at $s = 1$. Each choice of m then gives a different function, all built from the same J_α and distinguished only by how many of its oscillations fit into the interval.

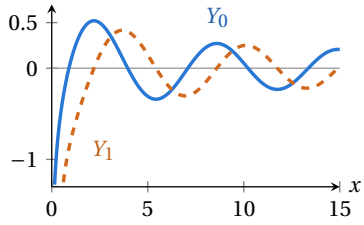


Figure 6.10 The Bessel functions of the second kind Y_0 (solid) and Y_1 (dashed). Both go to $-\infty$ as $x \to 0$.

Those functions are orthogonal over $0 \le s \le 1$ — provided the integral carries a factor of s:

$$\int_0^1 J_\alpha(u_{\alpha m} s)\, J_\alpha(u_{\alpha m'} s)\, s\, ds = \begin{cases} 0 & m \neq m' \\ \frac{1}{2}\left[J_{\alpha+1}(u_{\alpha m})\right]^2 & m = m' \end{cases}. \tag{6.20}$$

That extra s is called a **weight function**, and it is not arbitrary: it is the s that appears in the cylindrical volume element $d\tau = s\, ds\, d\phi\, dz$ (Figure 6.11). Orthogonality relations come with whatever weight the geometry supplies. Sines have one too, equal to 1, and so do Legendre polynomials: theirs is 1 in the variable x and $\sin\theta$ in the variable θ, as in (6.15).

6.4 Summary of similarities and differences

The three families are three versions of the same idea, and it helps to set them side by side, as in Table 6.2.

	SINES, COSINES	LEGENDRE	BESSEL
Arises from	Laplace's equation, Cartesian coordinates	Laplace's equation, spherical coordinates	the wave equation, cylindrical coordinates
Written	$\sin(n\omega_0 t)$, $\cos(n\omega_0 t)$	$P_\ell(\cos\theta)$, $Q_\ell(\cos\theta)$	$J_\alpha(u_{\alpha m}s)$, $Y_\alpha(u_{\alpha m}s)$
Typically used	both	P_ℓ	J_α
Labeled by	one integer n	one integer ℓ	two integers, α and m
Interval	one period, 0 to T	$-1 \le x \le 1$, i.e. all θ	$0 \le s \le 1$
Weight function	1	1 in x; $\sin\theta$ in θ	s
Orthogonality relation	$\int_0^T \sin(m\omega_0 t)\sin(n\omega_0 t)\,dt$ $= 0$ for $m \neq n$, $= T/2$ for $m = n$; the same for two cosines ($m, n \ge 1$); 0 for a sine and a cosine	$\int_{-1}^1 P_\ell(x)\,P_m(x)\,dx$ $= 0$ for $\ell \neq m$, $= 2/(2\ell+1)$ for $\ell = m$	$\int_0^1 J_\alpha(u_{\alpha m}s)\,J_\alpha(u_{\alpha m'}s)\,s\,ds$ $= 0$ for $m \neq m'$, $= \frac{1}{2}[J_{\alpha+1}(u_{\alpha m})]^2$ for $m = m'$
Zeros	evenly spaced	ℓ of them in the interval	not evenly spaced; looked up as $u_{\alpha m}$

Table 6.2 The three families of orthogonal functions side by side.

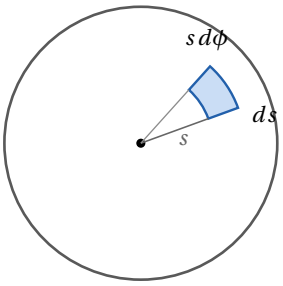


Figure 6.11 In a disk, the area element at radius s is $s\,ds\,d\phi$ (shaded). The extra factor s, which comes from the geometry, is the weight function of the Bessel orthogonality relation (6.20).

What is the same, in all three cases:

- an infinite family of functions, labeled by an integer;
- each member orthogonal to every other over the relevant interval;
- any reasonable function on that interval expandable as a sum over the family;
- and the coefficients of that sum extracted by Fourier's trick — multiply by one member, integrate, and watch every term but one vanish.

What differs is bookkeeping: which interval, what weight, what the normalization integral comes to, and how many integers it takes to label a member. If you understand Section 6.1.3 thoroughly, you understand all three families, and the rest is a matter of looking up constants.

6.5 Summary

Fourier series. A periodic function with period T contains only multiples of the fundamental frequency $\omega_0 = 2\pi/T$:

$$f(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos(n\omega_0 t) + \sum_{n=1}^{\infty} b_n \sin(n\omega_0 t)$$

$a_0 = \dfrac{1}{T}\displaystyle\int_0^T f\,dt$ the average value

$a_n = \dfrac{2}{T}\displaystyle\int_0^T f\cos(n\omega_0 t)\,dt$ the cosine coefficients

$b_n = \dfrac{2}{T}\displaystyle\int_0^T f\sin(n\omega_0 t)\,dt$ the sine coefficients

Orthogonality is what makes those formulas work. Sines and cosines are always orthogonal to each other; two sines or two cosines are orthogonal unless the integers match, in which case the integral gives $T/2$.

Fourier's trick. Multiply the series by one member of the family, integrate over the interval, and every term vanishes except the one that matches. This one technique works for all three families.

Colton's trick. When a function is already a finite sum of members of the family, or an identity makes it one, the coefficients can be read off directly, since the expansion is unique.

Symmetry first. An even function has only cosine terms; an odd function has only sine terms. Check before integrating.

The other two families work identically, differing only in interval, weight function and normalization — see the table in Section 6.4.

Exercises

Exercises for 6.1 Sines and cosines: Fourier series

Warm-up questions Read Section 6.1, then answer before class

W6.1 A function $f(t)$ has period T, so its Fourier series contains only whole-number multiples of $\omega_0 = 2\pi/T$. Add to it $\sin(3\omega_0 t/2)$, whose frequency lies halfway between two of them. Is the sum periodic? If so, what are its period and its fundamental frequency, which frequencies does its Fourier series contain, and is there a term at the fundamental frequency itself?

W6.2 Try Fourier's trick with the powers $1, x, x^2, x^3, \ldots$, which are not orthogonal on $-1 \le x \le 1$. Write $f(x) = c_0 + c_1 x + c_2 x^2 + \cdots$, multiply both sides by x^2, and integrate from -1 to 1. Which of the coefficients appear in the result? Can you solve it for c_2? How does this differ from the critical step of the derivation in Section 6.1.3?

W6.3 On $0 < t < \frac{1}{2}$ the square wave of Example 6.1 equals 1, so there $1 = (4/\pi)\left[\sin(2\pi t) + \frac{1}{3}\sin(6\pi t) + \frac{1}{5}\sin(10\pi t) + \cdots\right]$. (a) At $t = 0$, where the square wave jumps from -1 to $+1$, what value does the series give, and why is that a reasonable value? (b) Why can no finite number of the terms add up to 1 everywhere on $0 < t < \frac{1}{2}$?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P6.1 Verify the orthogonality relations by direct integration, taking $T = 1$ and $\omega_0 = 2\pi$ throughout.

(a) $\int_0^1 \sin(4\cdot 2\pi t)\cos(4\cdot 2\pi t)\,dt$

(b) $\int_0^1 \sin(3\cdot 2\pi t)\sin(8\cdot 2\pi t)\,dt$

(c) $\int_0^1 \cos(1\cdot 2\pi t)\cos(2\cdot 2\pi t)\,dt$

(d) $\int_0^1 \sin(3\cdot 2\pi t)\sin(3\cdot 2\pi t)\,dt$ — and explain in one sentence why this one is different from the others.

P6.2 Carry out Fourier's trick for the a_n coefficients, following the derivation in Section 6.1.3 but multiplying by $\cos(m\omega_0 t)$ instead. State clearly at each step which orthogonality relation you are using.

P6.3 Find the Fourier series of the triangle wave of period $T = 1$ that rises linearly from -1 at $t = -\frac{1}{2}$ to $+1$ at $t = 0$ and falls linearly back to -1 at $t = \frac{1}{2}$ (Figure 6.12).

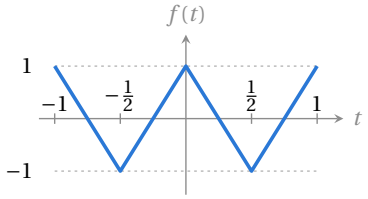


Figure 6.12 Two periods of the triangle wave of Exercise P6.3.

(a) Before integrating, use symmetry to say which coefficients must vanish.

(b) Find the surviving coefficients.

(c) Compare the rate at which your coefficients fall off with n against the $1/n$ found for the square wave in Example 6.1. Which function is "smoother", and how is that visible in the coefficients?

P6.4 Rewrite (6.3) and the coefficient formulas (6.5)–(6.6) for a function of position x with spatial period L. This is the form you will use in boundary-value problems, so it is worth writing out once by hand.

Exercises for 6.2 Legendre polynomials

Warm-up questions Read Section 6.2, then answer before class

W6.4 A classmate uses Rodrigues' formula and finds $P_6(x) = (231x^6 - 315x^4 + 105x^2 + 5)/16$. Without redoing the six derivatives, give two independent reasons why this cannot be right, using properties from this section. Which of the properties does the classmate's polynomial satisfy, so that checking them would not have caught the error?

W6.5 Figure 6.6 shows that P_2P_3 integrates to zero over $-1 \le x \le 1$ by symmetry alone, while for P_1P_3, which is even, the positive and negative areas cancel less obviously. (a) For which pairs $\ell \neq m$ does symmetry alone guarantee that $\int_{-1}^{1} P_\ell P_m \, dx = 0$? (b) Since $P_0 = 1$, what does (6.14) tell you about $\int_{-1}^{1} P_\ell \, dx$ for $\ell \ge 1$? So which coefficient of a sum $\sum_\ell c_\ell P_\ell(x)$ survives when the sum is integrated over $-1 \le x \le 1$?

W6.6 Every Legendre polynomial equals 1 at $x = 1$, yet the average value of P_ℓ^2 over $-1 \le x \le 1$ shrinks as ℓ grows (Figure 6.7: the areas are $2/3$, $2/5$, and $2/7$), while the average of $\sin^2(n\omega_0 t)$ over a period is $\frac{1}{2}$ for every n. What do the graphs of Figure 6.5 show that explains the difference?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P6.5 Each of the following boundary functions can be written as a finite combination of Legendre polynomials by inspection, without evaluating a single integral. Do so. You will need $P_0 = 1$, $P_1 = \cos\theta$, $P_2 = \frac{3}{2}\cos^2\theta - \frac{1}{2}$, and $P_3 = \frac{5}{2}\cos^3\theta - \frac{3}{2}\cos\theta$.

(a) $f(\theta) = \cos\theta \sin^2\theta$

(b) $f(\theta) = \sin^2\theta$

(c) $f(\theta) = 1 + \cos\theta + \cos^3\theta$

P6.6 A function on $-1 \le x \le 1$ is expanded as $f(x) = \sum_\ell c_\ell P_\ell(x)$. Derive (6.16) for the coefficients, using (6.14) and the same argument as in Section 6.1.3. Where exactly does the factor $(2\ell+1)/2$ come from, and what would go wrong if you assumed the normalization were the same for every ℓ, as it is for sines?

Exercises for 6.3 Bessel functions

Warm-up questions Read Section 6.3, then answer before class

W6.7 This section discards Y_α in any problem whose region includes the axis, by the same reasoning that discards a $1/r$ term inside a sphere, and the Legendre functions of the second kind, Q_ℓ, are discarded in the same way. For each of the three, describe a region in which it could *not* be discarded on these grounds, and say why.

W6.8 Using Figure 6.9 and the zeros listed in this section, sketch $J_0(u_{01}s)$, $J_0(u_{02}s)$, and $J_0(u_{03}s)$ for $0 \le s \le 1$. (a) Where does each one cross zero for $0 < s < 1$? (b) All three are zero at $s = 1$. What kind of boundary condition does that build in? Give a physical example.

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P6.7 The orthogonality relation (6.20) carries a weight factor of s that the sine and Legendre relations do not.

(a) Where does that factor come from?

(b) Sines and Legendre polynomials also have a weight function. What is it, and why is it easy not to notice?

P6.8 Using a table or Mathematica's `BesselJZero`:

(a) List the first four zeros of J_0 and the first four of J_1.

(b) Compute the spacing between consecutive zeros of J_0 and show that it approaches a constant for large x. What constant, and why is that value not a surprise given the shape of the function?

Exercises for 6.5 Summary

P6.9 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as the overshoot of a Fourier series near a jump in the function, the Fourier coefficients of a recorded musical note, how quickly the Fourier coefficients of smooth and of rough functions fall off, the expansion of a function of θ in Legendre polynomials, the vibration patterns of a circular drumhead, or the orthogonality of Bessel functions with and without the weight s.

Chapter 7

Boundary Conditions and Boundary Value Problems

Most of the problems in the preceding chapters started from a known charge distribution and asked for the field or the potential it produces. This chapter turns the question around. Often what we know is not where the charges are, but what the potential is on the edges of a region: a conductor held at a fixed voltage, a grounded plate, a surface on which the potential has been measured. The charges sit on those boundaries, wherever they need to be, and inside the region there is no charge at all. Finding the potential in such a region is called a **boundary value problem**.

The governing equation is Laplace's equation, and the first section shows why a boundary value problem has only one answer. That fact is what makes the rest of the chapter work. It lets us use three quite different methods — the method of images, the numerical relaxation method, and separation of variables — and be sure that whatever answer a method produces is *the* answer.

7.1 Boundary conditions and uniqueness theorems

Poisson's equation (Section 5.10) relates the potential to the charge density,

$$\nabla^2 V = -\frac{\rho}{\varepsilon_0},$$

and in a region where there is no charge it becomes Laplace's equation,

$$\nabla^2 V = 0.$$

At first sight Laplace's equation seems very limiting: if there is no charge, how can anything interesting happen? The point is that the charges can be nearby, just not inside the region we care about. Consider a capacitor made of two plates held at +100 V and −100 V. There is charge on the plates, but nowhere else. If we define our region of interest as all of space except the plates themselves, then Laplace's equation holds everywhere in that region, and the plates become part of its boundary (Figure 7.1).

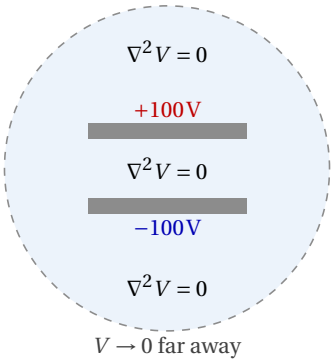


Figure 7.1 A capacitor with plates held at +100 V and −100 V. Everywhere except on the plates there is no charge, so Laplace's equation holds throughout the shaded region. The region's boundary is the surfaces of the plates, together with infinity, where $V \to 0$.

7.1.1 Boundary conditions

A solution of Laplace's equation is not determined by the equation alone. The functions $V = 0$, $V = x$, and $V = x^2 - y^2$ all satisfy (5.27), and so do infinitely many others. What picks out one of them is the set of **boundary conditions**: information about V on the surfaces that enclose the region.

The most common boundary condition is simply the value of V on each boundary surface. A conductor in equilibrium is an equipotential, so a conducting boundary supplies a condition of the form $V =$ constant over its whole surface. A *grounded* conductor has $V = 0$. Grounding does not mean that the conductor is uncharged: charge can flow between the conductor and the ground in whatever amount is needed to hold its potential at zero, and in Section 7.2.2 we will see a grounded plane that carries a substantial induced charge. For a region that extends to infinity, the condition $V \to 0$ far away also counts as a boundary condition.

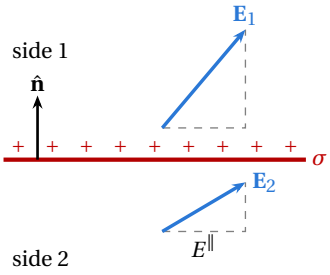


Figure 7.2 The fields just above ($\mathbf{E}_1$, on side 1, where $\hat{\mathbf{n}}$ points) and just below ($\mathbf{E}_2$) a surface carrying charge density $\sigma > 0$. Their parallel components are equal, and their perpendicular components differ by σ/ε_0, (7.1).

The other kind of boundary is a surface carrying a known surface charge density σ. Across such a surface the potential is continuous, the component of $\mathbf{E}$ parallel to the surface is continuous, and the perpendicular component jumps by σ/ε_0,

$$V_1 = V_2, \qquad \mathbf{E}_1^{\parallel} = \mathbf{E}_2^{\parallel}, \qquad E_1^{\perp} - E_2^{\perp} = \frac{\sigma}{\varepsilon_0}, \tag{7.1}$$

where side 1 is the side the normal points toward (Figure 7.2). The jump in $E^{\perp}$ follows from Gauss's law (Section 4.4), the continuity of $\mathbf{E}^{\parallel}$ from the curl of $\mathbf{E}$ (Section 5.2), and the continuity of V from the line integral of $\mathbf{E}$ across the surface (Section 5.3.4). Because $E^{\perp} = -\partial V/\partial n$, with n the coordinate along the normal, the last of these can be written in terms of the potential,

$$\frac{\partial V_1}{\partial n} - \frac{\partial V_2}{\partial n} = -\frac{\sigma}{\varepsilon_0}. \tag{7.2}$$

With words, it says that the slope of the potential changes abruptly at a surface charge, by an amount proportional to the charge density. At the surface of a conductor the field inside is zero, and (7.2) reduces to

$$\sigma = -\varepsilon_0 \frac{\partial V}{\partial n}, \tag{7.3}$$

with the derivative taken just outside, along the outward normal. Once a boundary value problem has been solved, (7.3) is how you find the charge that was induced on the conductors.

7.1.2 Laplace's equation and averages

Laplace's equation has a property that is easiest to see in one dimension. There (5.27) is just $d^2V/dx^2 = 0$, whose solutions are straight lines, $V = mx + b$. A straight line has no bumps or dips, and its value at any point is exactly the average of its values a distance h to either side,

$$V(x) = \tfrac{1}{2}\left[V(x-h) + V(x+h)\right].$$

The same is true in two and three dimensions, with a circle or a sphere in place of the two neighboring points: the value of V at any point equals the average of V over any sphere centered on that point, provided the sphere lies entirely within the charge-free region. Solutions of Laplace's equation are in this sense the smoothest possible functions that match the boundaries. A two-dimensional solution looks like a rubber sheet stretched over a frame whose height varies around its edge: the sheet meets the frame everywhere, and in between it varies smoothly, with no bumps (Figure 7.3).

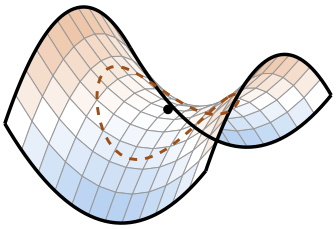


Figure 7.3 The solution $V = x^2 - y^2$ of Laplace's equation in two dimensions, drawn as a height above the square $|x|, |y| \le 1$. Like a rubber sheet stretched over a frame (black), it has no bumps or dips, and its largest and smallest values lie on the frame. Its heights around the dashed circle average to its height at the center of the circle, zero.

The average over a sphere

Let P be a point in a charge-free region, and let $\overline{V}(R)$ be the average value of V over a sphere of radius R centered on P. Using spherical coordinates centered on P, with $d\Omega = \sin\theta\, d\theta\, d\phi$ the element of solid angle,

$$\overline{V}(R) = \frac{1}{4\pi R^2} \oint V\, da = \frac{1}{4\pi} \oint V(R,\theta,\phi)\, d\Omega.$$

The radius appears only inside V, so the derivative with respect to R can be brought inside the integral,

$$\frac{d\overline{V}}{dR} = \frac{1}{4\pi} \oint \frac{\partial V}{\partial r}\, d\Omega.$$

On the sphere, $(\partial V/\partial r)\, R^2\, d\Omega$ is $\nabla V \cdot d\mathbf{a}$, so multiplying and dividing by R^2 and then applying the divergence theorem gives

$$\frac{d\overline{V}}{dR} = \frac{1}{4\pi R^2} \oint \nabla V \cdot d\mathbf{a} = \frac{1}{4\pi R^2} \int \nabla^2 V\, d\tau.$$

The volume integral is over the inside of the sphere, where $\nabla^2 V = 0$, so $d\overline{V}/dR = 0$. The average therefore does not depend on the radius. As R shrinks toward zero the sphere closes in on P and the average becomes $V(P)$, so

$$\overline{V}(R) = V(P) \tag{7.4}$$

for every sphere around P that fits inside the region. In two dimensions the same argument, with a circle in place of the sphere, shows that $V(P)$ is the average of V around any circle centered on P.

An important consequence of (7.4) is that *a solution of Laplace's equation has no local maxima or minima inside the region.* At a local maximum, V would be larger than at all the points around it, and so larger than their average, which (7.4) forbids. The largest and smallest values of V are always found on the boundaries.

To make that precise, suppose the largest value of V in the region, M, occurred at an interior point P. Every sphere around P that fits in the region has values of V that are all less than or equal to M, and their average is $V(P) = M$. The only way an average of numbers none of which exceeds M can equal M is for all of them to equal M. So $V = M$ on every such sphere, which means $V = M$ throughout a ball around P. Repeating the argument from points in that ball, the region where

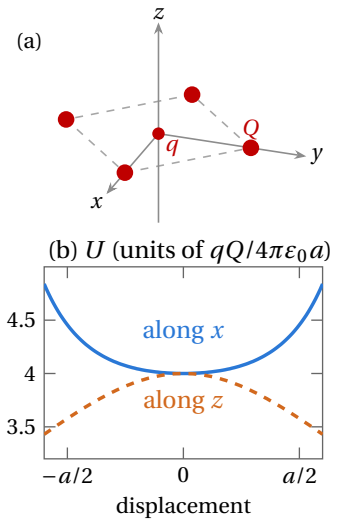


Figure 7.4 Earnshaw's theorem for a charge q at the center of four equal charges Q of the same sign, at the corners of a square (a). The potential energy of q (b) has a minimum for a displacement along the x axis (or the y axis) but a maximum along the z axis, so pushed along z, the charge does not return.

$V = M$ spreads outward until it reaches the boundary. Either way, the largest value of V is found on the boundary — and in the second case V was a constant all along. The same argument applies to the smallest value.

This also proves the result that Section 4.5.1 used for the Faraday cage. Suppose a charge-free region is enclosed by a single equipotential surface, as the empty cavity in a conductor is, and that the surface is at the potential V_0. The largest and the smallest values of V in the region are both found on the boundary, so both are V_0, and $V = V_0$ everywhere inside. With V constant, $\mathbf{E} = -\nabla V = 0$: there is no field anywhere in the region.

This has a physical consequence known as **Earnshaw's theorem**: a charged particle cannot be held in stable equilibrium by electrostatic forces alone. For a positive charge, a stable equilibrium point would have to be a local minimum of the potential, and in the charge-free space around the point there are none. (The particle's own field does not count, since it exerts no net force on the particle.) Push the particle slightly in some direction and, in at least one direction, it will move away rather than return (Figure 7.4).

7.1.3 The first uniqueness theorem

We can now prove the result that the rest of this chapter depends on.

First uniqueness theorem. *The solution of Laplace's equation in a region is uniquely determined if the value of V is specified on every part of the boundary of that region.*

Proof of the first uniqueness theorem

Suppose there were two different solutions, V_1 and V_2. Both satisfy Laplace's equation in the region,

$$\nabla^2 V_1 = 0 \qquad \text{and} \qquad \nabla^2 V_2 = 0,$$

and both take the specified values on the boundary. Consider their difference,

$$V_3 \equiv V_1 - V_2.$$

Because the Laplacian is a linear operator,

$$\nabla^2 V_3 = \nabla^2 V_1 - \nabla^2 V_2 = 0,$$

so V_3 also satisfies Laplace's equation in the region. That alone does not make V_3 zero; plenty of nonzero functions satisfy Laplace's equation. The second ingredient is the boundary. On every part of the boundary, V_1 and V_2 take the same specified values, so there

$$V_3 = V_1 - V_2 = 0.$$

Now use the result of Section 7.1.2: V_3 has no local maxima or minima inside the region, so its largest and smallest values are both found on the boundary. But V_3 is

zero everywhere on the boundary, so its largest value is 0 and its smallest value is 0. A function whose largest and smallest values are both zero is zero everywhere,

$$V_3 = 0 \qquad \text{throughout the region,}$$

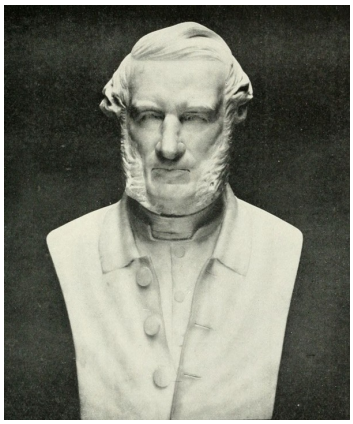
and so $V_1 = V_2$. There was only one solution after all.

Two remarks about the proof. First, it needs V to be specified on *every* part of the boundary; if a region extends to infinity, then "every part" includes infinity, and the usual condition is $V \to 0$ there. Second, the same proof works for Poisson's equation. If V_1 and V_2 both satisfy (5.26) with the same ρ, then

$$\nabla^2 V_3 = -\frac{\rho}{\varepsilon_0} - \left(-\frac{\rho}{\varepsilon_0}\right) = 0,$$

and the rest of the argument goes through unchanged. So the potential in a region is unique if the charge density inside the region and the potential on its boundary are both specified.

The practical importance of the theorem is this: *if you find a function that satisfies Laplace's equation and matches all the boundary conditions, then it is the solution*, however you found it. You can guess it, borrow it from a different problem that happens to have the same boundary conditions, compute it numerically, or build it up out of a series. Each of the methods in the rest of this chapter does one of those things, and the uniqueness theorem is what guarantees that the answer is right.



Samuel Earnshaw (1805–1888, English) Earnshaw, the son of a Sheffield file cutter, entered St John's College, Cambridge, in 1827 and in 1831 was Senior Wrangler and first Smith's Prizeman. His theorem comes from a paper read in 1839 on the luminiferous ether: he showed that the forces holding its particles in place could not vary as the inverse square of the distance, because a particle acted on only by such forces has no position of stable equilibrium. From 1847 he was assistant minister of the parish church in Sheffield, where he helped to found Firth College, a forerunner of the University of Sheffield. The picture is a bust of him by William Ellis. (Wikipedia)

7.1.4 The second uniqueness theorem

There is a second uniqueness theorem, which applies when what is known about a set of conductors is not their potentials but the total charge on each one.

Second uniqueness theorem. *In a region surrounded by conductors and containing a specified charge density ρ, the electric field is uniquely determined if the total charge on each conductor is given.*

In other words, if you put specified amounts of charge on a set of conductors (Figure 7.5), the charge distributes itself over their surfaces in exactly one way. Move one conductor away and then return it to the same place, and every charge returns to exactly where it was before. The proof is different from the one above: instead of the averaging property, it uses Gauss's law and the divergence theorem.

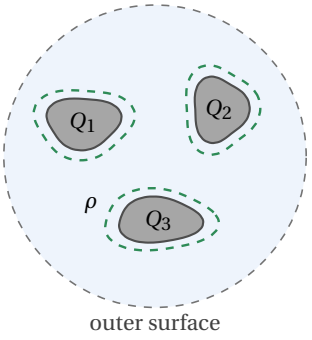


Figure 7.5 The setting of the second uniqueness theorem: conductors carrying given total charges Q_1, Q_2, Q_3, in a region (shaded) where ρ is given, closed off by an outer surface, which may be a conductor or a large sphere. In the proof, Gauss's law is applied to a surface wrapped tightly around each conductor (dashed).

Proof of the second uniqueness theorem

The region is the space between the conductors. It may be closed off on the outside by a conductor that surrounds everything, or it may extend to infinity. Suppose there were two different fields, $\mathbf{E}_1$ and $\mathbf{E}_2$, both satisfying the conditions of the theorem. Each satisfies Gauss's law in the region,

$$\nabla \cdot \mathbf{E}_1 = \frac{\rho}{\varepsilon_0} \qquad \text{and} \qquad \nabla \cdot \mathbf{E}_2 = \frac{\rho}{\varepsilon_0},$$

and each is the gradient of a potential, $\mathbf{E}_1 = -\nabla V_1$ and $\mathbf{E}_2 = -\nabla V_2$. Each conductor is an equipotential, so V_1 and V_2 are constant over the surface of each conductor — although the constants are not known, since only the charges were specified. Finally, apply Gauss's law to a closed surface that wraps tightly around conductor i, just outside its surface. The charge enclosed is the conductor's total charge Q_i, so

$$\oint_i \mathbf{E}_1 \cdot d\mathbf{a} = \frac{Q_i}{\varepsilon_0} \qquad \text{and} \qquad \oint_i \mathbf{E}_2 \cdot d\mathbf{a} = \frac{Q_i}{\varepsilon_0}.$$

Now consider the differences,

$$\mathbf{E}_3 \equiv \mathbf{E}_1 - \mathbf{E}_2 \qquad \text{and} \qquad V_3 \equiv V_1 - V_2,$$

so that $\mathbf{E}_3 = -\nabla V_3$. Subtracting the equations above term by term,

$$\nabla \cdot \mathbf{E}_3 = 0 \qquad \text{in the region,}$$

$$\oint_i \mathbf{E}_3 \cdot d\mathbf{a} = 0 \qquad \text{for each conductor,}$$

and V_3 is constant over the surface of each conductor, because it is the difference of two constants there.

The key step is a product rule for the divergence, which can be checked by writing out the components:

$$\nabla \cdot (V_3 \mathbf{E}_3) = V_3 \, (\nabla \cdot \mathbf{E}_3) + \mathbf{E}_3 \cdot \nabla V_3.$$

In the region the first term is zero, and in the second term $\nabla V_3 = -\mathbf{E}_3$, so

$$\nabla \cdot (V_3 \mathbf{E}_3) = -E_3^2.$$

Integrate both sides over the region, and use the divergence theorem on the left,

$$\oint_{\mathcal{S}} V_3 \mathbf{E}_3 \cdot d\mathbf{a} = -\int E_3^2 \, d\tau, \tag{7.5}$$

where $\mathcal{S}$ is the whole boundary of the region: the surfaces of all the conductors, plus the outer surface. Over the surface of conductor i, V_3 is a constant and comes out of the integral, leaving that constant times $\oint_i \mathbf{E}_3 \cdot d\mathbf{a}$, which is zero. (The direction of $d\mathbf{a}$, which points out of the region and therefore into the conductor, does not matter, since the integral is zero either way.) If the outer surface is a conductor, the same argument applies to it. If instead the region extends to infinity, take the outer surface to be a large sphere of radius r. With V_1 and V_2 both taken to be zero at infinity, V_3 falls off at least as fast as $1/r$ and E_3 at least as fast as $1/r^2$, while the area grows only as r^2, so the integral goes to zero as $r \to \infty$.

Every part of the left side of (7.5) is therefore zero, and

$$\int E_3^2 \, d\tau = 0.$$

The integrand is never negative, so the only way the integral can be zero is for E_3 to be zero everywhere in the region,

$$\mathbf{E}_3 = 0 \qquad \text{throughout the region,}$$

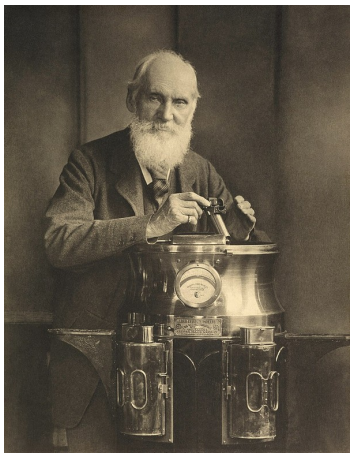
and so $\mathbf{E}_1 = \mathbf{E}_2$.

! **Uniqueness.** With ρ given in a region, the field there is fixed by V on every part of the boundary (first theorem), or by the total charge on each conductor (second theorem).

The proof shows that the *field* is unique. The potential is then unique up to an additive constant, which is fixed in the usual way, for example by $V \to 0$ far away.

7.2 Method of images

The uniqueness theorem says that any function satisfying the right equation and the right boundary conditions is the solution. The **method of images** takes advantage of that directly. If you can picture a *different* arrangement of charges whose potential happens to satisfy the same boundary conditions in the region of interest, then in that region its potential is the answer to the original problem, and it can often simply be written down.



William Thomson, Lord Kelvin (1824–1907, British) Born in Belfast, Thomson entered the University of Glasgow at 10, and in 1846, at 22, he became professor of natural philosophy there, a chair he held for 53 years. He proposed the absolute temperature scale (1848) and devised the method of images. His work on the transatlantic telegraph cable earned him a knighthood in 1866, and in 1892 he became Baron Kelvin, after the river that flows past the University of Glasgow. The first known statement of the curl theorem (Section 1.5.4) is in an 1850 letter from Thomson to Stokes, which is why it is sometimes called the Kelvin–Stokes theorem. (Wikipedia)

7.2.1 A point charge above a grounded plane

The classic example, which appears in essentially every textbook on the subject, is a point charge $+q$ held a distance d above an infinite, grounded, conducting plane. We want the potential everywhere above the plane.

Start with what happens physically. The plane is connected to ground, so charge can flow between it and the ground freely. The positive charge attracts negative charge onto the plane, and the negative charge arranges itself in a very specific way: whatever way is needed to hold the potential of the plane at zero. It is reasonable to expect the induced charge to be bunched up directly below $+q$, where the attraction is strongest, and to thin out with distance from that point. Finding that distribution directly would be hard. The method of images avoids having to.

Consider a different problem: no plane at all, just $+q$ at $z = d$ and a charge $-q$ at $z = -d$, as in Figure 7.6. Every point on the plane $z = 0$ is equidistant from the two charges, so their potentials cancel there, and the potential is zero on the whole plane. (The *field* on the plane is not zero: it points straight down, toward the negative charge. But potential is a scalar, and the two contributions simply add to zero.) The potential of the two charges also goes to zero far away. So the two-charge arrangement has exactly the boundary conditions of the original problem in the region $z > 0$: $V = 0$ on the plane, $V \to 0$ at infinity, and a point charge $+q$ at $z = d$. The charge $-q$ is called the **image charge**.

The uniqueness theorem for Poisson's equation now tells us that, for $z \geq 0$, the potential of the original problem is the potential of the two point charges. The only charge inside the region is $+q$, which is the same in both problems; the image charge lies outside the region, so it does not change the charge density there. With the field point at $\mathbf{r} = x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}}$ and the charges at $\pm d\hat{\mathbf{z}}$, the distances from the two charges to the field point are $\sqrt{x^2 + y^2 + (z-d)^2}$ and $\sqrt{x^2 + y^2 + (z+d)^2}$, so

$$V(x,y,z) = \frac{1}{4\pi\varepsilon_0}\left[\frac{q}{\sqrt{x^2 + y^2 + (z-d)^2}} - \frac{q}{\sqrt{x^2 + y^2 + (z+d)^2}}\right]. \qquad (z \geq 0) \quad (7.6)$$

That is the whole solution. The field above the plane is the field of two point charges, $\mathbf{E} = -\nabla V$, and below the plane the potential and field are both zero.

That last statement also follows from the uniqueness theorem. The region below the plane contains no charge, and it is bounded by the plane, where $V = 0$,

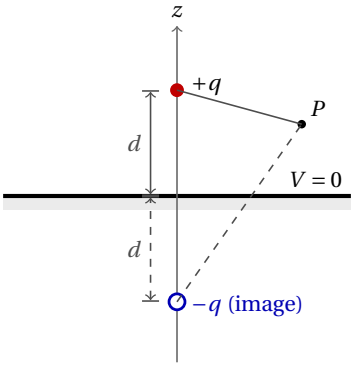


Figure 7.6 A point charge $+q$ a distance d above a grounded conducting plane, and its image $-q$ a distance d below. For points above the plane, the potential of the two point charges is the same as the potential of $+q$ and the charge induced on the plane. Animation of the charge approaching the plane.

and by infinity, where $V = 0$. The function $V = 0$ satisfies Laplace's equation and both boundary conditions, so it is the solution there. The image charge does not exist below the plane; it is only a device for finding the potential above it.

7.2.2 The induced surface charge

Now that we know V, we can find the charge induced on the plane from (7.3). The normal to the plane points up, in the $+z$ direction, so

$$\sigma = -\varepsilon_0 \left.\frac{\partial V}{\partial z}\right|_{z=0}.$$

Differentiating (7.6),

$$\frac{\partial V}{\partial z} = \frac{q}{4\pi\varepsilon_0}\left[\frac{-(z-d)}{\left[x^2+y^2+(z-d)^2\right]^{3/2}} + \frac{z+d}{\left[x^2+y^2+(z+d)^2\right]^{3/2}}\right].$$

At $z = 0$ the two denominators are equal, and the numerators are both d, so

$$\left.\frac{\partial V}{\partial z}\right|_{z=0} = \frac{q}{4\pi\varepsilon_0}\,\frac{2d}{\left(x^2+y^2+d^2\right)^{3/2}}.$$

Multiplying by $-\varepsilon_0$, and writing $s^2 = x^2 + y^2$ for the square of the distance from the point directly below the charge, gives

$$\sigma(s) = -\frac{qd}{2\pi\left(s^2+d^2\right)^{3/2}}. \tag{7.7}$$

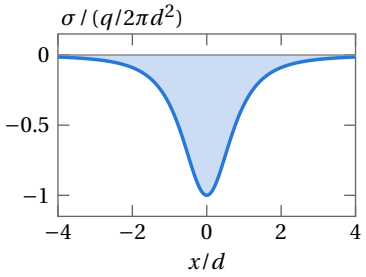


Figure 7.7 The induced surface charge density (7.7) along a line through the point directly below the charge, in units of $q/2\pi d^2$.

The induced charge is negative, largest directly below $+q$, and falls off with distance, as we expected; Figure 7.7 shows its shape. The total induced charge is found by integrating over the plane, using rings of radius s and area $2\pi s\,ds$,

$$\begin{aligned} Q_{\text{ind}} &= \int_0^\infty \sigma(s)\,2\pi s\,ds \\ &= -qd\int_0^\infty \frac{s\,ds}{\left(s^2+d^2\right)^{3/2}} \\ &= -qd\left[-\frac{1}{\sqrt{s^2+d^2}}\right]_0^\infty \\ &= -qd\cdot\frac{1}{d}, \end{aligned}$$

so

$$Q_{\text{ind}} = -q. \tag{7.8}$$

The plane carries exactly the charge of the image, which seems fitting: above the plane, the induced charge produces exactly the field that the image charge would.

The force on the charge $+q$ is also easy to find. It is the force exerted by the induced charge, and above the plane the induced charge acts just like the image charge, so the force is that of $-q$ a distance $2d$ away,

$$\mathbf{F} = -\frac{1}{4\pi\varepsilon_0}\frac{q^2}{(2d)^2}\hat{\mathbf{z}}. \tag{7.9}$$

The charge is attracted toward the plane. By Newton's third law the plane feels an equal and opposite force, and Exercise P7.1 asks you to confirm that by integrating the force on the induced charge directly.

7.2.3 Other image problems

The method works whenever an image arrangement can be found, and for a handful of geometries one can. The simplest extensions use the same plane. Two charges above a grounded plane, say $+q$ and $-q$ at different places, need two images, each the mirror image of its charge with the opposite sign: every point on the plane is equidistant from each charge and its image, so each pair contributes zero potential there. The same idea works for a line charge parallel to a grounded plane. A point charge outside a grounded conducting sphere also has an image, a smaller charge inside the sphere on the line joining the charge to the center: for a sphere of radius R and a charge q a distance a from the center, the image has charge $-qR/a$ and sits a distance R^2/a from the center. Exercise P7.2 asks you to derive this.

In each case, the skill is the same one: visualizing a set of charges, placed outside the region of interest, whose potential matches the boundary conditions. Once that arrangement is found, the uniqueness theorem does the rest.

7.3 Numerical solutions: the relaxation method

The method of images works only for a few special geometries, and separation of variables, in Section 7.4, only for boundaries that lie along the coordinate surfaces of some coordinate system. The **relaxation method** works for any boundary at all. It is a numerical method: rather than a formula, it produces values of V on a grid of points. It rests directly on the averaging property of Section 7.1.2.

7.3.1 Laplace's equation on a grid

Divide the region into a square grid with spacing h, as in Figure 7.8, and label the grid points with integers i and j, so that point (i, j) is at $x = ih$, $y = jh$. Write $V_{i,j}$ for the potential at that point. We want an equation relating $V_{i,j}$ to the potential at its four nearest neighbors.

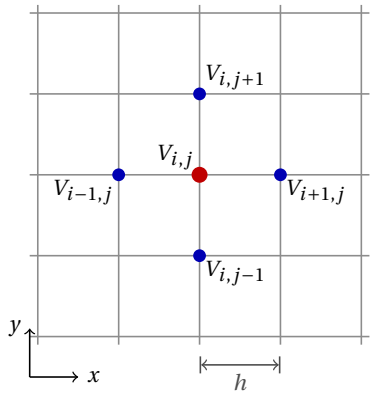


Figure 7.8 A point on the grid, in red, and its four nearest neighbors, in blue. In the relaxation method each point is repeatedly set equal to the average of its four neighbors.

Laplace's equation as an average

Expand the potential at the neighbor to the right in a Taylor series about (x, y),

$$V(x+h,y) = V + h\frac{\partial V}{\partial x} + \frac{h^2}{2}\frac{\partial^2 V}{\partial x^2} + \frac{h^3}{6}\frac{\partial^3 V}{\partial x^3} + \cdots,$$

where V and its derivatives on the right-hand side are evaluated at (x, y). The neighbor to the left is the same with h replaced by $-h$,

$$V(x-h,y) = V - h\frac{\partial V}{\partial x} + \frac{h^2}{2}\frac{\partial^2 V}{\partial x^2} - \frac{h^3}{6}\frac{\partial^3 V}{\partial x^3} + \cdots.$$

Adding the two, the odd-order terms cancel, leaving

$$V(x+h,y) + V(x-h,y) = 2V + h^2\frac{\partial^2 V}{\partial x^2} + \cdots,$$

where the next term is of order h^4. The two neighbors above and below give the same thing with y in place of x. Adding all four,

$$V(x+h,y) + V(x-h,y) + V(x,y+h) + V(x,y-h) = 4V + h^2\left(\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2}\right) + \cdots.$$

The quantity in parentheses is $\nabla^2 V$, which is zero. Dividing by 4 and neglecting terms of order h^4 gives

$$V(x,y) = \tfrac{1}{4}\left[V(x+h,y) + V(x-h,y) + V(x,y+h) + V(x,y-h)\right].$$

In the grid notation the result is

$$V_{i,j} = \tfrac{1}{4}\left(V_{i+1,j} + V_{i-1,j} + V_{i,j+1} + V_{i,j-1}\right). \tag{7.10}$$

This is the discrete version of the averaging property: on a grid, Laplace's equation says that the potential at each point is the average of the potential at its four nearest neighbors. In three dimensions the same argument gives the average of six neighbors. If there is charge in the region, Poisson's equation adds a term: in two dimensions the right-hand side of (7.10) becomes $\frac{1}{4}(\ldots) + \rho h^2/4\varepsilon_0$.

7.3.2 The method

Equation (7.10) is one equation for every interior grid point, so a grid of a few thousand points gives a few thousand simultaneous equations. Rather than solving them all at once, the relaxation method satisfies them by repetition.

1. Set up the grid, and fix the points on the boundaries at their specified potentials. These points are never changed.

2. Give every other point an initial guess. It does not need to be a good one; $V = 0$ works.

3. Go through the interior points one at a time, replacing each value by the average of its four neighbors, as in (7.10). This is called "relaxing" the grid. While doing it, keep track of the largest change made to any point.

4. Repeat step 3 until the largest change in a pass through the grid is smaller than some chosen tolerance.

When the loop stops, every point is the average of its neighbors to within the tolerance, so the grid values satisfy Laplace's equation, and they match the boundary conditions because the boundary points were never changed. By the uniqueness theorem, the result is the potential for the given boundary conditions, to the accuracy that the grid spacing and the tolerance allow.

There are two common ways to carry out step 3. The **Jacobi method** computes all the new values from the old ones and then replaces the whole grid at once. The **Gauss–Seidel method** replaces each value as soon as it is computed, so that later points in the same pass already use the updated values of their neighbors. For Laplace's equation, Gauss–Seidel needs about half as many passes as Jacobi to reach the same tolerance, and it is no harder to program. More sophisticated techniques, such as successive over-relaxation, converge much faster still, but the simple methods are enough for the problems in this book. More details and examples for solving Poisson's equation (of which Laplace's equation is a special case), including successive over-relaxation, can be found in Chapter 9 of R. L. Spencer and M. Ware, *Computational Methods for Scientists: Partial Differential Equations*, which is available as a free download.[1]

7.3.3 An example, and a caution about the tolerance

As a test, the relaxation method can be applied to the region of Example 7.1: two grounded plates at $y = 0$ and $y = a$, closed off at $x = 0$ by a strip held at 100 V. The region extends to $x \to \infty$, which a grid cannot do, so the grid was ended at $x = 3a$ with $V = 0$ there. That changes very little, because by $x = 3a$ the true potential has fallen to about 0.01 V. With a grid spacing of $h = a/20$, the Jacobi method needed roughly 1600 passes to reach a tolerance of 10^{-6} V, and Gauss–Seidel about 840.

Figure 7.9 compares the result with the exact series solution that will be found in Section 7.4.1. The two agree closely. At the center of the square $0 \le x, y \le a$, the relaxation method gives 26.104 V and the series gives 26.096 V, a difference of about 0.03%; the largest difference, about 0.7 V, occurs close to the corners where the potential jumps from 100 V to 0. Halving the grid spacing reduces the differences further.

The tolerance needs more care than it might seem to. Each pass changes the values by only a little, but many small changes add up, so a loop can stop while the grid is still far from relaxed. In the same calculation, stopping when no point changed by more than 0.01 V in a pass gave 25.08 V at the center, which is 4% too low — much larger than 0.01 V (Figure 7.10). Since you do not know in advance how small the tolerance must be, a practical approach is to start with a large tolerance and reduce it in steps, watching the solution. When reducing

[1] From https://physics.byu.edu/courses/computational/phys430. Direct link to the PDF: https://physics.byu.edu/courses/computational/docs/phys430/phys430python.pdf.

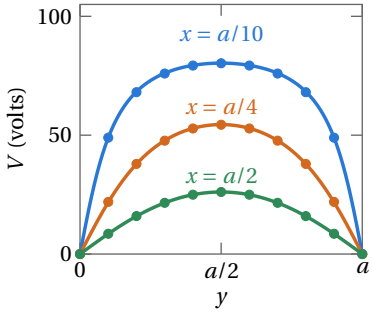


Figure 7.9 The potential in the region of Example 7.1, plotted against y at three distances from the 100 V strip. Curves: the series solution (7.26). Dots: the relaxation method, with grid spacing $h = a/20$ (every second grid point is shown).

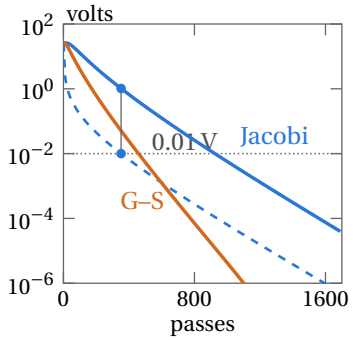


Figure 7.10 The relaxation calculation of Section 7.3.3 pass by pass, on a logarithmic scale: the error in the potential at the center of the square (solid; Jacobi in blue, Gauss–Seidel in orange), and the largest change made to any point in a Jacobi pass (dashed). When that change first falls below 0.01 V (dotted), the center is still about 1 V too low. Animation of the two methods, pass by pass.

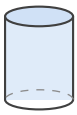
the tolerance by another factor of 2 no longer changes the solution visibly, the tolerance is small enough.



(a) planes
x, y, or z = const



(b) sphere r = const,
cone θ = const



(c) cylinder
s = const

Figure 7.11 Boundaries on which one coordinate is constant, where separation of variables works: (a) planes, such as the faces of a box; (b) a sphere, r = const, and a cone, θ = const; (c) a cylinder, s = const.

7.4 Separation of variables

Separation of variables turns Laplace's equation, a partial differential equation, into a set of ordinary differential equations, one for each coordinate. It is the method of choice when the boundaries of the region lie on surfaces where one coordinate is constant: planes in Cartesian coordinates, spheres and cones in spherical coordinates, cylinders in cylindrical coordinates (Figure 7.11). The solutions it produces are the orthogonal functions of Chapter 6 — sines and cosines in Cartesian coordinates, Legendre polynomials in spherical coordinates, and sines and cosines of the angle ϕ in cylindrical coordinates — and Fourier's trick is what fits them to the boundary conditions.

The method rests on the uniqueness theorem of Section 7.1.3. We guess a solution of a particular form, and if the guess satisfies both Laplace's equation and the boundary conditions, uniqueness guarantees that it is *the* solution. There is no other one to find.

7.4.1 2d Cartesian coordinates

We begin in two dimensions, with problems in which nothing depends on z. Example 7.1 below is a typical one. First we will find the general solution of Laplace's equation in two dimensions; then we will apply it to that particular situation.

In two dimensions, Laplace's equation is

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = 0. \tag{7.11}$$

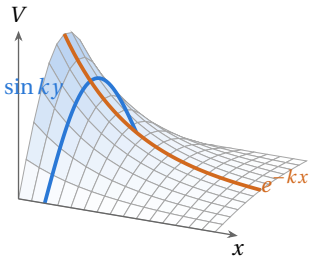


Figure 7.12 One product solution, $V = e^{-kx}\sin ky$, drawn as a height above the xy plane, with y running into the page. Along every line of constant x it is a sine of y (blue), and along every line of constant y it is an exponential in x (orange).

The key step is to look for solutions that are a product of a function of x alone and a function of y alone (Figure 7.12),

$$V(x,y) = X(x)\,Y(y). \qquad \text{(separation of variables)} \tag{7.12}$$

Nothing guarantees in advance that solutions of this form exist. We will find that they do, and that sums of them are enough to solve the problems in this section.

Separating Laplace's equation in two dimensions

Substituting (7.12) into (7.11) gives

$$Y\frac{d^2X}{dx^2} + X\frac{d^2Y}{dy^2} = 0.$$

The partial derivatives have become ordinary derivatives, because X depends only on x and Y depends only on y. Dividing through by XY gives

$$\frac{1}{X}\frac{d^2X}{dx^2} + \frac{1}{Y}\frac{d^2Y}{dy^2} = 0,$$

and moving the second term to the right-hand side gives

$$\frac{1}{X}\frac{d^2X}{dx^2} = -\frac{1}{Y}\frac{d^2Y}{dy^2}. \tag{7.13}$$

The equation is now "separated": the left-hand side depends only on x, and the right-hand side depends only on y.

That is a strong condition. The two sides must be equal for every x and every y, with x and y free to vary independently, and the only way that can happen is for both sides to equal the same constant. To see why, consider an equation such as $\cos x = e^y$. It is true for some pairs of x and y, but there is no way it can be true for all x and all y. Adding constants does not help: the equation

$$C_1 + C_2\cos x = C_3 + C_4 e^y$$

can be true for all x and all y only if $C_2 = C_4 = 0$ and $C_1 = C_3$ — that is, only if both sides are the same constant.

So each side of (7.13) must equal a constant, the same constant for both. That constant could be positive or negative. For now, let it be positive, and to make that explicit, call it k^2. Equation (7.13) then splits into two ordinary differential equations,

$$\frac{d^2X}{dx^2} = k^2X \qquad \text{and} \qquad \frac{d^2Y}{dy^2} = -k^2Y. \tag{7.14}$$

These are standard second-order equations, and each has two independent solutions,

$$X = \begin{cases} e^{kx} \\ e^{-kx} \end{cases} \qquad \text{and} \qquad Y = \begin{cases} \sin ky \\ \cos ky \end{cases}, \tag{7.15}$$

or any linear combination of the two.[2] You can check each one by substituting it back into (7.14).

If we had chosen a negative constant, the only difference would be that X would have been sines and cosines and Y would have been exponentials. This is the key: when doing separation of variables in 2d, in rectangular coordinates, one variable results in positive and negative exponentials, and the other results in sines and cosines. You need to select which is which based on the specifics of the problem. In summary, with the understanding that X and Y can be reversed,

$$\boxed{X = \begin{cases} e^{kx} \\ e^{-kx} \end{cases} \text{ or } \begin{cases} \cosh kx \\ \sinh kx \end{cases}, \qquad Y = \begin{cases} \sin ky \\ \cos ky \end{cases}}. \qquad \text{(2d Cartesian)} \tag{7.16}$$

[2] A separation constant of exactly zero is also allowed. It gives $X = A + Bx$ and $Y = C + Dy$, which are occasionally needed but not for the problems in this section.

The hyperbolic functions are combinations of the exponentials,

$$\cosh kx = \tfrac{1}{2}\left(e^{kx} + e^{-kx}\right), \qquad \sinh kx = \tfrac{1}{2}\left(e^{kx} - e^{-kx}\right),$$

and they are often the more convenient choice when the region is finite in that direction.

We can now apply these general solutions to Example 7.1.

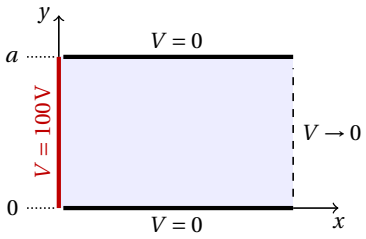


Figure 7.13 The region for Example 7.1: two grounded plates at $y = 0$ and $y = a$, closed off at $x = 0$ by a strip held at 100V. Nothing depends on z.

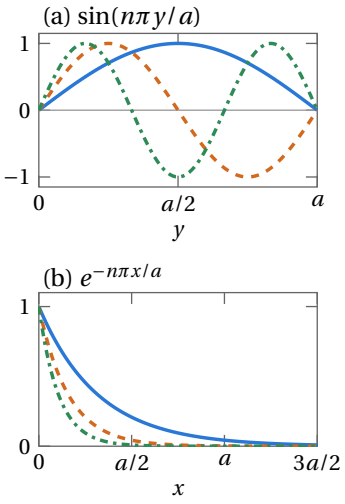
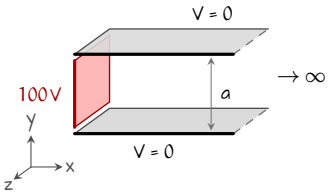


Figure 7.14 The two kinds of separated solutions in Example 7.1, for the first three values of n that BC4 will allow, $n = 1$ (solid), 2 (dashed), and 3 (dash-dot). (a) $\sin(n\pi y/a)$ vanishes on both plates, which an exponential cannot do; (b) $e^{-n\pi x/a}$ dies away as $x \to \infty$, which a sine cannot do.

Example 7.1

Find the potential $V(x, y)$ in the region shown in Figure 7.13.

Solution:



Our first sketch is three-dimensional, to show what Figure 7.13 stands for. Since nothing depends on z, the figure is a cross-section of a structure that extends infinitely into and out of the page: two grounded plates at $y = 0$ and $y = a$, running from $x = 0$ out to $x \to \infty$, closed off at $x = 0$ by a strip that is insulated from them and held at 100V.

Let's be explicit about the four boundary conditions:

$$\begin{aligned} &\text{BC1:} && V(x = 0) = 100\text{V}, \\ &\text{BC2:} && V(x \to \infty) = 0, \\ &\text{BC3:} && V(y = 0) = 0, \\ &\text{BC4:} && V(y = a) = 0. \end{aligned}$$

First recognize that we will want sines and cosines in the y direction and exponentials in the x direction. The potential has to vanish at both $y = 0$ and $y = a$, which a sine can do and an exponential cannot, and it has to die away as $x \to \infty$, which an exponential can do and a sine cannot (Figure 7.14). Next, consider the boundary conditions one at a time.

BC2. This means that we can throw out the $X = e^{kx}$ solution, which grows without bound as $x \to \infty$.

BC3. This means we can throw out the $Y = \cos ky$ solution, since $\cos 0 = 1$.

Already we've made great progress! Our function therefore has the form

$$V(x, y) = Ce^{-kx} \sin ky, \qquad \text{(satisfies BC2 and BC3)} \quad (7.17)$$

where C is some arbitrary constant. Incidentally, it might seem more intuitive to have chosen the origin to be the midpoint in the y direction, but if we had done that we would have had to include both sines and cosines in our answer.

BC4. This means that V must vanish at $y = a$ for every x, so

$$Ce^{-kx}\sin ka = 0. \qquad \text{(imposing BC4)} \quad (7.18)$$

The exponential is never zero, and choosing $C = 0$ would make V zero everywhere, so the only way to make that happen is to force $\sin ka = 0$. That requires $ka = n\pi$, which means

$$k = \frac{n\pi}{a}, \qquad n = 1, 2, 3, \ldots$$

(Negative values of n give nothing new, since $\sin(-ky) = -\sin ky$, and $n = 0$ gives $V = 0$.) So now we know our solution has the form

$$V(x, y) = Ce^{-n\pi x/a}\sin(n\pi y/a). \qquad \text{(satisfies BC2, BC3, and BC4)} \quad (7.19)$$

How about BC1? Let's try enforcing it the same way we did BC4, by setting $x = 0$ in (7.19) and requiring $V = 100\,\text{V}$. That means

$$100\,\text{V} = C\sin(n\pi y/a). \qquad \text{(attempt to impose BC1, failure)} \quad (7.20)$$

Is that true for every y? No. Is there any choice of C which can make it true? Also no. But does this mean we must give up? Also no!

Equation (7.19) is not just true for one n; it's true for any n. And since Laplace's equation is linear, linear combinations of solutions are also solutions. So actually, the general form of our solution is not (7.19) but rather

$$V(x, y) = \sum_{n=1}^{\infty} C_n e^{-n\pi x/a}\sin(n\pi y/a), \qquad \text{(general form)} \quad (7.21)$$

where C is now labeled with n because it can take a different value for each n.

So, back to BC1. The question is no longer "is there a choice of C which can make it true?" The question is now "is there a special combination of C_n values which can make it true?" And the answer to that question is yes! (Figure 7.15 shows such a combination at work.)

Setting $x = 0$ in (7.21), we need

$$100\,\text{V} = \sum_{n=1}^{\infty} C_n \sin(n\pi y/a). \qquad \text{(imposing BC1)} \quad (7.22)$$

To find the C_n values, use Fourier's trick (Section 6.1.3): multiply both sides by the orthogonal function $\sin(m\pi y/a)$, integrate over y from 0 to a, then use the orthogonality condition on the right-hand side. You can bring the integral inside the summation because the two are acting on different variables. That gives us

$$\int_0^a (100\,\text{V})\sin(m\pi y/a)\,dy = \sum_{n=1}^{\infty} C_n \int_0^a \sin(n\pi y/a)\sin(m\pi y/a)\,dy.$$

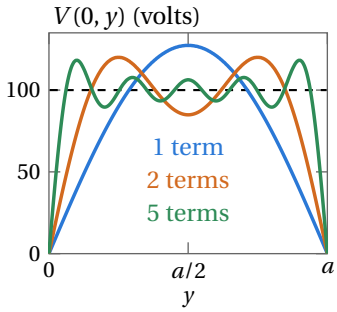


Figure 7.15 At $x = 0$ the sum (7.21) must equal 100 V for every y between the plates (dashed). With the coefficients (7.24) found below, the sums of its first one, two, and five nonzero terms come closer and closer to it, except right next to the corners.

The integrals on the right-hand side are given by the orthogonality relation for these sines,

$$\int_0^a \sin(n\pi y/a)\sin(m\pi y/a)\,dy = \begin{cases} 0, & n \neq m, \\ a/2, & n = m. \end{cases} \tag{7.23}$$

This is the same kind of relation as (6.8), except that the interval from 0 to a is half a period of the lowest sine, $\sin(\pi y/a)$, rather than a full period. (Why does the relation still hold over only half a period of the lowest sine? Answer in footnote.[3])

Because of orthogonality, the terms in the infinite summation on the right-hand side are zero for all of the terms except one. That's the one where $n = m$. For that term we are integrating $\sin^2(m\pi y/a)$, and that integral equals $a/2$. Thus we have

$$\int_0^a (100\,\mathrm{V})\sin(m\pi y/a)\,dy = C_m\frac{a}{2}.$$

The integral on the left-hand side is easily done,

$$\begin{aligned}\int_0^a (100\,\mathrm{V})\sin(m\pi y/a)\,dy &= (100\,\mathrm{V})\left[-\frac{a}{m\pi}\cos(m\pi y/a)\right]_0^a \\ &= (100\,\mathrm{V})\,\frac{a}{m\pi}\,(1-\cos m\pi)\,,\end{aligned}$$

giving us

$$(100\,\mathrm{V})\,\frac{a}{m\pi}\,(1-\cos m\pi) = C_m\frac{a}{2}.$$

Thus we can solve for C_m, which I'm now going to relabel as C_n because m and n are just dummy index variables,

$$C_n = 2\,(100\,\mathrm{V})\left(\frac{1-\cos n\pi}{n\pi}\right). \tag{7.24}$$

This is telling us that there is a specific set of coefficients C_n which will give us BC1. Since $\cos n\pi = (-1)^n$, the coefficients are zero for even n and equal to $400\,\mathrm{V}/n\pi$ for odd n. The first few are listed here in units of 100 V.

n	$C_n/(100\,\mathrm{V})$
1	$4/\pi$
2	0
3	$4/3\pi$
4	0
5	$4/5\pi$
6	0

[3] Each $\sin(n\pi y/a)$ is odd in y, so the product of two of them is even, and its integral from 0 to a is half its integral from $-a$ to a. That interval is a full period of $\sin(\pi y/a)$, so (6.8), with $T = 2a$, gives 0 for $n \neq m$ and a for $n = m$; half of these are the values in (7.23).

Written out, then, our solution is

$$V(x,y) = (100\,\mathrm{V})\Big[\frac{4}{\pi}\,e^{-\pi x/a}\sin(\pi y/a) + \frac{4}{3\pi}\,e^{-3\pi x/a}\sin(3\pi y/a) + \frac{4}{5\pi}\,e^{-5\pi x/a}\sin(5\pi y/a) + \frac{4}{7\pi}\,e^{-7\pi x/a}\sin(7\pi y/a) + \cdots\Big],$$

(BC1 now imposed, success!) (7.25)

or, more compactly,

$$V(x,y) = \frac{400\,\mathrm{V}}{\pi}\sum_{n=1,3,5,\ldots}\frac{1}{n}\,e^{-n\pi x/a}\sin(n\pi y/a). \tag{7.26}$$

It satisfies Laplace's equation and it now satisfies all four boundary conditions. So, by the uniqueness theorem of Section 7.1.3, it is the correct and unique solution to our problem.

Reality check: The slot is symmetric about its midplane, $y = a/2$, so the potential must be too, and it is: for odd n, $\sin\big(n\pi(a-y)/a\big) = \sin(n\pi y/a)$. The even-$n$ terms, which change sign across the midplane, are exactly the ones whose coefficients came out zero.

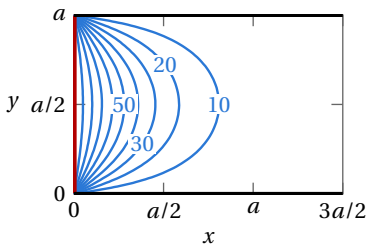


Figure 7.16 Equipotentials of the solution (7.26) for Example 7.1, every 10 V from 90 V (nearest the 100 V strip at $x = 0$) down to 10 V. All of them meet at the two corners, where the boundary potential jumps from 100 V to 0.

Figure 7.16 shows the equipotentials of this solution. The potential dies away quickly as x increases: at $x = 3a/2$, midway between the plates, it has already fallen to about 1 V.

7.4.2 3d Cartesian coordinates

In three dimensions the method works the same way, with one more function and one more separation constant. Laplace's equation is

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = 0, \tag{7.27}$$

and we look for solutions of the form

$$V(x,y,z) = X(x)\,Y(y)\,Z(z). \qquad \text{(separation of variables)} \tag{7.28}$$

Separating Laplace's equation in three dimensions

Substituting the product into (7.27) and dividing by XYZ gives

$$\frac{1}{X}\frac{d^2X}{dx^2} + \frac{1}{Y}\frac{d^2Y}{dy^2} + \frac{1}{Z}\frac{d^2Z}{dz^2} = 0.$$

Move the last two terms to the right-hand side,

$$\frac{1}{X}\frac{d^2X}{dx^2} = -\left(\frac{1}{Y}\frac{d^2Y}{dy^2} + \frac{1}{Z}\frac{d^2Z}{dz^2}\right).$$

The left-hand side is a function of x only, and the right-hand side does not depend on x at all. By the same argument as in two dimensions, both sides must equal a

constant. Call it $-k_x^2$, a negative number, so that the x solutions will be sines and cosines. Then

$$\frac{d^2X}{dx^2} = -k_x^2 X \qquad \text{and} \qquad \frac{1}{Y}\frac{d^2Y}{dy^2} + \frac{1}{Z}\frac{d^2Z}{dz^2} = k_x^2.$$

The second of these can be separated again,

$$\frac{1}{Y}\frac{d^2Y}{dy^2} = k_x^2 - \frac{1}{Z}\frac{d^2Z}{dz^2},$$

where the left-hand side depends only on y and the right-hand side only on z, so both equal a second constant. Call it $-k_y^2$. That leaves three ordinary differential equations,

$$\frac{d^2X}{dx^2} = -k_x^2 X, \qquad \frac{d^2Y}{dy^2} = -k_y^2 Y, \qquad \frac{d^2Z}{dz^2} = \left(k_x^2 + k_y^2\right) Z. \tag{7.29}$$

The first two have sines and cosines as their solutions, and because $k_x^2 + k_y^2$ is positive, the third has exponentials.

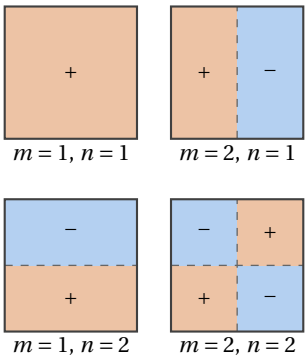


Figure 7.17 The x–y parts of four product solutions, $\sin(m\pi x/a)\sin(n\pi y/a)$, on the square $0 \le x, y \le a$, positive (orange) and negative (blue). In the third direction each one grows or decays with the constant $\kappa = (\pi/a)\sqrt{m^2+n^2}$ of (7.31): the more nodal lines (dashed), the faster.

Two of the directions get sines and cosines and the third gets exponentials (Figure 7.17). Which direction is which is up to you: with different choices of sign for the two separation constants, any one of the three directions can be the one with exponentials. Choose based on the boundary conditions. In summary, with the understanding that X, Y, and Z can be permuted,

$$X = \begin{cases} \sin k_x x \\ \cos k_x x \end{cases}, \quad Y = \begin{cases} \sin k_y y \\ \cos k_y y \end{cases}, \quad Z = \begin{cases} e^{\kappa z} \\ e^{-\kappa z} \end{cases} \text{ or } \begin{cases} \sinh \kappa z \\ \cosh \kappa z \end{cases}, \tag{3d Cartesian} \tag{7.30}$$

where

$$\kappa = \sqrt{k_x^2 + k_y^2}. \tag{7.31}$$

Notice that the exponentials are not independent of the sines and cosines: their decay constant is fixed by the two oscillation constants.

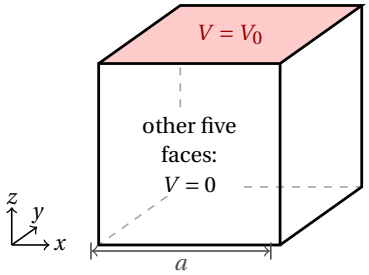
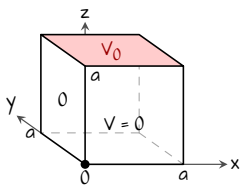


Figure 7.18 The cube of Example 7.2. The top face is held at V_0 and the other five faces are grounded.

Example 7.2

A hollow cube of side a has five of its faces grounded, while the sixth, the top face at $z = a$, is insulated from the others and held at a potential V_0, as shown in Figure 7.18. Find the potential inside the cube.

Solution:



We draw the cube of Figure 7.18 again, this time from the left, and add coordinate axes. Put the origin at one corner of the cube, so that the cube occupies $0 \le x, y, z \le a$.

There are six boundary conditions:

$$\begin{aligned} &\text{BC1:} \quad V(x=0)=0, &&\text{BC4:} \quad V(y=a)=0,\\ &\text{BC2:} \quad V(x=a)=0, &&\text{BC5:} \quad V(z=0)=0,\\ &\text{BC3:} \quad V(y=0)=0, &&\text{BC6:} \quad V(z=a)=V_0. \end{aligned}$$

The potential has to vanish on both sides of the cube in the x direction and on both sides in the y direction, which sines can do; so the x and y directions get the sines and cosines, and z gets the exponentials.

BC1 and BC3 rule out the cosines, which are not zero at $x = 0$ or $y = 0$. BC2 then requires $\sin k_x a = 0$, and BC4 requires $\sin k_y a = 0$. Just as in Example 7.1, that means

$$k_x = \frac{m\pi}{a} \qquad \text{and} \qquad k_y = \frac{n\pi}{a}, \qquad m, n = 1, 2, 3, \ldots,$$

with m and n two independent integers. (Why are m and n two independent integers, rather than a single one? Answer in footnote.[4]) So

$$X(x)\,Y(y) = \sin(m\pi x/a)\sin(n\pi y/a).$$

The four side faces are taken care of by the choice of functions alone.

Now the z direction, where by (7.31) the decay constant is

$$\kappa = \sqrt{k_x^2 + k_y^2} = \frac{\pi}{a}\sqrt{m^2+n^2}.$$

We need a function of z that is zero at $z = 0$ (BC5) and nonzero at $z = a$. Neither exponential can do that by itself, since neither $e^{\kappa z}$ nor $e^{-\kappa z}$ is ever zero. A linear combination can: with $Z = Ce^{\kappa z} + De^{-\kappa z}$, BC5 requires $C + D = 0$, so

$$Z = C\left(e^{\kappa z} - e^{-\kappa z}\right) = 2C\sinh\kappa z.$$

That is why the hyperbolic functions are included in (7.30): one of them is often exactly the combination a problem requires (Figure 7.19). The constant $2C$ is just another

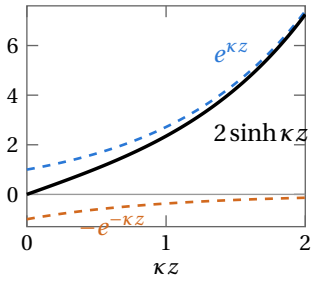


Figure 7.19 Neither $e^{\kappa z}$ nor $-e^{-\kappa z}$ (dashed) is zero at $z = 0$, but their sum, $2\sinh\kappa z$ (solid), is, and farther from $z = 0$ it grows like $e^{\kappa z}$.

[4] BC2 restricts only $X(x)$, and BC4 only $Y(y)$, so each condition fixes its own integer, and every pair m, n gives a product solution. All the pairs are needed: the coefficients C_{mn} found below are nonzero for every pair of odd integers, such as $m = 1$ and $n = 3$, not only for $m = n$.

arbitrary constant, so each product solution has the form

$$V(x,y,z) = C_{mn}\sin(m\pi x/a)\sin(n\pi y/a)\sinh\!\left(\pi\sqrt{m^2+n^2}\,z/a\right),$$

and, since sums of solutions are also solutions, the general form is a double sum over both integers,

$$V(x,y,z) = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} C_{mn}\sin(m\pi x/a)\sin(n\pi y/a)\sinh\!\left(\pi\sqrt{m^2+n^2}\,z/a\right). \tag{7.32}$$

That leaves BC6. Setting $z = a$ and $V = V_0$,

$$V_0 = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} C_{mn}\sinh\!\left(\pi\sqrt{m^2+n^2}\right)\sin(m\pi x/a)\sin(n\pi y/a).$$

(imposing BC6) (7.33)

At $z = a$ the hyperbolic sine is just a number for each m and n. To find the coefficients, use Fourier's trick twice: multiply both sides by $\sin(m'\pi x/a)\sin(n'\pi y/a)$, and integrate over x from 0 to a and over y from 0 to a. On the right-hand side the x integral is zero unless $m = m'$, and the y integral is zero unless $n = n'$, by (7.23). Only one term of the double sum survives, and its two integrals each give $a/2$:

$$V_0\int_0^a \sin(m'\pi x/a)\,dx\int_0^a \sin(n'\pi y/a)\,dy = C_{m'n'}\sinh\!\left(\pi\sqrt{m'^2+n'^2}\right)\frac{a}{2}\cdot\frac{a}{2}.$$

Each integral on the left is the one we did in Example 7.1,

$$\int_0^a \sin(m'\pi x/a)\,dx = \frac{a}{m'\pi}\left(1-\cos m'\pi\right),$$

so, dropping the primes,

$$V_0\,\frac{a}{m\pi}\left(1-\cos m\pi\right)\frac{a}{n\pi}\left(1-\cos n\pi\right) = C_{mn}\sinh\!\left(\pi\sqrt{m^2+n^2}\right)\frac{a^2}{4}.$$

Each factor $1-\cos m\pi$ is 2 for odd m and 0 for even m, so

$$C_{mn} = \begin{cases} \dfrac{16V_0}{\pi^2 mn\sinh\!\left(\pi\sqrt{m^2+n^2}\right)}, & m \text{ and } n \text{ both odd,} \\ 0, & \text{otherwise.} \end{cases} \tag{7.34}$$

Putting these into (7.32), the potential inside the cube is

$$V(x,y,z) = \frac{16V_0}{\pi^2}\sum_{m=1,3,5,\ldots}\ \sum_{n=1,3,5,\ldots}\frac{1}{mn}\sin(m\pi x/a)\sin(n\pi y/a) \times \frac{\sinh\!\left(\pi\sqrt{m^2+n^2}\,z/a\right)}{\sinh\!\left(\pi\sqrt{m^2+n^2}\right)}. \tag{7.35}$$

It satisfies Laplace's equation and all six boundary conditions, so by the uniqueness theorem it is the solution.

Reality check: The uniqueness theorem gives a check as well. Imagine six separate problems, each with a different face of the cube at V_0 and the other five grounded. Each solution is (7.35) with the coordinates relabeled. Adding all six gives a function that satisfies Laplace's equation and equals V_0 on every face. The constant $V = V_0$ also does that, so by uniqueness the sum of the six solutions is exactly V_0 everywhere inside. At the center of the cube all six problems look the same, so each must contribute one sixth: at the center, (7.35) should give $V_0/6$. Summing the series numerically confirms it. For $V_0 = 100\,\text{V}$ the sum converges to 16.667 V, and the first term alone, $m = n = 1$, gives 17.4 V.

The first term of (7.35), $m = n = 1$, is a reasonable picture of the whole solution away from the top face. The ratio of hyperbolic sines falls off roughly as $e^{-\pi\sqrt{m^2+n^2}\,(1-z/a)}$, so the higher terms die away quickly with distance below the top. The potential is largest in the middle of each horizontal slice and falls to zero at the walls, and it weakens and spreads out toward the bottom of the cube (Figure 7.20). If the potential were instead a temperature — the top face hot and the other faces held cold — the same equation and the same solution would describe how the heat reaches farther down in the center of the cube than near its sides.

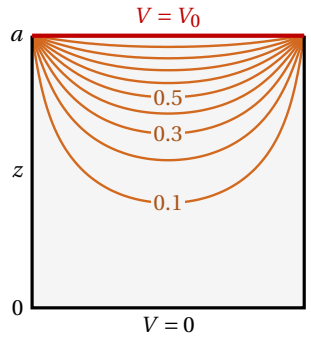


Figure 7.20 Equipotentials of the cube of Example 7.2 in its vertical midplane $y = a/2$, at $V/V_0 = 0.1, 0.2, \ldots, 0.9$, from (7.35). The potential reaches farthest down in the middle; at the center of the cube it is $V_0/6$.

7.4.3 Spherical coordinates, no ϕ dependence

When the boundaries are spheres, spherical coordinates are the natural choice. We will treat only problems with azimuthal symmetry, in which nothing depends on ϕ. That covers a great many problems, since a spherical boundary with a potential or charge that varies only with polar angle has that symmetry.

With no ϕ dependence, Laplace's equation in spherical coordinates is

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial V}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\,\frac{\partial V}{\partial\theta}\right) = 0, \tag{7.36}$$

and we look for solutions of the form

$$V(r,\theta) = R(r)\,\Theta(\theta). \qquad \text{(separation of variables)} \tag{7.37}$$

Here $R(r)$ is the radial function. In the examples below the radius of a sphere is also called R, as is common; when R appears without an argument, it is the radius.

Separating Laplace's equation in spherical coordinates

Substituting $R(r)\,\Theta(\theta)$ into (7.36), multiplying by r^2, and dividing by $R\Theta$ gives

$$\frac{1}{R}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) = -\frac{1}{\Theta\sin\theta}\frac{d}{d\theta}\left(\sin\theta\,\frac{d\Theta}{d\theta}\right).$$

The left-hand side depends only on r and the right-hand side only on θ, so both equal a constant. Instead of calling it k^2, we will call it $\ell(\ell+1)$, for a reason that will appear in a moment. The two ordinary differential equations are then

$$\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) = \ell(\ell+1)\,R \qquad \text{and} \qquad \frac{d}{d\theta}\left(\sin\theta\,\frac{d\Theta}{d\theta}\right) = -\ell(\ell+1)\sin\theta\,\Theta. \tag{7.38}$$

Start with the radial equation. Try a power of r, $R = r^p$. Then $r^2\,dR/dr = p\,r^{p+1}$, and

$$\frac{d}{dr}\left(p\,r^{p+1}\right) = p(p+1)\,r^p,$$

so $R = r^p$ is a solution if $p(p+1) = \ell(\ell+1)$. That quadratic equation for p has two roots, $p = \ell$ and $p = -(\ell+1)$, and the two solutions

$$R = r^\ell \qquad \text{and} \qquad R = \frac{1}{r^{\ell+1}}$$

are independent, so they are the two solutions of this second-order equation. Writing the constant as $\ell(\ell+1)$ is what makes these roots come out so simply.

The angular equation is Legendre's equation, in exactly the form (6.13) of Chapter 6. Its two solutions are $P_\ell(\cos\theta)$ and $Q_\ell(\cos\theta)$. It can be shown that a solution which stays finite at both $\theta = 0$ and $\theta = \pi$ exists only when ℓ is a non-negative integer, and then it is the Legendre polynomial $P_\ell(\cos\theta)$; the Q_ℓ are infinite on the z axis.

In summary,

$$\boxed{R = \begin{cases} r^\ell \\ \dfrac{1}{r^{\ell+1}} \end{cases}, \qquad \Theta = \begin{cases} P_\ell(\cos\theta) \\ Q_\ell(\cos\theta) \end{cases}} \quad \text{(spherical, no } \phi \text{ dependence)} \tag{7.39}$$

For any problem whose region includes the z axis, the Q_ℓ solutions are discarded. Taking linear combinations of the remaining products, the general solution is

$$V(r,\theta) = \sum_{\ell=0}^{\infty}\left(A_\ell\,r^\ell + \frac{B_\ell}{r^{\ell+1}}\right)P_\ell(\cos\theta). \tag{7.40}$$

This is a sum over the integer ℓ, just as the Cartesian solutions were sums over n. It is the starting point for every problem in this section.

Which terms survive depends on the region (Figure 7.21).

- *Inside a sphere*, the region includes $r = 0$, where every $1/r^{\ell+1}$ is infinite. So $B_\ell = 0$ for all ℓ.
- *Outside a sphere*, with $V \to 0$ as $r \to \infty$, every r^ℓ term with $\ell \geq 1$ grows without bound and the $\ell = 0$ term is a nonzero constant. So $A_\ell = 0$ for all ℓ.
- *Between two spheres*, neither $r = 0$ nor $r \to \infty$ is in the region, and both sets of terms are kept.

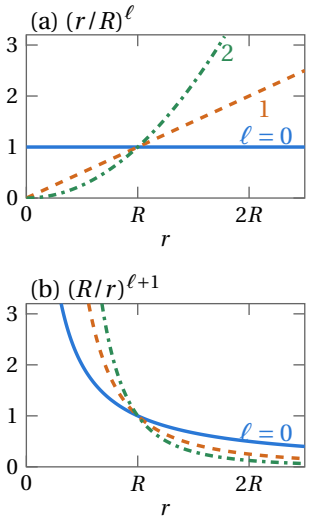


Figure 7.21 The radial solutions in (7.39) for $\ell = 0$ (solid), 1 (dashed), and 2 (dash-dot). (a) r^ℓ is finite at the origin but grows without bound; (b) $1/r^{\ell+1}$ goes to zero far away but is infinite at the origin.

- If the potential does *not* go to zero far away, as when an object sits in a uniform applied field, some of the r^ℓ terms survive outside, and the condition at infinity determines them.

To find the coefficients that remain, the boundary function has to be written as a sum of Legendre polynomials. Fourier's trick does that in general, using the orthogonality relation (6.15) and the coefficient formula (6.16). But in many problems the boundary function is already a polynomial in $\cos\theta$, or can be made into one with a trigonometric identity, and then it can be written as a finite combination of Legendre polynomials by inspection. Once both sides of a boundary condition are written as sums of $P_\ell(\cos\theta)$, the coefficient of each P_ℓ on one side must equal the coefficient of the same P_ℓ on the other side, because the expansion in Legendre polynomials is unique. This is **Colton's trick** (Section 6.2.4). It turns an infinite series into a finite one, often with only one or two terms, and it requires no integrals.

One caution: the functions that appear in (7.40) are Legendre polynomials of $\cos\theta$, so the boundary function has to be written in terms of $\cos\theta$, not of $\cos 2\theta$ or of $\cos(\theta/2)$ or anything else.

Three types of problems

Three types of spherical boundary value problems come up often enough to be worth working through one of each. In the first, the potential is specified on a sphere. In the second, a conductor sits in a field that is specified far away. In the third, the charge density is specified on a spherical shell. The examples that follow take them in that order (Figure 7.22).

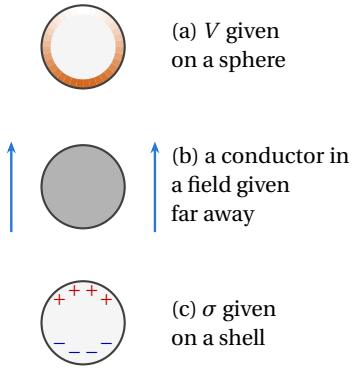


Figure 7.22 The three types of spherical problems of Examples 7.3–7.5: (a) the potential given on a sphere; (b) a conductor in a field that is given far away; (c) a charge density given on a shell.

Example 7.3

The potential on the surface of a hollow, insulating spherical shell of radius R is

$$V(R,\theta) = V_0 \sin^2(\theta/2),$$

which is zero at the north pole, $V_0/2$ at the equator, and V_0 at the south pole. (The insulating shell can hold any potential distribution if the right charge is fixed on it.) Find the potential inside and outside the shell, taking $V \to 0$ far away.

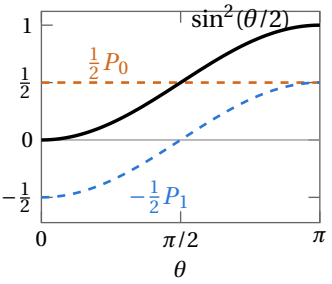
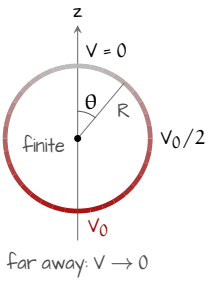


Figure 7.23 The boundary function of Example 7.3, $\sin^2(\theta/2)$ (black), is the sum of just two Legendre terms, $\frac{1}{2}P_0$ (orange) and $-\frac{1}{2}P_1(\cos\theta)$ (blue).

Solution:



We begin with a cross section of the shell, shaded by its potential, which rises from zero at the north pole to V_0 at the south pole. The sketch also marks the other two places where something is known about V: at the center, where it must be finite, and far away, where it goes to zero.

Let's be explicit about the boundary conditions:

$$\begin{aligned} &\text{BC1:} && V \text{ is finite at } r = 0, \\ &\text{BC2:} && V(R,\theta) = V_0 \sin^2(\theta/2), \\ &\text{BC3:} && V \to 0 \text{ as } r \to \infty. \end{aligned}$$

The inside and the outside are separate regions with separate solutions, so we use different letters for the coefficients in each.

Inside. BC1 eliminates the $1/r^{\ell+1}$ terms, leaving

$$V_{\text{in}}(r,\theta) = \sum_{\ell=0}^{\infty} A_\ell\, r^\ell P_\ell(\cos\theta).$$

At $r = R$, BC2 requires

$$\sum_{\ell=0}^{\infty} A_\ell R^\ell P_\ell(\cos\theta) = V_0 \sin^2(\theta/2). \qquad \text{(imposing BC2)} \quad (7.41)$$

The right-hand side is not yet written in terms of $\cos\theta$. The half-angle identity takes care of that,

$$\sin^2(\theta/2) = \tfrac{1}{2}(1-\cos\theta) = \tfrac{1}{2}P_0(\cos\theta) - \tfrac{1}{2}P_1(\cos\theta),$$

since $P_0 = 1$ and $P_1(\cos\theta) = \cos\theta$ (Figure 7.23). So (7.41) becomes

$$A_0 P_0 + A_1 R P_1 + A_2 R^2 P_2 + \cdots = \frac{V_0}{2}P_0 - \frac{V_0}{2}P_1.$$

Now match the coefficients of each Legendre polynomial. For P_0, $A_0 = V_0/2$; for P_1, $A_1 R = -V_0/2$; and for every other P_ℓ there is nothing on the right-hand side, so $A_\ell = 0$. The infinite series has collapsed to two terms,

$$V_{\text{in}}(r,\theta) = \frac{V_0}{2} - \frac{V_0}{2}\frac{r}{R}\cos\theta. \qquad (7.42)$$

Outside. BC3 eliminates the r^ℓ terms, leaving

$$V_{\text{out}}(r,\theta) = \sum_{\ell=0}^{\infty} \frac{B_\ell}{r^{\ell+1}} P_\ell(\cos\theta).$$

At $r = R$ the same boundary condition gives

$$\frac{B_0}{R}P_0 + \frac{B_1}{R^2}P_1 + \frac{B_2}{R^3}P_2 + \cdots = \frac{V_0}{2}P_0 - \frac{V_0}{2}P_1,$$

so $B_0 = V_0R/2$, $B_1 = -V_0R^2/2$, and all the other B_ℓ are zero. Thus

$$V_{\text{out}}(r,\theta) = \frac{V_0}{2}\frac{R}{r} - \frac{V_0}{2}\frac{R^2}{r^2}\cos\theta. \tag{7.43}$$

Both solutions satisfy Laplace's equation and the boundary conditions in their own regions, so by uniqueness they are correct.

Reality check: At $r = R$, both (7.42) and (7.43) reduce to $\frac{1}{2}V_0(1-\cos\theta) = V_0\sin^2(\theta/2)$, so the potential is continuous through the shell, as it must be. The potential at the center, $V_0/2$, is also the average of the potential over the shell, as the averaging property of Section 7.1.2 requires.

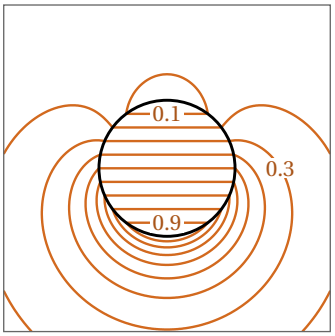


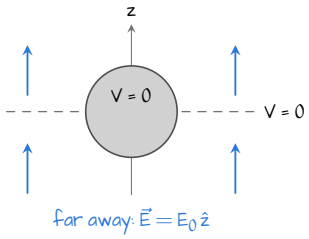
Figure 7.24 Equipotentials of Example 7.3, in a plane through the z axis, at $V/V_0 = 0.1, 0.2, \ldots, 0.9$. Inside the shell they are evenly spaced planes, the equipotentials of a uniform field; outside, the potential is that of a monopole plus a dipole.

Inside the shell, since $r\cos\theta = z$, the potential is $V_0/2 - (V_0/2R)\,z$, which is the potential of a uniform field $\mathbf{E} = (V_0/2R)\,\hat{\mathbf{z}}$, pointing from the high-potential south pole toward the low-potential north pole (Figure 7.24). Outside, the potential is a monopole term plus a dipole term. The monopole term has the form $Q/4\pi\varepsilon_0 r$, which identifies the net charge on the shell as $Q = 2\pi\varepsilon_0 RV_0$.

Example 7.4

An uncharged conducting sphere of radius R is placed in an otherwise uniform electric field $\mathbf{E} = E_0\,\hat{\mathbf{z}}$. Charge is induced on the sphere, and it distorts the field nearby. Find the potential outside the sphere and the induced surface charge density.

Solution:



The picture to draw first is the sphere in its field, with what we know about the potential marked on it. The sphere is a conductor, so it is an equipotential, and we are free to choose that potential to be zero. By symmetry, the whole plane $z = 0$ through the center of the sphere is then also at $V = 0$.

Far from the sphere, the field is the uniform field $E_0\,\hat{\mathbf{z}}$, whose potential is $-E_0 z$ plus a constant, and since the plane $z = 0$ is at zero potential, the constant is zero. So the boundary conditions are

BC1: $\quad V(R,\theta) = 0,$

BC2: $\quad V \to -E_0 z = -E_0 r\cos\theta$ as $r \to \infty$.

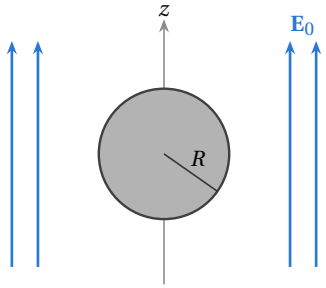


Figure 7.25 The conducting sphere of Example 7.4. The applied field is drawn only far from the sphere, where it is known.

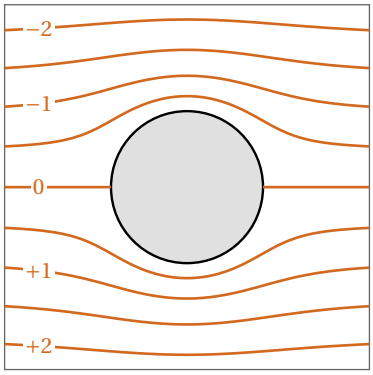


Figure 7.26 Equipotentials for Example 7.4, a conducting sphere in a uniform field that points up. The labels give V in units of $E_0 R$. The plane $z = 0$ and the sphere itself are both at $V = 0$.

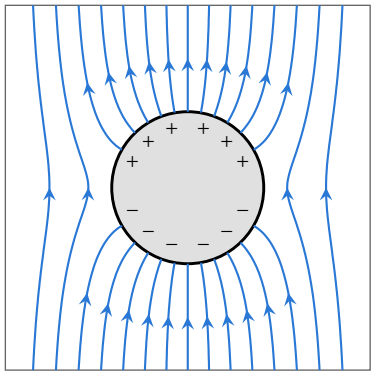


Figure 7.27 Field lines for Example 7.4. Far from the sphere they are the evenly spaced lines of the uniform field. The signs show the induced surface charge, (7.45).

Here the potential does not go to zero far away, so we cannot discard the r^ℓ terms outside. Both kinds of terms are kept,

$$V(r,\theta) = \sum_{\ell=0}^{\infty} \left(A_\ell\, r^\ell + \frac{B_\ell}{r^{\ell+1}} \right) P_\ell(\cos\theta).$$

BC2. Far from the sphere the $1/r^{\ell+1}$ terms die away and only the r^ℓ terms matter, so

$$\sum_{\ell=0}^{\infty} A_\ell\, r^\ell\, P_\ell(\cos\theta) = -E_0\, r\, P_1(\cos\theta).$$

Matching coefficients, $A_1 = -E_0$, and all the other A_ℓ are zero.

BC1. At $r = R$,

$$\sum_{\ell=0}^{\infty} \left(A_\ell R^\ell + \frac{B_\ell}{R^{\ell+1}} \right) P_\ell(\cos\theta) = 0,$$

and Colton's trick says that each coefficient must vanish separately: $A_\ell R^\ell + B_\ell / R^{\ell+1} = 0$, or

$$B_\ell = -A_\ell R^{2\ell+1}.$$

For $\ell = 1$ that gives $B_1 = E_0 R^3$, and for every other ℓ it gives $B_\ell = 0$, because the corresponding A_ℓ is zero.

This is also a good place to check the symmetry argument. Had we allowed a constant C in BC2, so that $V \to -E_0 r\cos\theta + C$, matching coefficients would have given $A_0 = C$, and BC1 would then require $B_0 = -CR$. The extra term $-CR/r$ is the potential of a net charge $-4\pi\varepsilon_0 RC$ on the sphere. The sphere is uncharged, so $C = 0$, as the symmetry argument said.

The potential outside the sphere is therefore

$$V(r,\theta) = -E_0 \left(r - \frac{R^3}{r^2} \right) \cos\theta. \tag{7.44}$$

Inside the sphere, $V = 0$.

The induced charge follows from (7.3), with the outward normal in the $\hat{\mathbf{r}}$ direction,

$$\sigma = -\varepsilon_0 \left. \frac{\partial V}{\partial r} \right|_{r=R} = \varepsilon_0 E_0 \left(1 + \frac{2R^3}{R^3} \right) \cos\theta,$$

so

$$\sigma(\theta) = 3\varepsilon_0 E_0 \cos\theta. \tag{7.45}$$

Reality check: The induced charge is positive on the upper hemisphere and negative on the lower one, as you would expect for a field pointing up, and its total is zero, as it must be for an uncharged sphere: the integral of $\cos\theta \sin\theta$ from 0 to π is zero.

The first term of (7.44) is the applied field. The second is the potential of the induced charge, and it has the form of a dipole potential, $p\cos\theta/4\pi\varepsilon_0 r^2$, with dipole moment $\mathbf{p} = 4\pi\varepsilon_0 R^3 E_0\, \hat{\mathbf{z}}$.

Figures 7.26 and 7.27 show the result. Far from the sphere the equipotentials are the evenly spaced planes of the uniform field. Near the sphere they bend so that the sphere's surface is one of them, and the field lines bend so that they meet the sphere at right angles. The field lines end on the negative charge induced on the lower hemisphere and begin again on the positive charge induced on the upper one; inside the sphere there is no field at all.

Example 7.5

An insulating spherical shell of radius R carries a fixed surface charge density

$$\sigma(\theta) = \sigma_0 \cos\theta.$$

Find the potential inside and outside the shell.

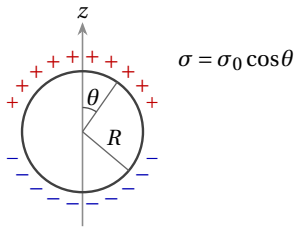
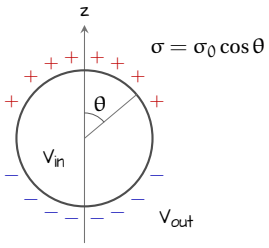


Figure 7.28 The shell of Example 7.5 with its surface charge: positive above the equator and negative below, densest at the poles.

Solution:



Here what is given is a charge density, so we sketch the shell with its charge: positive on the upper half and negative on the lower half, densest at the poles and zero at the equator. The shell divides space into two regions, inside and outside, each with its own potential.

This time there are four boundary conditions, and two of them connect the inside to the outside:

BC1:	V is finite at $r = 0$,
BC2:	$V_{\text{in}}(R,\theta) = V_{\text{out}}(R,\theta)$,
BC3:	$E^{\perp}_{\text{out}} - E^{\perp}_{\text{in}} = \sigma/\varepsilon_0$ at $r = R$,
BC4:	$V \to 0$ as $r \to \infty$.

BC2 and BC3 are the boundary conditions (7.1) at a surface charge. BC1 eliminates the $1/r^{\ell+1}$ terms inside, and BC4 eliminates the r^{ℓ} terms outside, so

$$V_{\text{in}}(r,\theta) = \sum_{\ell} A_\ell\, r^{\ell} P_\ell(\cos\theta) \qquad \text{and} \qquad V_{\text{out}}(r,\theta) = \sum_{\ell} \frac{B_\ell}{r^{\ell+1}} P_\ell(\cos\theta).$$

BC2. Setting $r = R$ in both and equating them,

$$\sum_{\ell} A_\ell R^{\ell} P_\ell(\cos\theta) = \sum_{\ell} \frac{B_\ell}{R^{\ell+1}} P_\ell(\cos\theta),$$

and matching the coefficients of each P_ℓ,

$$A_\ell R^{\ell} = \frac{B_\ell}{R^{\ell+1}}. \qquad \text{(from BC2)} \quad (7.46)$$

Hold that thought, and turn to BC3.

BC3. The perpendicular component of the field is $E^{\perp} = -\partial V/\partial r$, so BC3 says

$$-\left.\frac{\partial V_{\text{out}}}{\partial r}\right|_{r=R} + \left.\frac{\partial V_{\text{in}}}{\partial r}\right|_{r=R} = \frac{\sigma_0}{\varepsilon_0}\cos\theta = \frac{\sigma_0}{\varepsilon_0}P_1(\cos\theta).$$

Taking the derivatives term by term,

$$-\sum_\ell B_\ell\,[-(\ell+1)]\,R^{-\ell-2}P_\ell + \sum_\ell A_\ell\,\ell\,R^{\ell-1}P_\ell = \frac{\sigma_0}{\varepsilon_0}P_1,$$

or, combining the two sums,

$$\sum_\ell \left[(\ell+1)\frac{B_\ell}{R^{\ell+2}} + \ell A_\ell R^{\ell-1}\right] P_\ell(\cos\theta) = \frac{\sigma_0}{\varepsilon_0}P_1(\cos\theta).$$

Match coefficients again. For every $\ell \neq 1$ the right-hand side has no P_ℓ, so

$$(\ell+1)\frac{B_\ell}{R^{\ell+2}} + \ell A_\ell R^{\ell-1} = 0.$$

Using (7.46) to replace B_ℓ by $A_\ell R^{2\ell+1}$, this becomes $(2\ell+1)A_\ell R^{\ell-1} = 0$, so $A_\ell = 0$, and then $B_\ell = 0$ too. Only the $\ell = 1$ terms survive, and for them

$$\frac{2B_1}{R^3} + A_1 = \frac{\sigma_0}{\varepsilon_0}. \qquad \text{(from BC3)} \quad (7.47)$$

Now there are two equations for the two unknowns A_1 and B_1. Equation (7.46) with $\ell = 1$ gives $B_1 = A_1 R^3$, and substituting that into the BC3 equation gives

$$2A_1 + A_1 = \frac{\sigma_0}{\varepsilon_0}, \qquad \text{so} \qquad A_1 = \frac{\sigma_0}{3\varepsilon_0} \qquad \text{and} \qquad B_1 = \frac{\sigma_0 R^3}{3\varepsilon_0}.$$

All the other coefficients are zero, so with $P_1(\cos\theta) = \cos\theta$,

$$V_{\text{in}}(r,\theta) = \frac{\sigma_0}{3\varepsilon_0}\,r\cos\theta, \qquad V_{\text{out}}(r,\theta) = \frac{\sigma_0 R^3}{3\varepsilon_0}\,\frac{\cos\theta}{r^2}. \quad (7.48)$$

Reality check: Just inside the shell and just outside it, the parallel component of the field, $E_\theta = -(1/r)\,\partial V/\partial\theta$, is $(\sigma_0/3\varepsilon_0)\sin\theta$, so it is continuous, as (7.1) requires.

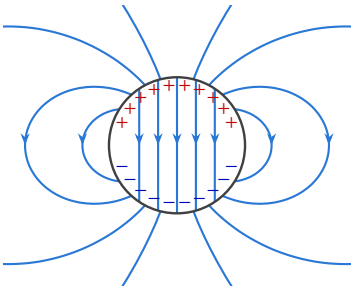


Figure 7.29 Field lines of the shell of Example 7.5. Inside, the field is uniform and points down, from the positive charge on the upper half toward the negative charge on the lower half; outside, it is a pure dipole field. The spacing of the lines is schematic.

Compare this with Example 7.4. Outside, the two answers have the same form, a pure dipole potential proportional to $\cos\theta/r^2$. Inside, they differ: here the potential is $(\sigma_0/3\varepsilon_0)\,z$, a uniform field $\mathbf{E} = -(\sigma_0/3\varepsilon_0)\,\hat{\mathbf{z}}$, while inside the conductor the field is zero (Figure 7.29). The reason for the difference becomes clear when the two problems are put side by side. The charge induced on the conducting sphere, (7.45), is itself a $\cos\theta$ distribution, with $\sigma_0 = 3\varepsilon_0 E_0$. By (7.48), that charge produces a uniform field $-E_0\,\hat{\mathbf{z}}$ inside the sphere, which exactly cancels the applied field there, and a dipole potential $E_0 R^3\cos\theta/r^2$ outside, which is exactly the second term of (7.44). The two examples are consistent with each other, and together they show how the induced charge on a conductor manages to make the field inside zero.

7.4.4 Cylindrical coordinates, no z dependence

The last case is a long cylinder, with boundaries on which nothing depends on z. With no z dependence, Laplace's equation in cylindrical coordinates is

$$\frac{1}{s}\frac{\partial}{\partial s}\left(s\frac{\partial V}{\partial s}\right)+\frac{1}{s^2}\frac{\partial^2 V}{\partial \phi^2}=0, \tag{7.49}$$

and we look for solutions of the form $V(s,\phi)=S(s)\,\Phi(\phi)$. Substituting, multiplying by s^2, and dividing by $S\Phi$ separates the equation into

$$s\frac{d}{ds}\left(s\frac{dS}{ds}\right)=k^2 S \qquad \text{and} \qquad \frac{d^2\Phi}{d\phi^2}=-k^2\Phi, \tag{7.50}$$

where the sign of the separation constant has been chosen so that the ϕ solutions are sines and cosines. Exercise P7.4 takes you through the derivation and the solution of both equations. The results are

$$\boxed{\begin{aligned} &k=0: && S\Phi=\begin{cases} C \\ \ln s \end{cases} \\ &k=1,2,\ldots,\infty: && S=\begin{cases} s^k \\ s^{-k} \end{cases}, \quad \Phi=\begin{cases} \sin k\phi \\ \cos k\phi \end{cases} \end{aligned}} \quad \text{(cylindrical, no } z\text{)} \tag{7.51}$$

Two features of (7.51) are worth understanding. First, k must be an integer. The angle ϕ and the angle $\phi+2\pi$ describe the same point in space, so $\Phi(\phi+2\pi)$ must equal $\Phi(\phi)$, and $\sin k\phi$ and $\cos k\phi$ repeat every 2π only when k is an integer. Second, the case $k=0$ is special. There the radial solutions are a constant and $\ln s$, and the ϕ equation, $d^2\Phi/d\phi^2=0$, has the solutions $\Phi=C$ and $\Phi=\phi$, the second of which is not periodic and is rejected. The $\ln s$ term is familiar: it is the potential of an infinite line charge, $-(\lambda/2\pi\varepsilon_0)\ln s$, plus a constant. Including it in a solution means including a net charge per unit length.

The general solution is

$$V(s,\phi)=A_0+B_0\ln s+\sum_{k=1}^{\infty}\left(A_k s^k+\frac{B_k}{s^k}\right)\left(C_k\cos k\phi+D_k\sin k\phi\right), \tag{7.52}$$

and its structure is close to that of the spherical solution: s^k plays the role of r^ℓ, and s^{-k} the role of $r^{-(\ell+1)}$. Inside a cylinder the s^{-k} and $\ln s$ terms are discarded, since they are infinite at $s=0$. Outside, the s^k terms are discarded, since they grow without bound. The angular part is a Fourier series in ϕ with period 2π, so the coefficients can be found with Fourier's trick, Section 6.1.3, or — when the boundary function is already a combination of sines and cosines of multiples of ϕ, or an identity makes it one — with Colton's trick, Section 6.2.4.

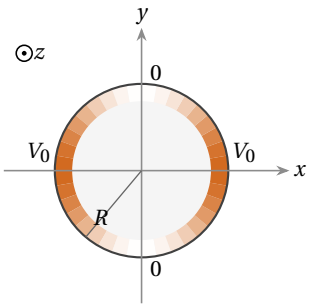


Figure 7.30 The cylindrical shell of Example 7.6 in cross section, shaded by its potential: V_0 where it crosses the x axis, zero where it crosses the y axis.

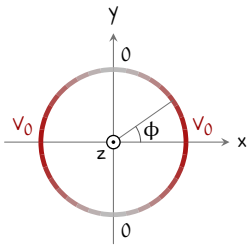
Example 7.6

A long, thin, insulating cylindrical shell of radius R carries no net charge, and the potential on its surface is

$$V(R,\phi) = V_0 \cos^2\phi.$$

Find the potential inside and outside the shell.

Solution:



Looking down the axis of the cylinder, we sketch its cross section and shade it by its potential: V_0 where it crosses the x axis, zero where it crosses the y axis. The pattern repeats twice as we go once around.

Next, write the boundary function as a combination of sines and cosines of multiples of ϕ. The double-angle identity does it,

$$V_0 \cos^2\phi = \frac{V_0}{2} + \frac{V_0}{2}\cos 2\phi,$$

so the boundary function has a $k = 0$ part and a $k = 2$ part, and nothing else.

Inside. The $\ln s$ and s^{-k} terms are infinite at $s = 0$, so

$$V_{\text{in}}(s,\phi) = A_0 + \sum_{k=1}^{\infty} s^k \left(C_k \cos k\phi + D_k \sin k\phi\right),$$

where the constants A_k have been absorbed into C_k and D_k. At $s = R$,

$$A_0 + \sum_{k=1}^{\infty} R^k \left(C_k \cos k\phi + D_k \sin k\phi\right) = \frac{V_0}{2} + \frac{V_0}{2}\cos 2\phi.$$

Matching coefficients: $A_0 = V_0/2$, $C_2 R^2 = V_0/2$, and all the other C_k and D_k are zero. So

$$V_{\text{in}}(s,\phi) = \frac{V_0}{2} + \frac{V_0}{2}\frac{s^2}{R^2}\cos 2\phi. \tag{7.53}$$

Outside. The s^k terms grow without bound, so they are discarded. The $\ln s$ term would mean a net charge per unit length on the shell, and the shell carries none, so it is discarded too. The same conclusion can be reached as a boundary condition: once the s^k terms are gone, $\ln s$ is the only term that does not stay finite as $s \to \infty$, so requiring V to remain finite far from the shell is equivalent to requiring the shell to carry no net charge. That leaves

$$V_{\text{out}}(s,\phi) = A_0' + \sum_{k=1}^{\infty} s^{-k} \left(C_k' \cos k\phi + D_k' \sin k\phi\right),$$

with primes to distinguish these coefficients from the inside ones. Matching coefficients at $s = R$ in the same way gives $A_0' = V_0/2$ and $C_2'/R^2 = V_0/2$, so

$$V_{\text{out}}(s,\phi) = \frac{V_0}{2} + \frac{V_0}{2}\frac{R^2}{s^2}\cos 2\phi. \tag{7.54}$$

Reality check: The solution is consistent with the shell being uncharged. From (7.2), the surface charge is

$$\sigma = -\varepsilon_0\left(\frac{\partial V_{\text{out}}}{\partial s} - \frac{\partial V_{\text{in}}}{\partial s}\right)_{s=R} = \frac{2\varepsilon_0 V_0}{R}\cos 2\phi,$$

which integrates to zero around the cylinder.

Notice that V_{out} does not go to zero far from the cylinder; it goes to $V_0/2$ (Figure 7.31). For an infinitely long object, the potential far away is not automatically zero, and here the value $V_0/2$ is fixed by the potential on the shell.

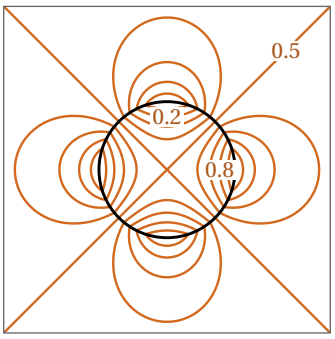


Figure 7.31 Equipotentials of Example 7.6 in a cross section, at $V/V_0 = 0.1, 0.2, \ldots, 0.9$. The $V_0/2$ equipotential is the pair of diagonals, and far from the cylinder the potential approaches $V_0/2$.

7.5 Summary

A boundary value problem asks for the potential in a region from what is known on the region's boundaries. Everything in this chapter rests on one fact: the answer is unique, so any method that produces a function satisfying the equation and the boundary conditions has produced the solution.

The equations, and the conditions at a boundary.

$\nabla^2 V = -\dfrac{\rho}{\varepsilon_0}$	Poisson's equation; Laplace's equation where $\rho = 0$
$V_1 = V_2, \quad \mathbf{E}_1^{\parallel} = \mathbf{E}_2^{\parallel}$	continuous across any surface
$E_1^{\perp} - E_2^{\perp} = \dfrac{\sigma}{\varepsilon_0}$	the jump at a surface charge
$\sigma = -\varepsilon_0\dfrac{\partial V}{\partial n}$	the charge on a conductor, once V is known

Averages and uniqueness. The value of a solution of Laplace's equation at any point is the average of its values over any surrounding sphere, so it has no local maxima or minima inside the region. That gives the *first uniqueness theorem*: specifying V on every part of the boundary determines V everywhere inside, and the same holds for Poisson's equation with a given ρ. The *second uniqueness theorem*: specifying the total charge on each conductor determines the field.

Method of images. Replace the boundaries by charges placed outside the region that reproduce the boundary conditions. For a charge q a height d above a grounded plane, the image is $-q$ a distance d below it, and the induced charge is

$$\sigma(s) = -\frac{qd}{2\pi\left(s^2 + d^2\right)^{3/2}},$$

which adds up to $-q$.

Relaxation. On a square grid, Laplace's equation says that each point is the average of its four neighbors,

$$V_{i,j} = \tfrac{1}{4}\left(V_{i+1,j} + V_{i-1,j} + V_{i,j+1} + V_{i,j-1}\right).$$

Fix the boundary points, apply this repeatedly to the others, and stop when the largest change falls below a tolerance — chosen small enough that reducing it further no longer changes the answer.

Separation of variables. The product solutions in each coordinate system, which are combined into sums and fitted to the boundary conditions with Fourier's trick or Colton's trick:
2d Cartesian, no z dependence (X and Y can be reversed):

$$X = \begin{cases} \sin kx \\ \cos kx \end{cases}, \qquad Y = \begin{cases} e^{ky} \\ e^{-ky} \end{cases} \text{ or } \begin{cases} \sinh ky \\ \cosh ky \end{cases}.$$

3d Cartesian (X, Y, and Z can be permuted), with $\kappa = \sqrt{k_x^2 + k_y^2}$:

$$X = \begin{cases} \sin k_x x \\ \cos k_x x \end{cases}, \qquad Y = \begin{cases} \sin k_y y \\ \cos k_y y \end{cases}, \qquad Z = \begin{cases} e^{\kappa z} \\ e^{-\kappa z} \end{cases} \text{ or } \begin{cases} \sinh \kappa z \\ \cosh \kappa z \end{cases}.$$

Spherical, no ϕ dependence (Q_ℓ is discarded if the region includes the z axis):

$$R = \begin{cases} r^\ell \\ 1/r^{\ell+1} \end{cases}, \qquad \Theta = \begin{cases} P_\ell(\cos\theta) \\ Q_\ell(\cos\theta) \end{cases}.$$

Cylindrical, no z dependence (k is an integer, and $\ln s$ means a net line charge):

$$k = 0:\ S\Phi = \begin{cases} C \\ \ln s \end{cases}; \qquad k = 1, 2, \ldots, \infty:\ S = \begin{cases} s^k \\ s^{-k} \end{cases}, \quad \Phi = \begin{cases} \sin k\phi \\ \cos k\phi \end{cases}.$$

Which terms to keep is decided by the boundary conditions: terms that are infinite somewhere in the region are discarded (such as $1/r^{\ell+1}$ inside a sphere, or r^ℓ outside one when $V \to 0$ far away), and the remaining coefficients are fixed by matching the conditions on the boundaries.

Exercises

Exercises for 7.1 Boundary conditions and uniqueness theorems

Warm-up questions Read Section 7.1, then answer before class

W7.1 A point charge q sits a distance D from the center of an imaginary sphere of radius $R < D$, and there is no other charge nearby. Without doing any integral, find the average of the charge's potential over the surface of the sphere. Why would your argument fail if the charge were inside the sphere?

W7.2 Two infinite, grounded, conducting plates lie in the planes $y = 0$ and $y = a$, and there is no charge between them. The potential $V = 0$ satisfies Laplace's equation between the plates and is zero on both of them. (a) Show that $V = (1\,\text{V})\, e^{\pi x/a} \sin(\pi y/a)$ does too. (b) Does this contradict the first uniqueness theorem? Explain.

W7.3 An isolated conducting sphere carries a total charge Q. A classmate says that the charge might settle unevenly, with more of it on the top half, for example, as long as the sphere ends up an equipotential. Use the second uniqueness theorem to show that it cannot.

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

Exercises for 7.2 Method of images

Warm-up questions Read Section 7.2, then answer before class

W7.4 A point charge q is held a distance d above a large grounded conducting plane. A classmate says that the work needed to pull it slowly away to infinity is $q^2/[4\pi\varepsilon_0(2d)]$, the work it would take to separate q from its image if both were real point charges. Is that right? (Think about the force on q at each height on the way out.)

W7.5 Two large grounded conducting planes meet at a right angle, like a floor and a wall, and a point charge q sits in the region between them. What image charges, and where, make the potential zero on both planes? Why are two images, one for each plane, not enough?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P7.1 Consider the image problem of Section 7.2.1: a charge $+q$ a distance d above a grounded conducting plane, which induces the surface charge density (7.7) on the plane. That surface charge experiences a force: each

patch of a conducting surface feels a force per unit area of $(\sigma^2/2\varepsilon_0)\,\hat{\mathbf{n}}$, where $\hat{\mathbf{n}}$ is the outward normal (Section 4.5.2). So the force of the charge on the plane can be calculated indirectly, through the force on the induced surface charge.

(a) Integrate the force per unit area over the whole plane to find the total force on the plane.

(b) Comment on your answer in light of the force on the charge, (7.9), which has magnitude $q^2/[4\pi\varepsilon_0(2d)^2]$ because the charge "sees" its image a distance $2d$ away.

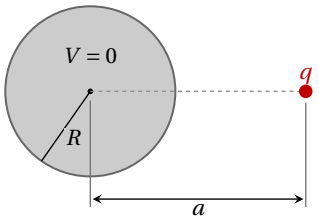


Figure 7.32 The grounded sphere and the point charge of Exercise P7.2, in cross section.

P7.2 A point charge q sits a distance a from the center of a grounded conducting sphere of radius R, with $a > R$ (Figure 7.32). Section 7.2.3 states that the image is a charge $q' = -qR/a$ a distance $b = R^2/a$ from the center, on the line joining the center to q. In this problem you derive that result.

(a) Suppose the image is a charge q' on that line, a distance $b < R$ from the center, with q' and b unknown. The line crosses the sphere at two points, one nearest to q and one farthest from it. Require the potential of q and q' together to be zero at both points, and solve the two resulting equations for q' and b.

(b) Part (a) guarantees only that $V = 0$ at two points of the sphere. Show that with these values of q' and b, $V = 0$ at *every* point of the sphere. (Hint: use the law of cosines to write the distances from a point on the sphere at polar angle θ, measured from the line, to q and to q'.)

(c) Explain why the uniqueness theorem now guarantees that the potential of q and q' is the correct potential everywhere outside the sphere.

(d) Find the force on q. Is it attractive or repulsive?

(e) Suppose q is very close to the sphere: $a = R + d$, with $d \ll R$. Show that your answer to (d) reduces to the force on a charge a distance d above a grounded plane, (7.9), and explain why that makes sense.

Exercises for 7.3 Numerical solutions: the relaxation method

Warm-up questions Read Section 7.3, then answer before class

W7.6 Relax the region of Example 7.1 by the Jacobi method, starting from $V = 0$ at every grid point not on a boundary. After the first pass, which points have changed: all of them, only those next to the 100 V strip, or those next to any boundary? After the second pass? When does a point 40 grid spacings from the strip first change, and what does that suggest about the number of passes needed?

W7.7 In the test of Section 7.3.3, the relaxation method and the series disagree most near the two corners where the boundary potential jumps from 100 V to 0. Why would a grid have the most trouble there? (Look at which terms the derivation of (7.10) neglects.)

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P7.3 Use the relaxation method of Section 7.3 to find the potential around a finite parallel-plate capacitor (Figure 7.33), in a programming language of your choice.

(a) Set up a two-dimensional array of at least 100×100 grid points to represent a region of space.

(b) Create two parallel rows of points separated by about 10% of the total grid size, to represent the plates. These should not be complete rows: each should take up only about 1/3 of the width of your domain (for example, from point 33 to point 67 of 100). Fix the points of one row at $V = +100$ V and the points of the other at $V = -100$ V.

(c) Fix the points on the outer boundary of the array at $V = 0$, to mimic the convention that $V = 0$ at infinity.

(d) Give every other point an initial value (zero will work), and then relax the array: go through the points that are not fixed, setting each equal to the average of its four neighbors, and repeat until the largest change in a pass is smaller than a tolerance you choose. Explain how you decided that your tolerance was small enough.

(e) Turn in your program, a plot of your initial potential, and a plot of the final, relaxed potential. A three-dimensional surface plot or a contour plot will work.

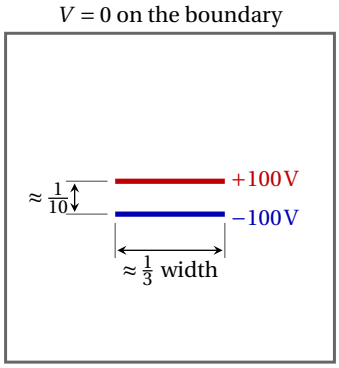


Figure 7.33 The region of Exercise P7.3, with its boundary at $V = 0$ and the two plates.

Exercises for 7.4 Separation of variables

Warm-up questions Read Section 7.4, then answer before class

W7.8 (7.4.1) Far from the strip in Example 7.1, which term of (7.26) dominates, and over what distance does the potential midway between the plates fall by a factor of e? If the plates were twice as far apart, with the strip still at 100 V, how would that distance change?

W7.9 (7.4.1) Why can't a product solution in two dimensions decay exponentially in both directions, like $e^{-x/a}e^{-y/b}$, instead of oscillating in one of them? How does that fit with the rubber-sheet picture of Section 7.1.2?

W7.10 (7.4.2) The text finds the potential at the center of the cube of Example 7.2 from six problems, one for each face. A hollow regular tetrahedron has one face held at V_0, insulated from the other three, which are grounded. (a) Without solving Laplace's equation, find the potential at its center, the point equally far from all four faces. (b) Why does this kind of argument give the potential at the center, but not at other points?

W7.11 (7.4.3) In Example 7.4, each patch of the sphere's surface is pulled outward with a force $\sigma^2/2\varepsilon_0$ per unit area (Section 4.5.2). Is there a net force on the sphere? If the sphere were a flexible conductor, such as a drop of water, how would these pulls change its shape?

W7.12 (7.4.3) A classmate says that Example 7.5 must be wrong, because Section 3.5.3 showed that the field inside a charged spherical shell is zero. What has the classmate overlooked?

W7.13 (7.4.4) Suppose the shell of Example 7.6 carried a net charge λ per unit length, while its surface potential stayed $V_0\cos^2\phi$. Would the potential inside change? Which term would you add outside, and how would the potential behave far from the shell?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P7.4 Derive the results (7.51) for separation of variables in cylindrical coordinates with no z dependence.

(a) Separate Laplace's equation (7.49) by writing $V(s,\phi) = S(s)\,\Phi(\phi)$, to obtain two ordinary differential equations, one for $S(s)$ and one for $\Phi(\phi)$, which must both equal the same constant.

(b) Solve the Φ equation first. Choose the sign of the constant to give sines and cosines, and show that the solutions are $\Phi = \cos k\phi$ and $\Phi = \sin k\phi$. Then show that the requirement $\Phi(\phi+2\pi) = \Phi(\phi)$ means

that k must be an integer. (Think about whether, for example, $\sin(1.9\cdot 30°)$ equals $\sin(1.9\cdot 390°)$, and then whether $\sin(2\cdot 30°)$ equals $\sin(2\cdot 390°)$.)

(c) Solve the S equation next, and show that its solutions are $S = s^k$ and $S = s^{-k}$.

(d) Show that when the constant is *zero*, the equations also have solutions. Specifically, show that the solutions of the Φ equation are then $\Phi = C$ and $\Phi = \phi$, but that $\Phi = \phi$ must be rejected because it does not satisfy $\Phi(\phi + 2\pi) = \Phi(\phi)$; and show that the solutions of the S equation are $S = C$ and $S = \ln s$.

P7.5 Use the results of separation of variables in cylindrical coordinates with no z dependence, (7.51), to solve the following problem. An infinite cylindrical shell has the potential $V = V_0 \sin 5\phi$ on its surface at $s = R$ (Figure 7.34). Find the potential inside and outside the shell.

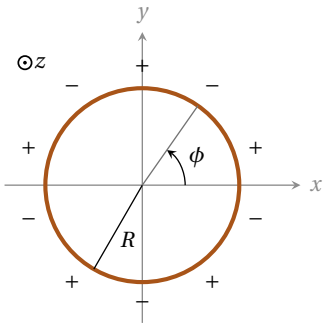


Figure 7.34 The cylindrical shell of Exercise P7.5, end on. The signs show where its potential is positive and negative.

Exercises for 7.5 Summary

P7.6 If the shell of Example 7.5 carried a uniform surface charge σ_0 instead of $\sigma_0 \cos\theta$, which terms would survive the matching at $r = R$?

(a) $\ell = 0$ only (b) $\ell = 1$ only (c) $\ell = 0$ and 1 (d) every ℓ

Explain your choice, and say what it implies for the field inside.

P7.7 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as the potential and the induced charge of a point charge near a grounded conducting sphere, the speed of the relaxation method and how over-relaxation improves it, the potential in a rectangular slot summed from many terms of its series, the field lines around a conducting sphere in a uniform field, the potential inside a cylinder whose surface potential is any function of ϕ, or the average of a solution of Laplace's equation over a sphere compared with its value at the center.

Chapter 8

Vectors, Tensors, and the Multipole Expansion

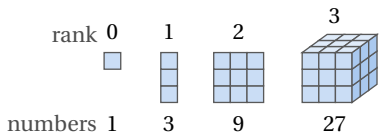


Figure 8.1 How many numbers it takes: one for a scalar (rank 0), three for a vector (rank 1), nine for a rank 2 tensor, which can be written as a 3×3 array, and twenty-seven for a rank 3 tensor, a $3 \times 3 \times 3$ cube. Each added index multiplies the count by three.

Up to now, every quantity in this book has been either a single number or a set of three numbers: charge is a single number, and the electric field at a point is three. In this chapter we meet a quantity that needs *nine* numbers to specify it, and later one that needs twenty-seven (Figure 8.1). Quantities like these are called *tensors*, and scalars and vectors turn out to be the first two members of the same family.

The physics that requires them is the **multipole expansion**: a way of describing the potential far from a localized charge distribution without doing the full integral. It is one of the most useful approximations in electromagnetism, and it shares a chapter with tensors because its third term cannot be written down without one.

8.1 What is a scalar? What is a vector?

You have been using both words since your first physics class, probably with definitions along these lines: a scalar is a quantity with magnitude, and a vector is a quantity with magnitude and direction. Those definitions are serviceable, and for most purposes they are all you need. But they are not the most precise definitions, and it is the precise ones that generalize.

The precise definition is about *what happens when you rotate your coordinate system.* The physics cannot depend on which way you chose to point your x axis. The charge in a region is the same number whichever way you point it. The electric field at a point is the same physical arrow whichever way you point it — but the three *numbers* you use to describe that arrow are not the same, because they are components along axes you just moved.

So:

- A **scalar** is a quantity described by one number, and that number is unchanged when you rotate the axes. Charge, mass, temperature, electric potential.

- A **vector** is a quantity described by three numbers, and those three numbers transform in a particular way when you rotate the axes — the way the components of a displacement arrow transform.

With words, the second one says: a vector is a thing that behaves like an arrow. Three numbers written in a column are not automatically a vector. The prices of three different stocks are three numbers, and rotating your coordinate axes does nothing to them at all. They are not a vector; they are just a list.

8.1.1 The transformation rule, written out

"Transform in a particular way" can be made exact. A rotation of the coordinate axes is described by a 3×3 **rotation matrix** R, and the new components of a vector come from the old ones by matrix multiplication:

$$\begin{pmatrix} a'_x \\ a'_y \\ a'_z \end{pmatrix} = \begin{pmatrix} R_{xx} & R_{xy} & R_{xz} \\ R_{yx} & R_{yy} & R_{yz} \\ R_{zx} & R_{zy} & R_{zz} \end{pmatrix} \begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix}, \tag{8.1}$$

where the primed components are the ones measured in the rotated system.

For the concrete case of a rotation by an angle ϕ about the z axis (Figure 8.2), the matrix is

$$R = \begin{pmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{8.2}$$

Equation (8.1) **is the definition of a vector.** A triple of numbers is a vector if and only if it transforms that way under every rotation R. Applying (8.2) to three stock prices gives numbers that mean nothing: stock prices do not transform that way, so they are not a vector.

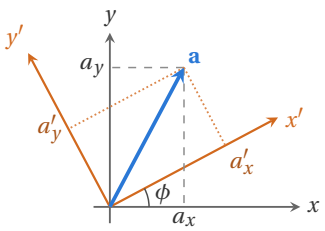


Figure 8.2 Rotating the axes by an angle ϕ about the z axis, (8.2). The vector **a** is the same arrow before and after; only its components change, from its projections a_x, a_y on the old axes to its projections a'_x, a'_y on the new ones.

8.2 Generalization—tensors

Once the definition is about transformation rather than about arrows, the generalization is straightforward. What distinguishes the members of the family is the number of indices a quantity carries, as Table 8.1 shows.

NAME	INDICES	COMPONENTS	WRITTEN
rank 0 tensor — a scalar	none	$3^0 = 1$	q
rank 1 tensor — a vector	one	$3^1 = 3$	p_i, or $\mathbf{p}$
rank 2 tensor	two	$3^2 = 9$	Q_{ij}, or a 3×3 matrix
rank 3 tensor	three	$3^3 = 27$	O_{ijk}, a $3 \times 3 \times 3$ array

Table 8.1 Tensors of rank 0 to 3.

Each index runs over the three Cartesian directions, so each one you add multiplies the number of components by three. A rank 2 tensor can be written as a matrix; a rank 3 tensor needs a cube of numbers, which cannot be printed conveniently on a page.

8.2.1 Index notation

Index notation is worth getting comfortable with now, because by the time we reach the quadrupole you will need it. Writing p_i rather than $\mathbf{p}$ means "the ith Cartesian component of $\mathbf{p}$", where i stands for any one of x, y, z. A dot product becomes a sum over a repeated index,

$$\mathbf{a}\cdot\mathbf{b} = \sum_i a_i b_i,$$

and the Kronecker delta

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

is the index-notation way of writing the identity — it is what you use when you need a term that survives only on the diagonal.

That last point deserves its own warning, which is why it is in the margin. We will shortly write expressions like $\sum_{i,j} \hat{r}_i \hat{r}_j Q_{ij}$. Here $\hat{r}_i$ is a *component* of the unit vector $\hat{\mathbf{r}}$, not a vector in its own right. If $\hat{\mathbf{r}} = \left(\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}\right)$, then $\hat{r}_x = 1/\sqrt{2}$, $\hat{r}_y = 0$, and $\hat{r}_z = 1/\sqrt{2}$ — three ordinary numbers.



Leopold Kronecker (1823–1891, German) Kronecker was born in Liegnitz (now Legnica, Poland), where Ernst Kummer taught him at the gymnasium. He studied in Berlin under Peter Gustav Lejeune Dirichlet, with semesters at Bonn and Breslau, and received his doctorate in 1845. For the next ten years he helped to run his uncle's banking business and the family estate, and in 1855 he returned to Berlin. Elected to the Berlin Academy in 1861, he lectured at the university by right of his membership, and in 1883 he succeeded Kummer as professor there. He worked in number theory, algebra, and the theory of elliptic functions. He held that every mathematical object should be built from the integers in a finite number of steps, a view summed up in the remark attributed to him, "God made the integers, all else is the work of man," and he opposed Georg Cantor's theory of infinite sets. The delta is named for him. (Wikipedia)

☞ A hat on a variable with explicit indices, as in $\hat{r}_i$, does *not* mean "a unit vector". It means the ith component *of* the unit vector $\hat{\mathbf{r}}$. So $\hat{r}_x$ is a number. This is a common source of confusion.

8.2.2 How tensors transform

Now write the rotation rule (8.1) in index notation. Matrix multiplication sums over the second index of the matrix against the index of the vector, so

$$a'_i = \sum_j R_{ij}\, a_j. \tag{8.3}$$

That is exactly (8.1) written compactly: one index on the quantity, so one factor of R and one summation.

This leads to the general definition of a **tensor**. A rank 2 tensor is a set of nine numbers that transforms with *two* factors of R and *two* summations, one for each index:

$$\boxed{T'_{ij} = \sum_k \sum_l R_{ik} R_{jl}\, T_{kl}}. \tag{8.4}$$

A rank 3 tensor takes three factors of R and three summations, and so on. A scalar takes none at all, which is just the statement that it does not change.

So the transformation rule has the same form for every rank: *a rank n tensor transforms with n factors of R.* That shared rule is what makes these objects a single family, and it is why an arbitrary 3×3 array of numbers is not automatically a rank 2 tensor — it has to transform like one.

8.3 Application: the multipole expansion of V

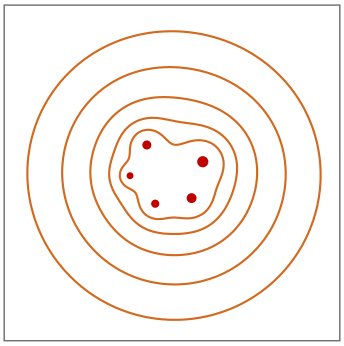


Figure 8.3 Equipotentials of five positive point charges, in their plane (dot size grows with charge). Close to the charges the equipotentials follow the lumps; farther out they become circles about the cluster, like those of a single point charge.

Suppose you have some localized distribution of charge — a molecule, a nucleus, a charged object of any shape — and you want the potential at a point far away. In principle you already know how to find it:

$$V(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\int \frac{\rho(\mathbf{r}')}{\mathscr{r}}\, d\tau'. \tag{8.5}$$

In practice this integral is often difficult, and sometimes impossible in closed form. It is also more work than the situation needs: if you are a long way away, the detailed shape of the distribution should hardly matter (Figure 8.3). A football and a beach ball look equally like points from a kilometer off.

8.3.1 What the expansion says, before any algebra

Here is the idea, stated before we do anything technical, because the algebra is much easier to follow once you know what it is going to produce.

The integral in (8.5) is going to turn into a *sum* — an infinite one, though in practice you stop after a term or two. Each term in that sum is a qualitatively different *kind* of field, and the kinds have names (Figure 8.4):

- the **monopole** field, which looks like the field of a single point charge sitting at the origin: purely radial, the same in every direction, with a potential falling off as $1/r$;
- the **dipole** field, which looks like the field of a positive and a negative charge very close together. It has an axis, it is cylindrically symmetric about that axis, and its potential falls off as $1/r^2$;
- the **quadrupole** field, with a more intricate angular pattern again, falling off as $1/r^3$;
- and so on, each more structured than the last and each falling off faster.

What the expansion tells you is that *any* localized charge distribution produces a potential that is some weighted combination of these standard shapes: some amount of monopole, plus some amount of dipole pointing in a particular direction, plus some amount of quadrupole with a particular orientation, and so on. The terms of the expansion are exactly those weightings, and the directions and orientations are built into them. Figure 8.5 shows an example.

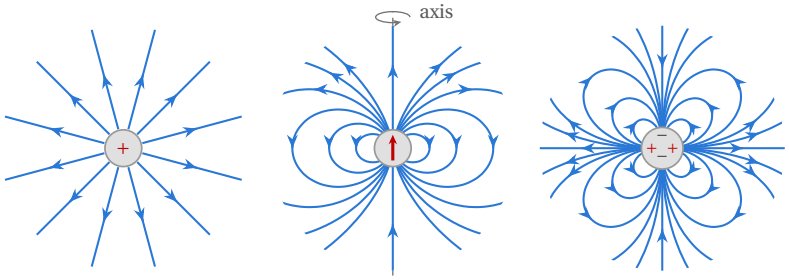


Figure 8.4 Field lines of the first three multipoles, in a plane through the source: (a) a point charge; (b) a pure dipole, whose pattern is the same in every plane that contains its axis; (c) a pure quadrupole, here four charges at the corners of a tiny square (positive on the horizontal axis, negative on the vertical axis), in the plane of the square. Each pattern has more angular structure than the one before, and its potential falls off faster with distance. The gray disks stand in for the sources.

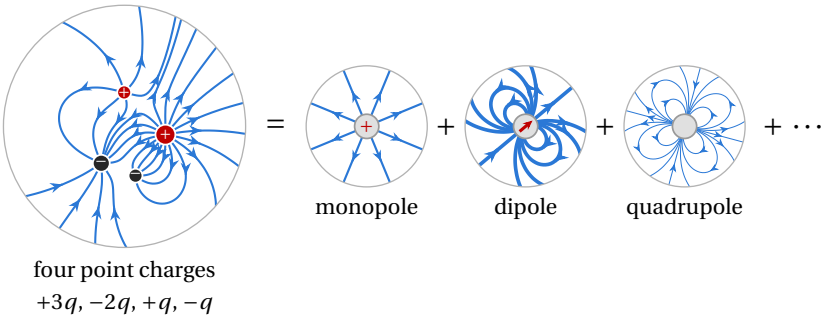


Figure 8.5 The multipole expansion as a picture. Left: field lines of four point charges, in their plane. Right: the first three terms of the expansion of their potential about the center of the circle. The net charge is $+q$, so there is a monopole term; the dipole moment points in a direction set by the charges, about 40° above the horizontal; and the quadrupole term has an orientation of its own. Line weights show the size of each term's potential on the edge of the left-hand circle, whose radius is a little more than twice the size of the arrangement: mostly dipole, some monopole, a little quadrupole. Farther out the monopole term takes over, because the others fall off faster.

Pure dipoles and physical dipoles. A caution belongs here, while the pictures are in mind.

When you draw a dipole as a single arrow, you are describing a **pure** (or *point*) dipole: two equal and opposite charges separated by a distance that has been taken to zero while the product qd is held fixed (Figure 8.6). The dipole field is exact for a pure dipole everywhere except at the origin itself.

A **physical** dipole is two actual charges a finite distance apart. For it, the dipole formula is only the leading term of an expansion that assumed you were

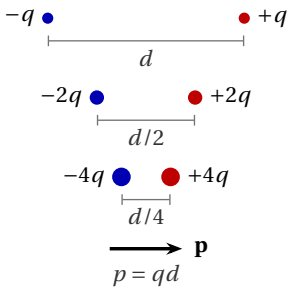


Figure 8.6 From a physical dipole to a pure one. Halving the separation while doubling the charges keeps $p = qd$ the same; in the limit $d \to 0$ the pair becomes a pure dipole, drawn as a single arrow.

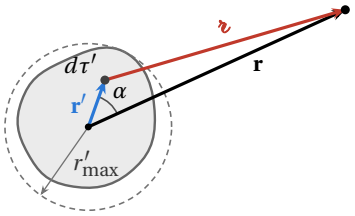


Figure 8.7 The geometry of the expansion. A bit of the charge distribution sits at $\mathbf{r}'$ and the field point at $\mathbf{r}$, and $\boldsymbol{\mathscr{r}}$ runs from one to the other. The expansion converges, and is exact, everywhere outside the sphere centered on the origin that just encloses the charge (dashed, radius $r'_{\max}$). Far away, where $r' \ll r$, a few terms are enough.

far away, so it is a good approximation at large distance and a bad one close up. The distinction matters whenever you are tempted to use the dipole formula near the object — which is exactly the situation in which it fails.

8.3.2 The mathematics of the expansion

Everything hinges on the separation distance. Let α be the angle between $\mathbf{r}$ (origin to the field point) and $\mathbf{r}'$ (origin to the charge), as in Figure 8.7. The law of cosines gives

$$\mathscr{r} = \sqrt{r^2 + r'^2 - 2rr'\cos\alpha} = r\sqrt{1 + \left(\frac{r'}{r}\right)^2 - 2\left(\frac{r'}{r}\right)\cos\alpha}.$$

Far away means $r' \ll r$, so the ratio

$$\varepsilon \equiv \frac{r'}{r}$$

is a small number, and

$$\frac{1}{\mathscr{r}} = \frac{1}{r}\left(1 + \varepsilon^2 - 2\varepsilon\cos\alpha\right)^{-1/2}. \tag{8.6}$$

Anything of the form $(1+u)^{-1/2}$ with u small is begging to be expanded in a Taylor series.



Brook Taylor (1685–1731, English) Taylor was educated at home before going to St John's College, Cambridge, where he took a degree in law in 1709. He was elected a Fellow of the Royal Society in 1712 and was its secretary from 1714 to 1718. His *Methodus incrementorum directa et inversa* (1715) contains the series now named for him, along with a treatment of the vibrating string, and in the same year he published a book on linear perspective. James Gregory, Isaac Newton, Gottfried Leibniz, and others had found versions of the series before him. Its importance went unrecognized until 1772, when Joseph-Louis Lagrange called it the main foundation of the differential calculus. (Wikipedia)

Where the Legendre polynomials come from

Write $u = \varepsilon^2 - 2\varepsilon\cos\alpha$, which is small because ε is, and use the Taylor series of $(1+u)^{-1/2}$ about $u = 0$ (the binomial series),

$$(1+u)^{-1/2} = 1 - \tfrac{1}{2}u + \tfrac{3}{8}u^2 - \tfrac{5}{16}u^3 + \cdots.$$

Now substitute $u = \varepsilon(\varepsilon - 2\cos\alpha)$ and collect terms by how small they are — that is, by their power of ε. This is the step worth being careful about: a single power of u contributes to *several* orders in ε at once, so the orders have to be gathered up across the whole series rather than read off term by term.

Order ε^0. Only the leading 1 contributes, giving simply 1.

Order ε^1. Only $-\frac{1}{2}u$ can produce a single power of ε, and only through its $-2\varepsilon\cos\alpha$ piece:

$$-\tfrac{1}{2}(-2\varepsilon\cos\alpha) = \varepsilon\cos\alpha.$$

Order ε^2. Two sources now: the ε^2 piece of $-\frac{1}{2}u$, and the square of the $-2\varepsilon\cos\alpha$ piece in $\frac{3}{8}u^2$,

$$-\tfrac{1}{2}u \longrightarrow -\tfrac{1}{2}\varepsilon^2,$$
$$\tfrac{3}{8}u^2 \longrightarrow \tfrac{3}{8}\left(4\varepsilon^2\cos^2\alpha\right) = \tfrac{3}{2}\varepsilon^2\cos^2\alpha.$$

Together they give

$$\varepsilon^2\left(\tfrac{3}{2}\cos^2\alpha - \tfrac{1}{2}\right).$$

Order ε^3. Again two sources: the cross term in $\frac{3}{8}u^2$, and the cube of the $-2\varepsilon\cos\alpha$ piece in $-\frac{5}{16}u^3$,

$$\begin{aligned}\tfrac{3}{8}u^2 &\longrightarrow \tfrac{3}{8}\left(-4\varepsilon^3\cos\alpha\right) = -\tfrac{3}{2}\varepsilon^3\cos\alpha,\\ -\tfrac{5}{16}u^3 &\longrightarrow -\tfrac{5}{16}\left(-8\varepsilon^3\cos^3\alpha\right) = \tfrac{5}{2}\varepsilon^3\cos^3\alpha.\end{aligned}$$

Together they give

$$\varepsilon^3\left(\tfrac{5}{2}\cos^3\alpha - \tfrac{3}{2}\cos\alpha\right).$$

Line the four results up and look at the coefficient of each ε^n:

$$1, \qquad \cos\alpha, \qquad \tfrac{3}{2}\cos^2\alpha - \tfrac{1}{2}, \qquad \tfrac{5}{2}\cos^3\alpha - \tfrac{3}{2}\cos\alpha.$$

Those are P_0, P_1, P_2 and P_3 — the Legendre polynomials from Section 6.2, evaluated at $\cos\alpha$. They were introduced there as the angular solutions of Laplace's equation in spherical coordinates, and they have now turned up from what looks like an entirely different direction. It is worth noticing that two unrelated-looking problems have produced the same set of functions.

The pattern continues to all orders, which is the statement

$$\frac{1}{\sqrt{1-2\varepsilon x+\varepsilon^2}} = \sum_{n=0}^{\infty}\varepsilon^n P_n(x). \tag{8.7}$$

The left-hand side of (8.7) is called the **generating function** of the Legendre polynomials, because expanding it in powers of ε generates them one after another, as we just saw (Figure 8.8).

Applying it to (8.6) with $\varepsilon = r'/r$ and $x = \cos\alpha$,

$$\frac{1}{\mathscr{r}} = \frac{1}{r}\sum_{n=0}^{\infty}\left(\frac{r'}{r}\right)^n P_n(\cos\alpha), \tag{8.8}$$

and substituting into (8.5), with everything that does not depend on the integration variable pulled out of the integral, gives

$$\boxed{V(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\sum_{n=0}^{\infty}\frac{1}{r^{n+1}}\int (r')^n P_n(\cos\alpha)\,\rho(\mathbf{r}')\,d\tau'}. \tag{8.9}$$

This is the multipole expansion. It is exact — no approximation has been made, provided the field point is farther from the origin than any of the charge — but its value is that you can stop after one or two terms.

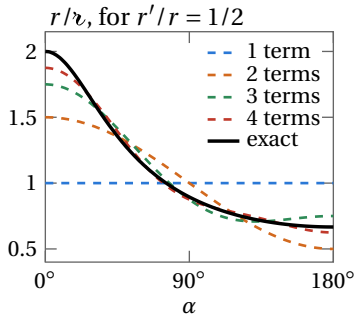


Figure 8.8 The generating function at work, for $r'/r = 1/2$: the exact $r/\mathscr{r} = \left(1+\varepsilon^2-2\varepsilon\cos\alpha\right)^{-1/2}$ (black), and the sums of the first one to four terms of (8.7). Each added term brings the sum closer, and for smaller r'/r the sums close in faster.

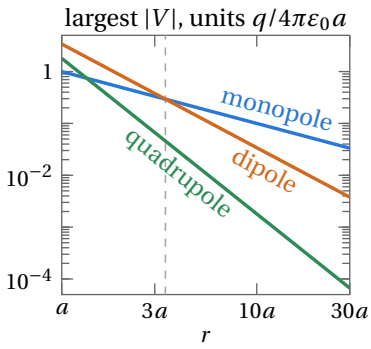


Figure 8.9 The first three terms for the four charges of Figure 8.5: the largest value of each term's potential at distance r, where a is roughly the radius of the arrangement. On log–log axes $1/r$, $1/r^2$ and $1/r^3$ are straight lines with slopes -1, -2 and -3. The dipole term is the biggest out to $r = 3.4a$ (dashed); beyond that the monopole term takes over. Animation of the exact potential and its first three terms as the field point moves out.

8.3.3 What the expansion is for

Two features of (8.9) matter, and they are easy to miss on a first reading.

First, successive terms fall off faster. The nth term carries $1/r^{n+1}$, so as you move away the terms decrease at different rates, and far enough away the first nonzero term dominates (Figure 8.9). If the monopole term is nonzero, then far enough away it does not matter what dipole or quadrupole terms you have: the monopole dominates, because everything else is smaller by factors of r'/r. If the monopole vanishes but the dipole does not, the same argument applies to the dipole: far enough away, it is all that matters.

This is what makes the expansion so useful. For most practical purposes you need exactly one term: *the first one that does not vanish.* Finding which one that is, and computing it, is usually far less work than the exact integral.

Second, each term separates cleanly into source and field point. Inside the integral is everything about the charge distribution; outside is everything about the field point. That separation is what makes the first point usable: you compute the integral once, for your object, and get a handful of numbers that characterize it. After that, finding the potential anywhere far away requires no further integration.

8.3.4 The first two terms

Take $n = 0$. Since $P_0 = 1$, the integral is just the total charge:

$$q = \int \rho(\mathbf{r}')\, d\tau', \qquad V_{\text{mono}}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\frac{q}{r}. \tag{8.10}$$

The **monopole moment** is the net charge, and the monopole potential is the potential of a point charge q at the origin.

Take $n = 1$. Now $P_1(\cos\alpha) = \cos\alpha$, and $r'\cos\alpha$ is the projection of $\mathbf{r}'$ onto the direction $\hat{\mathbf{r}}$, which is $\hat{\mathbf{r}}\cdot\mathbf{r}'$. So

$$\int r'\cos\alpha\,\rho(\mathbf{r}')\, d\tau' = \hat{\mathbf{r}}\cdot\int \mathbf{r}'\,\rho(\mathbf{r}')\, d\tau',$$

and the integral on the right depends only on the charge distribution. Give it a name:

$$\mathbf{p} = \int \mathbf{r}'\,\rho(\mathbf{r}')\, d\tau'. \tag{8.11}$$

The **dipole moment p** is a vector — three numbers.

The dipole potential can then be written in two equivalent ways: as a summation over components, which is the form that will generalize,

$$V_{\text{dip}}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0 r^2}\sum_i \hat{r}_i\, p_i, \tag{8.12}$$

or, recognizing that summation as nothing but a dot product, in the compact form

$$\boxed{V_{\text{dip}}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\frac{\hat{\mathbf{r}}\cdot\mathbf{p}}{r^2}}. \tag{8.13}$$

Both say the same thing. Equation (8.13) is the one to remember and the one to use; (8.12) is the one whose shape reappears at the quadrupole, where there is no simple dot-product shorthand and the summation is the natural form.

8.3.5 Where θ is measured from

One more warning, and it is the one that causes the most trouble. Write the dipole potential out in spherical coordinates and you get

$$V_{\text{dip}}(r,\theta) = \frac{1}{4\pi\varepsilon_0}\frac{p\cos\theta}{r^2}, \tag{8.14}$$

with the corresponding field (Figure 8.10)

$$\mathbf{E}_{\text{dip}}(r,\theta) = \frac{p}{4\pi\varepsilon_0 r^3}\left(2\cos\theta\,\hat{\mathbf{r}} + \sin\theta\,\hat{\boldsymbol{\theta}}\right). \tag{8.15}$$

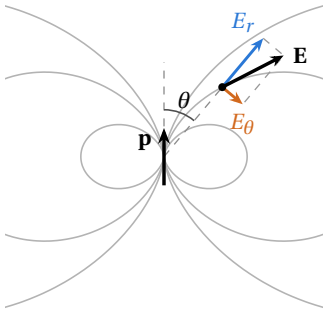
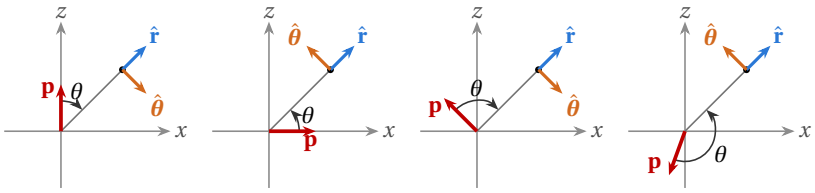


Figure 8.10 The dipole field at one point, 40° from **p**. Its components, from (8.15), are $E_r \propto 2\cos\theta$ (blue, along $\hat{\mathbf{r}}$) and $E_\theta \propto \sin\theta$ (orange, along $\hat{\boldsymbol{\theta}}$). Their sum **E** is tangent to the field line through the point (gray).

Equation (8.15) is not a new physical input. It is just $\mathbf{E} = -\nabla V$ applied to (8.14), using the gradient in spherical coordinates: the $\hat{\mathbf{r}}$ term comes from $-\partial V/\partial r$ and the $\hat{\boldsymbol{\theta}}$ term from $-(1/r)\,\partial V/\partial\theta$. Working it through is a good exercise, and is set as one below. Notice in passing that V goes as $1/r^2$ while **E** goes as $1/r^3$ — taking the derivative brings in one more power of $1/r$, and that relationship holds for every term of the expansion.

The important point is this. In (8.14) and (8.15), θ is measured *from the direction of* **p** — not from the z axis of whatever coordinate system you happened to set up. In effect you are building a new polar axis along **p**, and $\hat{\boldsymbol{\theta}}$ is defined relative to *that* axis. If **p** happens to point along $\hat{\mathbf{z}}$ the two agree, which is why the distinction is easy to miss. Figure 8.11 shows one field point and four directions of **p**.



(a) **p** along z, $\theta = 45°$ (b) **p** along x, $\theta = 45°$ (c) **p** at 135° from x; $\theta = 90°$ (d) **p** at 250° from x; $\theta = 155°$

Figure 8.11 One field point (the dot), and a dipole moment **p** pointing four different ways. $\hat{\mathbf{r}}$ (blue) points from the origin toward the field point, and is the same in every panel. θ is measured from **p**, and $\hat{\boldsymbol{\theta}}$ (orange) points the way θ increases, so both change when **p** does: in (a) and (b) θ is 45° both times, but $\hat{\boldsymbol{\theta}}$ points in opposite directions. Each panel is drawn in the plane that contains **p** and the field point, which is where $\hat{\boldsymbol{\theta}}$ always lies.

The dipole field has cylindrical symmetry about the **p** axis: rotate about **p** and nothing changes. Compare that with the monopole field, which has no angular dependence at all. Each successive multipole has a more intricate angular structure than the last — which is the same statement as "it takes more numbers to specify it".

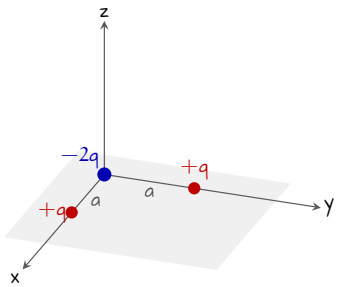
8.4 Worked example: how good is the dipole approximation?

An approximation is only useful if you know when to trust it. A good way to find out is to take a charge distribution simple enough that the exact potential can be written down, compute the dipole approximation alongside it, and see where the two agree.

Example 8.1

Three point charges sit near the origin: $+q$ at $(a,0,0)$, $+q$ at $(0,a,0)$, and $-2q$ at the origin. Find the monopole and dipole moments, then compare the dipole approximation with the exact potential along each of the three coordinate axes.

Solution:



We sketch the charges and the axes before computing anything. All three charges lie in the xy plane, and the dipole approximation is to be compared with the exact potential far out along each of the three axes.

The moments. The net charge is $q+q-2q=0$, so the monopole moment vanishes and $V_{\text{mono}}=0$ everywhere. For the dipole moment, the integral (8.11) becomes a sum over the point charges:

$$\mathbf{p} = q\,(a,0,0) + q\,(0,a,0) - 2q\,(0,0,0) = qa\left(\hat{\mathbf{x}}+\hat{\mathbf{y}}\right).$$

So the leading behavior is dipole, with $\mathbf{p}$ pointing along the diagonal between the x and y axes - a direction picked out by the charges themselves, not by our choice of coordinates (Figure 8.12).

Along the x axis. Here $\hat{\mathbf{r}}=\hat{\mathbf{x}}$, so $\hat{\mathbf{r}}\cdot\mathbf{p}=qa$ and (8.13) gives

$$V_{\text{dip}} = \frac{1}{4\pi\varepsilon_0}\frac{qa}{x^2}.$$

The exact potential is the sum of three Coulomb terms, with the distances read off the geometry:

$$V_{\text{exact}} = \frac{q}{4\pi\varepsilon_0}\left[\frac{1}{x-a}+\frac{1}{\sqrt{x^2+a^2}}-\frac{2}{x}\right].$$

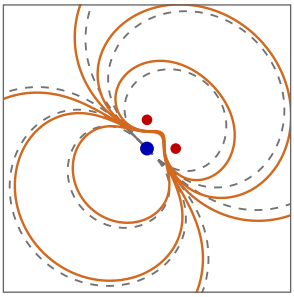


Figure 8.12 Equipotentials of the three charges in the xy plane (orange) and of their dipole approximation (dashed), at the same six values of V: three positive (upper right) and three negative. Far from the charges the two sets have nearly the same shape; close in they part company.

Expanding each term for $x \gg a$,

$$V_{\text{exact}} \approx \frac{q}{4\pi\varepsilon_0 x}\left[\left(1+\frac{a}{x}+\frac{a^2}{x^2}\right)+\left(1-\frac{a^2}{2x^2}\right)-2\right]$$
$$= \frac{q}{4\pi\varepsilon_0 x}\left(\frac{a}{x}+\frac{a^2}{2x^2}\right),$$

whose leading term, $qa/4\pi\varepsilon_0 x^2$, agrees with the dipole result, as it must.

Along the y axis. The arrangement is unchanged by swapping x and y, so this case is identical with y in place of x.

Along the z axis. Now something more interesting happens. Here $\hat{\mathbf{r}} = \hat{\mathbf{z}}$, and since $\mathbf{p}$ has no z component,

$$\hat{\mathbf{r}}\cdot\mathbf{p} = 0 \qquad \Longrightarrow \qquad V_{\text{dip}} = 0.$$

The dipole term vanishes along this direction - but the exact potential does not:

$$V_{\text{exact}} = \frac{q}{4\pi\varepsilon_0}\left[\frac{2}{\sqrt{z^2+a^2}}-\frac{2}{z}\right] \longrightarrow -\frac{1}{4\pi\varepsilon_0}\frac{qa^2}{z^3} \quad \text{for } z \gg a.$$

Along the z axis the monopole is zero and the dipole is zero, so the leading behavior is the next term - the quadrupole, falling as $1/z^3$. This is the "first nonvanishing term" rule of Section 8.3.3 in action, and it shows that which term dominates can depend on which direction you travel, not only on the charge distribution.

Reality check: The sign on the z axis makes sense. Every point of the axis is a distance z from the charge $-2q$ but a larger distance, $\sqrt{z^2+a^2}$, from each charge $+q$. Together the two charges $+q$ equal the $-2q$ in size, so the closer, negative charge dominates, and V_{exact} is negative, as found.

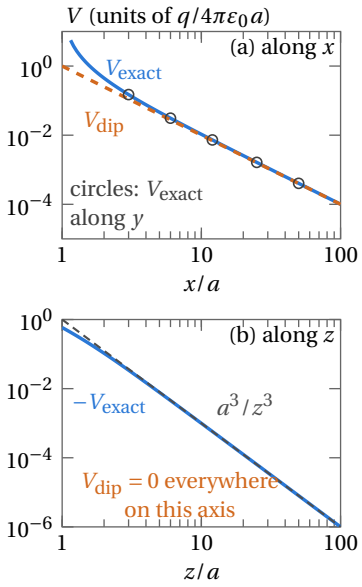


Figure 8.13 Exact and dipole potentials for the three charges of Example 8.1, on log–log axes. (a) Along the x axis the two merge as x grows, both falling as $1/x^2$: they differ by 15% at $x = 5a$ and by 1% at $x = 50a$. Near $x = a$ the exact potential climbs toward the charge sitting there. The circles are the exact potential along the y axis, which lands on the curve for the x axis, as the symmetry of the arrangement requires. (b) Along the z axis the dipole potential is zero. The exact potential is negative (the plot shows $-V_{\text{exact}}$) and falls off as $1/z^3$: the quadrupole term is the first one that survives.

Figure 8.13 shows the comparison on a graph.

8.5 The electric quadrupole moment tensor

We now come to the term that motivated this chapter.

Set $n = 2$ in (8.9). The Legendre polynomial is $P_2(\cos\alpha) = \frac{3}{2}\cos^2\alpha - \frac{1}{2}$, so

$$V_{\text{quad}}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0 r^3}\int r'^2\left(\tfrac{3}{2}\cos^2\alpha - \tfrac{1}{2}\right)\rho(\mathbf{r}')\,d\tau'. \tag{8.16}$$

We want to do to this what we did to the dipole: pull all the $\mathbf{r}$ dependence outside the integral, leaving a pure property of the source behind. Take the two terms separately.

Separating source from field point

First term. Using $r'\cos\alpha = \hat{\mathbf{r}}\cdot\mathbf{r}' = \sum_i \hat{r}_i r'_i$ as before — but now it appears squared, so we need two independent summation indices:

$$\tfrac{3}{2} r'^2 \cos^2\alpha = \tfrac{3}{2}\left(r'\cos\alpha\right)^2 = \tfrac{3}{2}\sum_{i,j} \hat{r}_i \hat{r}_j \, r'_i r'_j .$$

Second term. This one has no angular dependence at all, which makes it awkward: we need it in the same $\sum_{i,j}\hat{r}_i\hat{r}_j$ form as the first term before the two can be combined. The trick is to notice that $\hat{\mathbf{r}}\cdot\hat{\mathbf{r}} = 1$, so multiplying by it changes nothing:

$$\begin{aligned} \tfrac{1}{2} r'^2 &= \tfrac{1}{2} r'^2 \,(\hat{\mathbf{r}}\cdot\hat{\mathbf{r}}) \\ &= \tfrac{1}{2} r'^2 \sum_i \hat{r}_i \hat{r}_i \\ &= \sum_{i,j} \tfrac{1}{2} r'^2 \, \hat{r}_i \hat{r}_j \, \delta_{ij} . \end{aligned}$$

The Kronecker delta is what makes this possible: it promotes a single sum to a double sum without changing the value, because every off-diagonal term it creates is zero.

Putting them together. Both terms now carry $\hat{r}_i\hat{r}_j$, so they can be collected under one double sum, and since the sum is over indices while the integral is over primed coordinates, the two can be interchanged:

$$V_{\text{quad}}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0 r^3} \sum_{i,j} \hat{r}_i \hat{r}_j \left\{ \int \left(\tfrac{3}{2} r'_i r'_j - \tfrac{1}{2} r'^2 \delta_{ij} \right) \rho(\mathbf{r}')\, d\tau' \right\}.$$

The quantity in braces has no $\mathbf{r}$ dependence left in it. That was the goal.

So we define the **quadrupole moment**

$$\boxed{Q_{ij} = \int \left(\tfrac{3}{2} r'_i r'_j - \tfrac{1}{2} r'^2 \delta_{ij} \right) \rho(\mathbf{r}')\, d\tau'} , \tag{8.17}$$

and the potential it produces is

$$\boxed{V_{\text{quad}}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0 r^3} \sum_{i,j} \hat{r}_i \hat{r}_j \, Q_{ij}} . \tag{8.18}$$

Set (8.18) beside (8.12) and the family resemblance is clear: one index and a factor of $1/r^2$ for the dipole, two indices and a factor of $1/r^3$ for the quadrupole. This is where writing the dipole as a summation is useful. The compact form (8.13) has no simple analog here: a single dot product of $\hat{\mathbf{r}}$ with a nine-component object gives a vector, not a number, so $\hat{\mathbf{r}}$ has to appear twice — which is exactly what the double sum does.

Because i and j each run over x, y, z, the quadrupole moment is a collection of nine numbers — a rank 2 tensor, transforming under rotations according to

(8.4). In matrix form it is often written $\overleftrightarrow{\mathbf{Q}}$:

$$\overleftrightarrow{\mathbf{Q}} = \begin{pmatrix} Q_{xx} & Q_{xy} & Q_{xz} \\ Q_{xy} & Q_{yy} & Q_{yz} \\ Q_{xz} & Q_{yz} & Q_{zz} \end{pmatrix}. \tag{8.19}$$

Notice that (8.17) is unchanged if you swap i and j — both $r'_i r'_j$ and δ_{ij} are symmetric. So $Q_{ij} = Q_{ji}$, the matrix is symmetric, and only six of the nine components are independent (Figure 8.14).

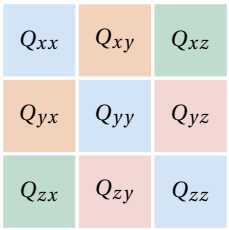


Figure 8.14 The nine components of $\overleftrightarrow{\mathbf{Q}}$. Each one off the diagonal equals its mirror image across the diagonal (same color), so symmetry leaves at most six independent components.

8.5.1 Two of the components, written out

For $i = j = x$, the delta is 1 (Figure 8.15):

$$Q_{xx} = \int \left(\tfrac{3}{2} x' x' - \tfrac{1}{2} r'^2\right) \rho(\mathbf{r}')\, d\tau'.$$

For $i = x$, $j = y$, the delta vanishes and the second term goes away entirely:

$$Q_{xy} = \int \tfrac{3}{2} x' y' \, \rho(\mathbf{r}')\, d\tau'.$$

For a set of discrete point charges q_k at positions $\mathbf{r}_k$, rather than a continuous distribution, every integral becomes a sum over the charges in the usual way,

$$Q_{ij} = \sum_k q_k \left(\tfrac{3}{2}\, r_{k,i}\, r_{k,j} - \tfrac{1}{2}\, r_k^2 \delta_{ij}\right),$$

where $r_{k,i}$ is the ith Cartesian component of $\mathbf{r}_k$.

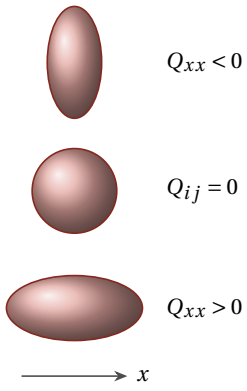


Figure 8.15 What Q_{xx} measures. Uniformly charged spheroids (positive charge) with their axes along x: squashed along x, $Q_{xx} < 0$; a sphere, every $Q_{ij} = 0$; stretched along x, $Q_{xx} > 0$. For semi-axes b along x and c across, $Q_{xx} = q\left(b^2 - c^2\right)/5$.

8.5.2 The pattern continues

The next term, $n = 3$, gives an **octopole** moment O_{ijk} — a rank 3 tensor with 27 components — and a potential built from a 27-term summation and falling off as $1/r^4$:

$$O_{ijk} = \int \Big[\tfrac{5}{2}\, r'_i r'_j r'_k - \tfrac{1}{2}\, r'^2 \left(r'_i \delta_{jk} + r'_j \delta_{ik} + r'_k \delta_{ij}\right)\Big] \rho(\mathbf{r}')\, d\tau', \tag{8.20}$$

$$V_{\text{oct}}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0 r^4} \sum_{i,j,k} \hat{r}_i \hat{r}_j \hat{r}_k \, O_{ijk}. \tag{8.21}$$

You will not need the octopole in this course. The point is the pattern, which can now be read off directly, as in Table 8.2.

NAME	SYMBOL	RANK	COMPONENTS	FALLS OFF AS
monopole	q	0	1	$1/r$
dipole	$\mathbf{p}$ or p_i	1	3	$1/r^2$
quadrupole	$\overleftrightarrow{\mathbf{Q}}$ or Q_{ij}	2	9	$1/r^3$
octopole	O_{ijk}	3	27	$1/r^4$

Table 8.2 The first four multipole moments.

Each successive moment needs a larger collection of numbers to specify it, and each successive potential falls off one power of r faster.

8.5.3 A note on conventions

Equation (8.17) is not the only definition of Q_{ij} in circulation, and you will meet the alternative. Many sources define Q_{ij} with 3 and 1 in place of $\frac{3}{2}$ and $\frac{1}{2}$, which makes it twice as large, and put an extra factor of $\frac{1}{2}$ in the expression for V_{quad} instead. The physics is identical; only the bookkeeping differs.

This book uses the convention above because it makes the Legendre polynomial visible: the $\frac{3}{2}$ and $\frac{1}{2}$ in (8.17) are exactly the coefficients of P_2, which is where they came from. As with Fourier series, none of this changes any physics as long as you are consistent within a convention, so be sure to know which convention a given reference is using.

! **The other convention.** There $Q^{\text{alt}}_{ij} = 2Q_{ij}$, with 3 and 1 in place of $\frac{3}{2}$ and $\frac{1}{2}$ in (8.17), and V_{quad} becomes

$$\frac{1}{2}\,\frac{1}{4\pi\varepsilon_0 r^3}\sum_{i,j}\hat{r}_i\hat{r}_j\,Q^{\text{alt}}_{ij}.$$

Tutorial 8.1 puts the pieces together: what the individual components of Q_{ij} do, how a charge distribution produces them, and the potential that results.

Tutorial 8.1 Electric quadrupole moment Printable version

Background

The quadrupole moment, (8.17), is a collection of nine components which can be represented by a matrix called a "rank 2 tensor" (i and j vary from 1 to 3, or equivalently from x to z). $\mathbf{r}'$ represents the locations where the charges are, r'_i and r'_j represent its Cartesian components (that is, x', y' or z'), and δ_{ij} is the Kronecker delta. If you have discrete charges instead of a continuous charge distribution, the integrals become summations over all the charges.

The potential produced by Q_{ij} at the location $\mathbf{r}$ is given by (8.18), where $\hat{\mathbf{r}}$ is a unit vector pointing toward the field point and (important!) $\hat{r}_i$ represents the ith Cartesian component of the unit vector. For example, if $\hat{\mathbf{r}} = \left(\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}\right)$, then $\hat{r}_x = \frac{1}{\sqrt{2}}$, $\hat{r}_y = 0$, and $\hat{r}_z = \frac{1}{\sqrt{2}}$.

Part 1 — How the quadrupole moment tensor produces a field

To better understand what the different Q_{ij} tensor elements mean, first consider a field point on the x axis, $\mathbf{r}_1 = (x, 0, 0)$.

(a) Draw the quadrupole at the origin and the vector $\mathbf{r}_1$ to your field point, and identify the three Cartesian components of $\hat{\mathbf{r}}_1$. These should just be three numbers, which then become $\hat{r}_x$, $\hat{r}_y$ and $\hat{r}_z$ in the formula.

(b) Write out the formula for V_{quad} with all nine terms of the summation in (8.18) explicitly written out, and cross out the terms which are zero. What is the form of V_{quad} for this field point? Which component or components of Q_{ij} does it depend on?

(c) Similarly determine what the formula for V_{quad} becomes for these other field points: $\mathbf{r}_2 = (0, y, 0)$, $\mathbf{r}_3 = (0, 0, z)$, and $\mathbf{r}_4 = \frac{1}{\sqrt{2}}(r, r, 0)$.

(d) Describe in words what you have learned about when the tensor elements Q_{xx}, Q_{yy}, Q_{zz} and Q_{xy} affect the potential at a given point. Then complete the following sentence: *The tensor element Q_{ij} affects the potential at a given point* $\mathbf{r}$ *if* _____.

Part 2 — How a charge distribution produces the quadrupole moment tensor

Consider the following charge distribution, a sphere of radius R:

$$\rho(r,\theta) = \rho_0 \frac{r^2}{R^2}\left(3\cos^2\theta - 1\right).$$

(a) Make a sketch of where the charges are positive and negative within the sphere.

As hopefully seems plausible from the symmetry of your sketch, this charge distribution has no net charge, and hence no monopole moment q. This can also be shown algebraically:

$$q = \int \rho(\mathbf{r}')\, d\tau' = \int \rho_0 \frac{r'^2}{R^2}\left(3\cos^2\theta' - 1\right)\left(r'^2 \sin\theta'\, dr'\, d\theta'\, d\phi'\right)$$
$$= 0 \qquad \text{(from the } \theta' \text{ integral)}.$$

(b) Make a physical argument from your sketch that the dipole moment should also be zero. One can show that that is actually the case by integrating $\mathbf{p} = \int \rho(\mathbf{r}')\,\mathbf{r}'\, d\tau'$, but you don't need to do that.

(c) Even though the monopole and dipole moments are zero for this charge configuration, the quadrupole moment will be nonzero. Why is that significant?

(d) Determine the nine components of the quadrupole moment tensor for this charge distribution, and write them in matrix form, $\overleftrightarrow{\mathbf{Q}}$. Mathematica (or any computer algebra system) is fair game for the integrals.

(e) What do you think would happen to $\overleftrightarrow{\mathbf{Q}}$ if the charge distribution were rotated to point in some arbitrary direction? It's fine if you don't have enough intuition yet to give a completely correct answer to this question; just give it your best guess.

Part 3 — Piecing together: V_{quad} of a specific charge distribution at a specific point

(a) Combine your results from Parts 1 and 2 to determine the quadrupole potential produced by the $\overleftrightarrow{\mathbf{Q}}$ of Part 2 at the specific point $\mathbf{r}_3 = (0,0,z)$.

(b) Check whether your answer is reasonable, using at least one of the standard physics methods: limiting cases, units, comparison to other problems, and so on.

8.6 Summary

The multipole expansion rewrites the potential of a localized charge distribution as a series whose terms fall off at different rates, so that far from the object a single

term — the first that does not vanish — is all you need.

What makes something a tensor is how it transforms. Rank n means n factors of the rotation matrix and n summations:

$a'_i = \sum_j R_{ij} a_j$ — a vector: one index, one factor of R

$T'_{ij} = \sum_k \sum_l R_{ik} R_{jl} T_{kl}$ — a rank 2 tensor: two of each

The expansion. Exact, and the starting point for everything else:

$\frac{1}{\mathscr{r}} = \frac{1}{r} \sum_n \left(\frac{r'}{r}\right)^n P_n(\cos\alpha)$ — the Legendre generating function, with α the angle between $\mathbf{r}$ and $\mathbf{r}'$

$V = \frac{1}{4\pi\varepsilon_0} \sum_n \frac{1}{r^{n+1}} \int (r')^n P_n(\cos\alpha)\rho\, d\tau'$ — each term separates source from field point

The first three moments and their potentials.

$q = \int \rho\, d\tau'$ $\qquad V_{\text{mono}} = \frac{1}{4\pi\varepsilon_0}\frac{q}{r}$

$\mathbf{p} = \int \mathbf{r}'\rho\, d\tau'$ $\qquad V_{\text{dip}} = \frac{1}{4\pi\varepsilon_0}\frac{\hat{\mathbf{r}}\cdot\mathbf{p}}{r^2} = \frac{1}{4\pi\varepsilon_0 r^2}\sum_i \hat{r}_i p_i$

$Q_{ij} = \int \left(\frac{3}{2} r'_i r'_j - \frac{1}{2} r'^2 \delta_{ij}\right)\rho\, d\tau'$ $\qquad V_{\text{quad}} = \frac{1}{4\pi\varepsilon_0 r^3}\sum_{i,j} \hat{r}_i \hat{r}_j Q_{ij}$

The dipole field, obtained from $\mathbf{E} = -\nabla V$, with θ measured from $\mathbf{p}$ and not from the z axis:

$$\mathbf{E}_{\text{dip}} = \frac{p}{4\pi\varepsilon_0 r^3}\left(2\cos\theta\,\hat{\mathbf{r}} + \sin\theta\,\hat{\boldsymbol{\theta}}\right).$$

Three things to carry forward. A tensor of rank n has 3^n components, and it is how a quantity that needs more than three numbers is described. The symbol $\hat{r}_i$ is a *component* of a unit vector, not a vector. And the leading nonvanishing moment is the one that matters far away — which is the reason the expansion is worth doing.

Exercises

Exercises for 8.1 What is a scalar? What is a vector?

Warm-up questions Read Section 8.1, then answer before class

W8.1 A classmate says that E_x, the x component of the electric field at a point, is a scalar, because it is a single number. Is the classmate right? What about the magnitude $|\mathbf{E}|$ of the field at that point? Answer each in terms of what happens when you rotate the axes.

W8.2 Figure 8.2 shows the new components a'_x and a'_y as the projections of $\mathbf{a}$ onto the new axes. Use that to explain why the first row of the matrix (8.2) is $(\cos\phi, \sin\phi, 0)$: what vector is it? What are the other two rows?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

Exercises for 8.2 Generalization—tensors

Warm-up questions Read Section 8.2, then answer before class

W8.3 Section 8.2.1 mentions sums like $\sum_{i,j} \hat{r}_i \hat{r}_j Q_{ij}$. A classmate says that such a sum must be a vector, since $\hat{\mathbf{r}}$ is a unit vector, or else a tensor, since Q_{ij} is one. What kind of quantity is it, and why?

W8.4 Equation (8.4) packs many equations into one line. How many separate equations does it stand for, and how many terms are on the right side of each? Which of its indices are summed over, and which label the equations?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P8.1 Three numbers do not make a vector. For each of the following triples, decide whether it is a vector, and say why in one sentence.

(a) The components (E_x, E_y, E_z) of the electric field at a point.

(b) The populations of the three largest cities in Utah, listed in order.

(c) The three diagonal elements (Q_{xx}, Q_{yy}, Q_{zz}) of a quadrupole moment tensor.

P8.2 Evaluate the following, where all indices run over x, y, z.

(a) $\sum_i \delta_{ii}$

(b) $\sum_{i,j} \delta_{ij}\delta_{ij}$

(c) $\sum_j \delta_{ij} a_j$, for an arbitrary vector $\mathbf{a}$.

P8.3 Take the rotation matrix (8.2) for a rotation by ϕ about the z axis.

(a) Apply it to the vector $\mathbf{a} = a\hat{\mathbf{x}}$ and verify that the result is what you would draw.

(b) Use (8.4) to work out T'_{xx} for a general rank 2 tensor under the same rotation, in terms of the unprimed components and ϕ.

(c) Show that a tensor proportional to δ_{ij} is unchanged by this rotation. (Such a tensor is called *isotropic*: it is the same in every orientation.)

Exercises for 8.3 Application: the multipole expansion of V

Warm-up questions Read Section 8.3, then answer before class

W8.5 Along the line $\alpha = 0$, where the field point lies straight out beyond the charge at $\mathbf{r}'$, the distance is $\mathscr{r} = r - r'$. Expand $1/\mathscr{r}$ in powers of r'/r (a geometric series) and compare it term by term with (8.8). What does the comparison tell you about the Legendre polynomials? Why does the series fail if $r < r'$?

W8.6 A single point charge q sits at $(0,0,d)$. Find its monopole moment and its dipole moment about the origin. Then sketch three fields: the field of the charge itself, the field of the monopole term alone, and the field of the dipole term alone. Why can the monopole term alone not reproduce the field of the charge?

W8.7 In panels (a) and (b) of Figure 8.11, the field point is the same and $\theta = 45°$ in both. Is the dipole potential at the field point the same in the two panels? Is the field? Explain.

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P8.4 Continue the expansion in Section 8.3.2 one order further, to ε^4, and confirm that the coefficient is $P_4(\cos\alpha) = \frac{35}{8}\cos^4\alpha - \frac{15}{4}\cos^2\alpha + \frac{3}{8}$. You will need the next term of the binomial series; keep careful track of which powers of u contribute.

P8.5 A charge distribution has $q = 0$ but $\mathbf{p} \neq 0$. A second distribution of the same overall size has $q \neq 0$. Both are observed from a distance r large compared with the objects themselves.

(a) By what factor does the ratio of their potentials change if you double r?

(b) Roughly how far away must you be before the second object's potential exceeds the first's by a factor of ten? Express your answer in terms of the object size r' and whatever else you need.

P8.6 Starting from the dipole potential (8.14), take $\mathbf{E} = -\nabla V$ using the gradient in spherical coordinates, and show that you obtain (8.15). Explain in one sentence why there is no $\hat{\boldsymbol{\phi}}$ component.

P8.7 A pure dipole $\mathbf{p} = p\hat{\mathbf{x}}$ sits at the origin (Figure 8.16). Find $\mathbf{E}$ at the point $(0,0,a)$ on the z axis, being careful about where θ is measured from. Then state the answer again in terms of $\hat{\mathbf{r}}$ and $\hat{\boldsymbol{\theta}}$ referred to the dipole's own axis, and confirm the two agree.

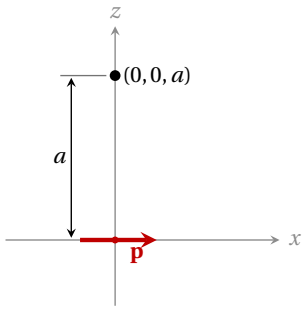


Figure 8.16 The dipole and the field point of Exercise P8.7, in the xz plane.

P8.8 For the three-charge arrangement of Example 8.1, expand V_{exact} along the y axis for $y \gg a$ and confirm that it agrees with the dipole approximation. Then explain, without further calculation, why the answer had to match the x-axis result.

Exercises for 8.4 Worked example: how good is the dipole approximation?

Warm-up questions — Read Section 8.4, then answer before class

W8.8 For the three charges of Example 8.1, along which line through the origin in the xy plane does the dipole term vanish? Far out along that line, how would you expect the potential to fall off, and why?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

Exercises for 8.5 The electric quadrupole moment tensor

Warm-up questions — Read Section 8.5, then answer before class

W8.9 Figure 8.15 shows that stretching a positive charge distribution along x makes Q_{xx} positive. Use (8.17) to find the sign of Q_{xx} for a thin rod with a uniform positive charge, centered on the origin and lying (a) along the x axis, (b) along the y axis. Do your answers fit the pattern of the figure?

W8.10 Two tables give the quadrupole moment Q_{zz} of the same molecule, and one value is exactly twice the other. Can both be right? Which would you use in (8.18), and how could you tell?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

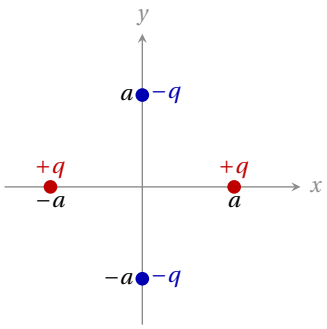


Figure 8.17 The four charges of Exercise P8.10.

P8.9 For the quadrupole moment defined in (8.17), show that $\sum_i Q_{ii} = 0$ — that is, the tensor is traceless. Combined with the symmetry argument already given in the text, how many of the nine components are genuinely independent?

P8.10 Four point charges lie in the xy plane (Figure 8.17): $+q$ at $(\pm a, 0, 0)$ and $-q$ at $(0, \pm a, 0)$. Far away, their field has the four-lobed pattern of Figure 8.4(c).

(a) Show that the monopole and dipole moments are both zero.

(b) Find all nine components of Q_{ij}, and write $\overleftrightarrow{\mathbf{Q}}$ in matrix form. Check that it is symmetric and that its diagonal elements add to zero.

(c) Use (8.18) to find V_{quad} at the field points $(x, 0, 0)$, $(0, 0, z)$ and $\frac{1}{\sqrt{2}}(r, r, 0)$.

(d) Write the exact potential on the x axis for $x > a$, expand it for $x \gg a$, and confirm that its leading term agrees with your answer to part (c).

(e) On the z axis, and on the line $x = y$ in the xy plane, the exact potential is zero — not just its quadrupole term. Explain why, from the positions of the charges.

P8.11 Many charge distributions are symmetric about an axis. Take it to be the z axis. Then $\overleftrightarrow{\mathbf{Q}}$ is diagonal with $Q_{xx} = Q_{yy}$, and since the diagonal elements add to zero (Exercise P8.9),

$$\overleftrightarrow{\mathbf{Q}} = Q_0 \begin{pmatrix} -\frac{1}{2} & 0 & 0 \\ 0 & -\frac{1}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

for some constant Q_0.

(a) Show that at a field point in the direction given by the spherical angles θ and ϕ, (8.18) gives $V_{\text{quad}} = Q_0 P_2(\cos\theta)/4\pi\varepsilon_0 r^3$ — the Legendre polynomial of (8.9) reappears.

(b) The sphere of Tutorial 8.1 has this symmetry. Use its $\overleftrightarrow{\mathbf{Q}}$ to write its quadrupole potential at any distant point, and check that the result agrees with Part 3 of the tutorial on the z axis.

(c) At what angles θ is that potential zero?

Exercises for 8.6 Summary

P8.12 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as the rotation of a vector and of a tensor by a rotation matrix, the terms of the multipole expansion of an irregular cluster of charges, the error of the dipole approximation as a function of distance, the quadrupole moment tensor of a molecule such as water from the charges of its atoms, the principal axes of a quadrupole tensor found by diagonalizing it, or the equipotentials of a charge distribution at distances where different multipoles dominate.

Chapter 9

Polarization and Dielectrics

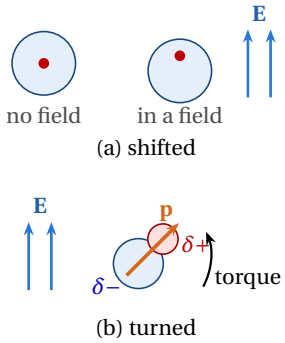


Figure 9.1 The two ways a field polarizes an insulator. (a) The field pushes an atom's nucleus along **E** and its electron cloud the other way, so the atom becomes a small dipole (the shift is exaggerated). (b) The field exerts a torque on a molecule that is already lopsided, a polar molecule, and turns its dipole moment **p** toward **E**.

Chapters 3–8 dealt with charges in vacuum and with conductors, in which charge moves freely until the field inside is zero. Most materials are neither. They are insulators, in which every electron stays bound to its own atom or molecule. A field cannot move those charges far, but it can shift them slightly, the nuclei one way and the electrons the other, and it can turn molecules that are already lopsided (Figure 9.1). Either way, each atom or molecule becomes a small dipole, and the material as a whole becomes **polarized**. A polarized material produces a field of its own, which adds to the field that polarized it. In this context insulators are called **dielectrics**;[1] the two words are used interchangeably in this chapter. (The line between conductors and insulators is not as sharp as this makes it sound: every real material is partly both, as Section 22.4 explains.)

The chapter begins with dipoles in external fields, and with the ways atoms and molecules come to have dipole moments. It then describes a polarized material by its dipole moment per unit volume, shows that the field of a polarized object is the field of a particular distribution of charge, called bound charge, and develops two ways to find that field. The second way uses a new field, the displacement **D**, whose divergence depends only on the charge that we put in place ourselves. The chapter ends with the boundary conditions at the surface of a dielectric, numerical methods, and the energy stored in a dielectric and the forces on it.

[1] Faraday introduced the word, from the Greek *dia-*, "through." In a paper written in 1837 he described electric induction as "an action of the contiguous particles of the insulating medium or di-electric." In the same paper he reported that a capacitor made of a metal ball inside a hollow metal sphere holds more charge when part of the gap is filled with shellac, glass, or sulfur than when it is filled with air. He called the property of the material responsible for this its "specific inductive capacity," today's dielectric constant (Section 9.5.1), and from his measurements he estimated it to be roughly 2 for shellac, above 1.76 for flint glass, and about 2.24 for sulfur. Cavendish had found the same effect in the 1770s but never published it; his notes became widely known only when Maxwell edited them in 1879.

9.1 Dipoles again

Chapter 8 defined the dipole moment of a charge distribution, (8.11),

$$\mathbf{p} = \int \mathbf{r}'\,\rho(\mathbf{r}')\,d\tau',$$

which for two charges $\pm q$ separated by the vector $\mathbf{d}$, pointing from the negative charge to the positive one, is $\mathbf{p} = q\mathbf{d}$. It also found the potential and field of a pure dipole, (8.13) and (8.15), with θ measured from the direction of $\mathbf{p}$. Those are the fields a dipole *produces*. This section is about the other half of the story: what a field made by other charges does to a dipole placed in it. Throughout, $\mathbf{E}$ means that external field. A dipole's own field exerts no net force or torque on the dipole itself, just as a point charge exerts no force on itself.

9.1.1 Torque on a dipole

Put a dipole, $+q$ and $-q$ held a fixed distance apart, in a uniform field $\mathbf{E}$. The positive charge feels the force $q\mathbf{E}$ and the negative charge feels $-q\mathbf{E}$, so the net force is zero. The two forces do not act along the same line, however, so they exert a torque. About the center of the dipole, the positive charge is at $\mathbf{d}/2$ and the negative charge at $-\mathbf{d}/2$, and the total torque is

$$\boldsymbol{\tau} = \frac{\mathbf{d}}{2} \times (q\mathbf{E}) + \left(-\frac{\mathbf{d}}{2}\right) \times (-q\mathbf{E}) = q\,\mathbf{d} \times \mathbf{E}.$$

Since $q\mathbf{d}$ is the dipole moment,

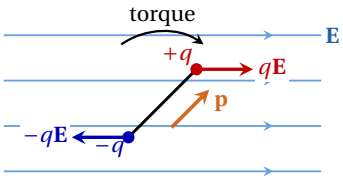
$$\boldsymbol{\tau} = \mathbf{p} \times \mathbf{E}. \tag{9.1}$$

(The τ here is a torque, not the volume element $d\tau$.) The same result holds for any charge distribution in a uniform field, not just two charges: the torque $\sum \mathbf{r}_i \times q_i\mathbf{E}$ is $\left(\sum q_i \mathbf{r}_i\right) \times \mathbf{E} = \mathbf{p} \times \mathbf{E}$ (Exercise P9.1). In a uniform field, only the dipole moment matters.

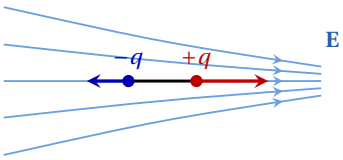
The torque is perpendicular to both $\mathbf{p}$ and $\mathbf{E}$, and by the right-hand rule it turns $\mathbf{p}$ toward the direction of $\mathbf{E}$ (Figure 9.2a). Its magnitude is $pE\sin\theta$, where θ is the angle between $\mathbf{p}$ and $\mathbf{E}$. If nothing resisted the motion, a dipole released at an angle to the field would swing past the direction of $\mathbf{E}$ and back again, oscillating about it the way a pendulum oscillates about the vertical. In a material, collisions with neighboring molecules damp the motion, and the dipole tends to end up pointing along the field.

9.1.2 Energy of a dipole in a field

The potential energy of a dipole in a field is the work needed to turn it, slowly, from a reference orientation to its final orientation. To turn it slowly we must



(a) uniform field



(b) nonuniform field

Figure 9.2 (a) A dipole in a uniform field. The forces on the two charges are equal and opposite, so the net force is zero, but they exert a torque that turns $\mathbf{p}$ toward $\mathbf{E}$; here the torque points into the page. (b) In a nonuniform field the forces do not cancel. Here the field is stronger at the positive charge, and the dipole is pulled toward the stronger field.

apply a torque equal and opposite to $\mathbf{p} \times \mathbf{E}$, of magnitude $pE\sin\theta$, in the direction of increasing θ. The work is

$$U(\theta) = \int_{\theta_{\text{ref}}}^{\theta} pE\sin\theta'\, d\theta'.$$

Any reference angle will do, since only differences in energy matter, but one choice gives a particularly simple formula: $\theta_{\text{ref}} = \pi/2$, with the dipole perpendicular to the field. Then $U = -pE\cos\theta$, or

$$U = -\mathbf{p}\cdot\mathbf{E}. \tag{9.2}$$

The energy is lowest, $-pE$, when the dipole points along the field, and highest, $+pE$, when it points the other way. (With the reference at $\theta = 0$ instead, the energy is $pE(1-\cos\theta)$, which differs from (9.2) only by the constant pE.)

The choice of reference is more natural than it looks. By (5.12) applied to each charge, the potential energy of the two charges in the potential V of the other charges is $U = qV(\mathbf{r}_+) - qV(\mathbf{r}_-)$. For a small separation, $V(\mathbf{r}_+) - V(\mathbf{r}_-) = \mathbf{d}\cdot\nabla V$, so $U = q\mathbf{d}\cdot\nabla V = -\mathbf{p}\cdot\mathbf{E}$: the same result, with the same zero. The reference point of V drops out, since the dipole is neutral, and the zero now has a simple meaning: U is the work needed to bring the dipole, without turning it, from a place where there is no field.

Equation (9.2) is the energy of a dipole of fixed size placed in a field made by other charges, the dipole version of qV. It is not the energy stored in the dipole's own field, which is the subject of Section 5.9.

9.1.3 Force on a dipole

In a uniform field the net force on a dipole is zero. In a nonuniform field it is not, because the two charges sit at points where the field is different (Figure 9.2b). With the negative charge at $\mathbf{r}$ and the positive charge at $\mathbf{r}+\mathbf{d}$, the net force is

$$\mathbf{F} = q\mathbf{E}(\mathbf{r}+\mathbf{d}) - q\mathbf{E}(\mathbf{r}).$$

For a small separation, the change in each component of the field across $\mathbf{d}$ is given by the gradient, (1.19): $E_x(\mathbf{r}+\mathbf{d}) - E_x(\mathbf{r}) = \mathbf{d}\cdot\nabla E_x$, and likewise for E_y and E_z. So $\mathbf{F} = q(\mathbf{d}\cdot\nabla)\mathbf{E}$, or

$$\mathbf{F} = (\mathbf{p}\cdot\nabla)\,\mathbf{E} = p_x\frac{\partial\mathbf{E}}{\partial x} + p_y\frac{\partial\mathbf{E}}{\partial y} + p_z\frac{\partial\mathbf{E}}{\partial z}. \tag{9.3}$$

The operator $\mathbf{p}\cdot\nabla$ is shorthand for the combination of derivatives on the right; it is not a dot product of two vectors. In words, the force is p times the rate at which $\mathbf{E}$ changes as you move in the direction of $\mathbf{p}$.

☞ *Caution*: In $\nabla(\mathbf{p}\cdot\mathbf{E})$, $\mathbf{p}$ is the dipole moment of a single dipole, not a field. It does not depend on position and is not differentiated.

The force can also be found from the energy, as $\mathbf{F} = -\nabla U$:

$$\mathbf{F} = \nabla\big(\mathbf{p}\cdot\mathbf{E}\big). \tag{9.4}$$

Here **p** is the dipole moment of one particular dipole, a constant vector, and only **E** is differentiated. The two expressions agree because, by the product rule for the gradient of a dot product (Section 1.7), $\nabla(\mathbf{p}\cdot\mathbf{E}) = (\mathbf{p}\cdot\nabla)\mathbf{E} + \mathbf{p}\times(\nabla\times\mathbf{E})$ for constant **p**, and the curl of an electrostatic field is zero, (5.2). In electrostatics either form may be used. (For a magnetic dipole, Section 13.1.3, the two forms are not equivalent in general, and it is the gradient form that carries over.) Equation (9.4) also gives a quick way to see which way the force points: a dipole that is lined up with the field has $U = -pE$, which is lowest where the field is strongest, so it is pulled toward the region of stronger field.

Equations (9.1)–(9.4) assume that the dipole moment is fixed. For an induced dipole, whose moment is itself proportional to the field (below), the force is still (9.3) with **p** evaluated at the dipole's location, but the energy is $-\frac{1}{2}\alpha E^2$ rather than $-\mathbf{p}\cdot\mathbf{E}$ (Exercise P9.3).



Peter Debye (1884–1966, Dutch-American) Born in Maastricht, Debye trained as an electrical engineer at Aachen, where he became an assistant to the physicist Arnold Sommerfeld, and he followed Sommerfeld to Munich for a doctorate (1908). In 1912 he extended Einstein's theory of the heat capacity of solids to low temperatures, and in the same year he proposed that some molecules carry permanent dipole moments, which an applied field lines up only partly against the randomizing effect of thermal motion. That explained why such substances have large dielectric constants that fall as the temperature rises. With Paul Scherrer he developed X-ray diffraction from powders (1916), and with Erich Hückel a theory of solutions of electrolytes (1923). After professorships in Zurich, Utrecht, Göttingen, and Leipzig, he directed the Kaiser Wilhelm Institute for Physics in Berlin, and in 1940 he moved to Cornell University, where he spent the rest of his career.[2] He received the 1936 Nobel Prize in Chemistry for his work on dipole moments and on the diffraction of X-rays and electrons in gases. (Wikipedia)

9.1.4 Polar molecules

Some molecules have a dipole moment even when there is no field. In a water molecule the oxygen atom pulls the shared electrons toward itself, away from the two hydrogen atoms, so the oxygen end is negative and the hydrogen end positive (Figure 9.3). The dipole moment of an isolated water molecule is 6.19×10^{-30} C m.[3] Such molecules are called **polar molecules**.

A field exerts a torque on each polar molecule and tends to line it up, while thermal motion keeps knocking the molecules into random orientations. At room temperature the thermal motion wins by a wide margin. In a field of 10^6 V/m, which is a third of the field at which air breaks down, the energy difference between a water molecule lined up with the field and one pointing the other way is $2pE = 1.2\times10^{-23}$ J, while the typical thermal energy kT at 20°C is 4.0×10^{-21} J, more than 300 times larger. So the average alignment is very slight. But a liter of

[2] One aspect of Debye's history is a bit controversial, namely his dealings with Nazi Germany. As one study of the case put it, Debye "had to face several political dilemmas." In December 1938, as chairman of the German Physical Society, he signed a letter telling the society's German Jewish members that their membership could no longer be maintained and asked them to resign. For this he posthumously came under a large amount of criticism when the letter drew wide attention in 2006. In January 1940, Debye, having refused to become a German citizen, left Germany for the United States, where he became a professor at Cornell University, chaired the chemistry department for 10 years, and in 1946 became an American citizen. In 2006 following the publication of *Einstein in Nederland* by Sybe Izaak Rispens, Utrecht University took his name off its Debye Institute; but Cornell University found no evidence that he had been a Nazi sympathizer or collaborator, and a study by the Netherlands Institute for War Documentation concluded that he was neither a Nazi nor an anti-Semite. His defenders also pointed to how in July 1938, at risk to himself, he had helped Jewish physicist Lise Meitner escape from Germany after the German Anschluss stripped her of the protection of her Austrian citizenship. In 2008 Utrecht restored the name, accepting an independent committee's conclusion that "since no bad faith on Debye's part has been demonstrated, his good faith must be assumed." The quotations are from "Peter Debye: Nazi Collaborator or Secret Opponent?" J. Reiding, Ambix 57, 275–300 (2010). https://doi.org/10.1179/174582310X12849808295706.

[3] Chemists measure dipole moments in debyes, after Peter Debye: $1\,\mathrm{D} = 10^{-18}$ in Gaussian units (statcoulomb centimeters), or 3.336×10^{-30} C m. The moment of a water molecule is 1.855 D.

water contains about 3×10^{25} molecules, and even a slight average alignment of that many dipoles produces a large effect, as Section 9.5.1 shows.

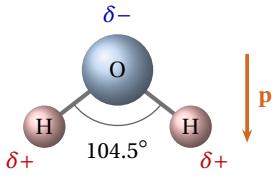


Figure 9.3 A water molecule. The oxygen atom pulls the shared electrons toward itself, so the oxygen end is negative ($\delta-$) and the hydrogen end positive ($\delta+$), and the dipole moment **p** points from the oxygen toward the hydrogens. The two bonds make an angle of about 104.5°.

9.1.5 Induced dipoles: polarizability

An isolated atom has no dipole moment: on average, its electrons are distributed symmetrically around the nucleus. In a field, the nucleus is pushed along **E** and the electrons are pushed the other way, and the atom acquires an **induced dipole moment**. For fields much weaker than the fields inside the atom, which is the case for any static field that can be applied in a laboratory, the induced dipole moment is proportional to the field:

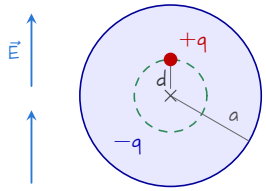
$$\mathbf{p} = \alpha \mathbf{E}. \tag{9.5}$$

The constant α is the **atomic polarizability**. Its SI unit is $\mathrm{C\,m^2/V}$, and $\alpha/4\pi\varepsilon_0$ has the units of volume. The following example shows why.

Example 9.1

A simple model of an atom is a point nucleus of charge $+q$ at the center of a uniformly charged ball of electron charge $-q$, with radius a. Assume that in a field **E** the ball keeps its shape and the nucleus moves until the forces on it balance. Find the polarizability.

Solution:



We draw the model atom in a field pointing up, with a sphere of radius d (dashed) about the center of the cloud. The nucleus moves a distance d along the field. There it feels the applied field and the field of the electron cloud, which by Gauss's law (Chapter 4) is that of the charge enclosed within radius d: a fraction d^3/a^3 of the cloud.

That field is $(1/4\pi\varepsilon_0)\,qd/a^3$, pointing back toward the center of the cloud. The forces balance when

$$E = \frac{1}{4\pi\varepsilon_0}\frac{qd}{a^3},$$

so the dipole moment is $p = qd = 4\pi\varepsilon_0 a^3 E$. (Why is the dipole moment simply qd, although the electron charge is spread through a ball? Answer in footnote.[4]) Since $p = \alpha E$,

$$\alpha = 4\pi\varepsilon_0\, a^3.$$

[4]The cloud is symmetric about its center, so in (8.11) each pair of its points on opposite sides of the center contributes the pair's charge times the position of the center, and the whole cloud counts as a point charge $-q$ at its center. The atom therefore has the dipole moment of two point charges $\pm q$ a distance d apart.

Reality check: In this model $\alpha/4\pi\varepsilon_0$ is the cube of the atom's radius, a volume, as the units of α require. For hydrogen, with a equal to the Bohr radius $a_0 = 5.29\times10^{-11}\,\mathrm{m}$, the model gives $\alpha/4\pi\varepsilon_0 = a_0^3 = 0.148\times10^{-30}\,\mathrm{m}^3$. The exact quantum-mechanical value is $4.5\,a_0^3 = 0.667\times10^{-30}\,\mathrm{m}^3$, so the model gets the order of magnitude right.

Table 9.1 lists measured or calculated polarizabilities of some atoms, as $\alpha/4\pi\varepsilon_0$ in units of $10^{-30}\,\mathrm{m}^3$.[5] As the simple model suggests, $\alpha/4\pi\varepsilon_0$ is of the order of the cube of the atom's radius, so larger atoms are easier to polarize. The noble gases, whose electrons are tightly bound in closed shells, have small polarizabilities. The alkali atoms have one loosely bound outer electron, far from the nucleus, which a field can shift easily; their polarizabilities are 15 to 120 times those of the noble gases that precede them in the periodic table.

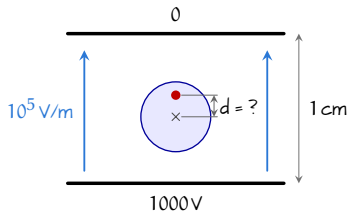
Atom	$\alpha/4\pi\varepsilon_0$
H	0.667
He	0.205
Ne	0.396
Ar	1.64
Kr	2.48
Xe	4.04
Li	24.3
Na	24.1
Cs	59.4

Table 9.1 Polarizabilities of some atoms, as $\alpha/4\pi\varepsilon_0$ in units of $10^{-30}\,\mathrm{m}^3$: hydrogen, the noble gases, and three alkali atoms (sources in the footnote).

Example 9.2

A hydrogen atom is placed in a field of $10^5\,\mathrm{V/m}$, the field produced by 1000 V across a gap of 1 cm. How far apart are the centers of the positive and negative charge?

Solution:



Even a number problem starts with a sketch. Here it shows the atom between the plates, its stretch greatly exaggerated. The field pulls the nucleus one way and the electron cloud the other, and we want the distance d between their centers.

With $\alpha = 4\pi\varepsilon_0\,(0.667\times10^{-30}\,\mathrm{m}^3) = 7.42\times10^{-41}\,\mathrm{C\,m^2/V}$, the induced dipole moment is $p = \alpha E = 7.42\times10^{-36}\,\mathrm{C\,m}$. Writing $p = ed$,

$$\begin{aligned} d &= \frac{\alpha E}{e} \\ &= \frac{7.42\times10^{-36}\,\mathrm{C\,m}}{1.602\times10^{-19}\,\mathrm{C}} \\ &= 4.6\times10^{-17}\,\mathrm{m}, \end{aligned}$$

about a millionth of the Bohr radius. The shift is tiny, which is why the response is so accurately linear. But a solid contains some 10^{23} atoms per cubic centimeter, and their combined effect is large.

Reality check: The units work out: $(\mathrm{C\,m^2/V})(\mathrm{V/m})/\mathrm{C} = \mathrm{m}$.

[5] Values are from "Theory and applications of atomic and ionic polarizabilities," J. Mitroy, M. S. Safronova, and C. W. Clark, J. Phys. B 43, 202001 (2010), https://doi.org/10.1088/0953-4075/43/20/202001, and "2018 Table of static dipole polarizabilities of the neutral elements in the periodic table," P. Schwerdtfeger and J. K. Nagle, Mol. Phys. 117, 1200–1225 (2019), https://doi.org/10.1080/00268976.2018.1535143.

Most molecules are not spherical, and they can be easier to polarize in some directions than in others. A nitrogen molecule, for example, is about one and a half times as easy to polarize along its axis as across it, so a field at an angle to the axis induces a dipole moment tilted toward the axis. Then **p** is not parallel to **E**, and the relationship between them is a matrix equation,

$$\begin{pmatrix} p_x \\ p_y \\ p_z \end{pmatrix} = \begin{pmatrix} \alpha_{xx} & \alpha_{xy} & \alpha_{xz} \\ \alpha_{yx} & \alpha_{yy} & \alpha_{yz} \\ \alpha_{zx} & \alpha_{zy} & \alpha_{zz} \end{pmatrix} \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix},$$

or $p_i = \sum_j \alpha_{ij} E_j$ in the index notation of Chapter 8. The matrix α_{ij} is the **polarizability tensor**. In very strong fields, such as those of intense laser light, the response also stops being linear. This chapter assumes isotropic, linear behavior, (9.5), unless it says otherwise.

9.2 Dielectric materials: polarization and susceptibility

Put a slab of metal and a slab of an insulator side by side in the same uniform field $\mathbf{E}_0$, pointing up (Figure 9.4). In the metal, free electrons move down until the charge on the two faces produces a field inside that exactly cancels $\mathbf{E}_0$: the field inside a conductor is zero (Section 4.5). In the insulator, each atom or molecule polarizes slightly. Inside the slab the positive end of each small dipole sits next to the negative end of its neighbor, and on average they cancel. At the top face, however, there is a layer of uncanceled positive ends, and at the bottom face a layer of uncanceled negative ends. These two layers of charge produce a field inside the slab that points down, opposite to the applied field. It partly cancels the applied field, so the field inside is smaller than $\mathbf{E}_0$ but not zero. Three fields are involved: the *applied* field $\mathbf{E}_0$, the *counter field* of the polarized material, and the total field *inside*, which is their sum. The reduction of the field inside a material by its own polarization is called **screening**.

A conductor is an extreme case of the same thing: its charges move all the way to the surface, and the screening is complete. In a dielectric they move by a small fraction of an atomic diameter, and the screening is partial.

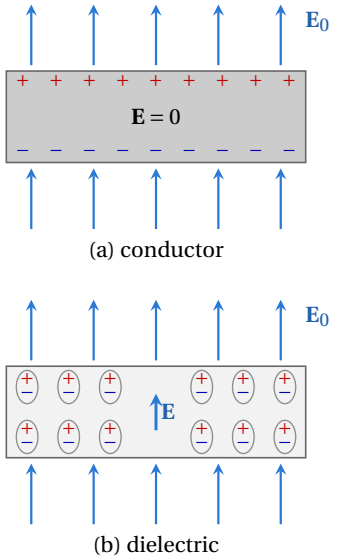


Figure 9.4 A conductor (a) and a dielectric (b) in the same applied field $\mathbf{E}_0$. In the conductor the induced surface charge cancels the applied field completely. In the dielectric each molecule polarizes slightly; the charges inside cancel on average, and the layers left on the two faces produce a counter field that cancels the applied field only partly.

9.2.1 Polarization

A piece of material contains far too many dipoles to keep track of one by one. Instead we describe them by their density. The **polarization P** is the dipole moment per unit volume:

$$\mathbf{P} = \frac{\text{dipole moment of a small volume}}{\text{volume}}. \tag{9.6}$$

The "small volume" is small compared with the object but large enough to contain many molecules, so **P** is a smooth function of position, just as the charge density ρ smooths out the individual charges of Chapter 3. If the material has N molecules

per unit volume, each with dipole moment **p** (or each with the same average **p**), then

$$\mathbf{P} = N\mathbf{p}. \tag{9.7}$$

The total dipole moment of an object is the integral of its polarization,

$$\mathbf{p}_{\text{total}} = \int \mathbf{P}\, d\tau, \tag{9.8}$$

the same way a total charge is the integral of ρ. (Lowercase **p** is the integral of uppercase **P**, the reverse of what you might guess, and the two letters are easy to confuse in handwriting.) The SI unit of **P** is that of a dipole moment divided by a volume, $\text{C m}/\text{m}^3 = \text{C}/\text{m}^2$, the unit of a surface charge density. Section 9.3 shows that this is not a coincidence.

9.2.2 How a material becomes polarized

A material can become polarized in three ways.

1. *Induced dipoles.* A field polarizes each atom or molecule, as in Section 9.1.5. This happens in every material. In an ionic crystal, such as salt, the field also pushes the positive and negative ions slightly apart, which polarizes the crystal in the same way.

2. *Orientation of polar molecules.* A field partly lines up molecules that already have dipole moments, as in Section 9.1.4. In a material made of polar molecules, such as water, this is usually the larger effect.

3. *Built-in polarization.* In some materials the polarization is there even with no applied field.

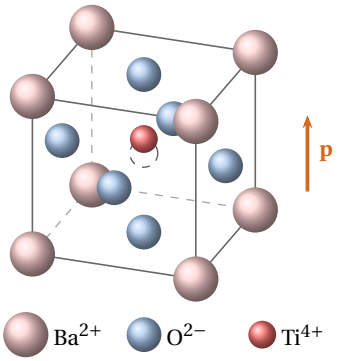


Figure 9.5 A unit cell of barium titanate below about 120°C, drawn as a cube: a barium ion at each corner, an oxygen ion at the center of each face, and a titanium ion that sits slightly above the center (dashed circle), which gives the cell a dipole moment **p**. The shift is exaggerated.

In the first two, the polarization is caused by the field and disappears when the field is removed. In the third it is a property of the material, the electrical counterpart of the magnetization of a permanent magnet. Two kinds of material with built-in polarization are especially important. In **ferroelectrics**, such as barium titanate ($BaTiO_3$), the crystal structure itself is polarized: below about 120°C, the titanium ion in each unit cell sits slightly off center (Figure 9.5), and the cells line up with each other in regions called domains. A strong field can reverse the polarization of a domain, and the polarization stays reversed when the field is removed. The name was chosen by analogy with ferromagnets; ferroelectrics have nothing to do with iron.[6] **Electrets** are materials in which a polarization has

[6] Joseph Valasek, a graduate student at the University of Minnesota, showed in 1920 and 1921 that the polarization of Rochelle salt, plotted against the applied field, traces out a hysteresis loop like the loop that the magnetization of iron traces out in a magnetic field. The word "ferroelectric" had appeared earlier: Erwin Schrödinger used it in 1912, in quotation marks, with an apology for the word.

been frozen in, for example by letting melted wax solidify in a strong field.[7] Both are the electrical analogs of permanent magnets, and the field of a polarized bar of either looks, outside the bar, like the field of a bar magnet, at least until stray charge from the air collects on its faces and cancels most of that field.

9.2.3 Linear dielectrics and the susceptibility

In most materials, as long as the field is not too strong, the polarization is proportional to the field that causes it:

$$\mathbf{P} = \varepsilon_0 \chi_e \mathbf{E}. \tag{9.9}$$

The dimensionless constant χ_e is the **electric susceptibility** of the material; the factor ε_0 is there so that χ_e has no units. It plays the same role for a bulk material that the polarizability α plays for a single atom. A material that obeys (9.9) is called a **linear dielectric**.

One point in (9.9) needs care. The **E** in it is the total field inside the material: the applied field plus the field produced by the polarization itself. That makes problems with dielectrics self-referential, since the polarization depends on a field that depends on the polarization. Sections 9.5–9.8 show how to get around this.

☞ *Caution*: In $\mathbf{P} = \varepsilon_0 \chi_e \mathbf{E}$, **E** is the total field inside the material, including the field of the polarization itself, not the applied field.

Equation (9.9) contains three assumptions, and a material that obeys it is often described by all three adjectives:

- *linear*: **P** is proportional to **E**;
- *isotropic*: **P** is parallel to **E**, whatever the direction of **E**;
- *homogeneous*: χ_e is the same everywhere in the material.

Unless it says otherwise, this book assumes all three whenever it calls a material a linear dielectric. Each assumption can fail. In an anisotropic crystal, **P** is related to **E** by a susceptibility tensor, $P_i = \varepsilon_0 \sum_j \chi_{ij} E_j$, just as for the polarizability tensor of a molecule. In strong fields the response becomes nonlinear, $P = \varepsilon_0 \left(\chi^{(1)} E + \chi^{(2)} E^2 + \cdots\right)$, the starting point of nonlinear optics. And the susceptibility depends on how fast the field changes: a field that oscillates at the frequency of visible light, about 5×10^{14} Hz, is far too fast for water molecules to follow by turning, and the only polarization that can keep up with it is the shifting of electrons. This chapter is about static fields, and the values of χ_e in it are static values. Ferroelectrics and electrets do not obey (9.9) as it stands, since they are

[7] Oliver Heaviside coined the word "electret" for an imagined permanently polarized body, the electrical counterpart of a magnet. Mototarô Eguchi, at the Higher Naval College in Tokyo, made real ones from 1919 on by letting a melted mixture of carnauba wax and resin solidify between electrodes held at a high voltage; in 1925 he reported that their charges had not noticeably decayed in three years. In 1962 Gerhard Sessler and James West at Bell Laboratories used a thin metal-coated polymer electret as the membrane of a microphone that needs no external voltage; electret microphones have since been made by the billions.

polarized even with no field. For small changes in the field, though, a ferroelectric ceramic responds nearly linearly, and the dielectric constant that describes that response can be very large (Section 9.5.1).

9.3 Bound charges

Suppose we know the polarization $\mathbf{P}(\mathbf{r}')$ throughout some object. What field does the object produce? Each small piece of it, of volume $d\tau'$ at $\mathbf{r}'$, is a dipole with moment $d\mathbf{p} = \mathbf{P}\,d\tau'$. The potential of a dipole at the origin is (8.13); for a dipole at $\mathbf{r}'$, the vector from the dipole to the field point is $\boldsymbol{\mathscr{r}} = \mathbf{r} - \mathbf{r}'$, so $\mathbf{r}$ and r are replaced by $\boldsymbol{\mathscr{r}}$ and $\mathscr{r}$. Adding up the pieces,

$$V(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\int \frac{\mathbf{P}(\mathbf{r}')\cdot\hat{\boldsymbol{\mathscr{r}}}}{\mathscr{r}^2}\,d\tau'. \tag{9.10}$$

This integral can be evaluated as it stands, but it can also be rewritten in a much more useful form. The rewriting uses two facts and the divergence theorem.

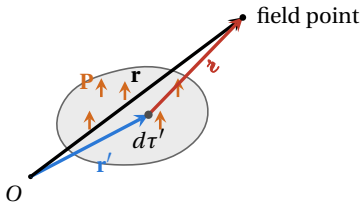


Figure 9.6 A polarized object. The small piece of volume $d\tau'$ at $\mathbf{r}'$ is a dipole with moment $\mathbf{P}\,d\tau'$, and $\boldsymbol{\mathscr{r}} = \mathbf{r} - \mathbf{r}'$ is the separation vector from it to the field point $\mathbf{r}$.

The first fact is that

$$\nabla'\left(\frac{1}{\mathscr{r}}\right) = \frac{\hat{\boldsymbol{\mathscr{r}}}}{\mathscr{r}^2}, \tag{9.11}$$

where ∇' means the gradient with respect to the source coordinates (x', y', z'). To see this, take the x' component:

$$\frac{\partial}{\partial x'}\left[(x-x')^2 + (y-y')^2 + (z-z')^2\right]^{-1/2} = -\frac{1}{2}\left[\cdots\right]^{-3/2}\cdot 2(x-x')(-1) = \frac{x-x'}{\mathscr{r}^3},$$

which is the x component of $\boldsymbol{\mathscr{r}}/\mathscr{r}^3 = \hat{\boldsymbol{\mathscr{r}}}/\mathscr{r}^2$. (The gradient with respect to the field coordinates, $\nabla(1/\mathscr{r})$, is the negative of this; the chain rule supplies the extra minus sign here.) The second fact is the product rule (1.52), $\nabla'\cdot(f\mathbf{A}) = f\,\nabla'\cdot\mathbf{A} + \mathbf{A}\cdot\nabla' f$, which with $f = 1/\mathscr{r}$ and $\mathbf{A} = \mathbf{P}$ gives

$$\mathbf{P}\cdot\nabla'\left(\frac{1}{\mathscr{r}}\right) = \nabla'\cdot\left(\frac{\mathbf{P}}{\mathscr{r}}\right) - \frac{1}{\mathscr{r}}\nabla'\cdot\mathbf{P}.$$

Using the first fact and then the second in (9.10),

$$V(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\int_{\mathcal{V}} \nabla'\cdot\left(\frac{\mathbf{P}}{\mathscr{r}}\right)d\tau' + \frac{1}{4\pi\varepsilon_0}\int_{\mathcal{V}} \frac{-\nabla'\cdot\mathbf{P}}{\mathscr{r}}\,d\tau',$$

where $\mathcal{V}$ is the volume of the object. The divergence theorem, (1.36), turns the first integral into an integral over the object's surface $\mathcal{S}$, with $\hat{\mathbf{n}}$ the outward unit normal:

$$V(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\oint_{\mathcal{S}} \frac{\mathbf{P}\cdot\hat{\mathbf{n}}}{\mathscr{r}}\,da' + \frac{1}{4\pi\varepsilon_0}\int_{\mathcal{V}} \frac{-\nabla'\cdot\mathbf{P}}{\mathscr{r}}\,d\tau'. \tag{9.12}$$

Compare this with the potential of a surface charge and a volume charge, (5.15) and (5.16). The first term is the potential of a surface charge density $\mathbf{P}\cdot\hat{\mathbf{n}}$, and the second is the potential of a volume charge density $-\nabla\cdot\mathbf{P}$. A polarized

object produces exactly the potential, and therefore exactly the field, of the **bound charges**

$$\sigma_b = \mathbf{P}\cdot\hat{\mathbf{n}} \tag{9.13}$$

on its surface and

$$\rho_b = -\nabla\cdot\mathbf{P} \tag{9.14}$$

throughout its volume. This is why $\mathbf{P}$ has the units of a surface charge density. (Where $\mathbf{P}$ stops, at the surface, $-\nabla\cdot\mathbf{P}$ is σ_b times a delta function; Exercise P1.21.)

9.3.1 What the bound charges are

Bound charge is not a mathematical device. It is real charge, made of the same electrons and nuclei as any other charge, and it is there because polarization moves those charges slightly. Figure 9.7 shows how.

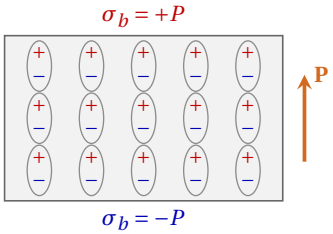
If the polarization is uniform (Figure 9.7a), the dipoles line up head to tail. Inside the material, every positive end is next to a negative end, and the charge cancels on average, so $\rho_b = 0$, consistent with $-\nabla\cdot\mathbf{P} = 0$ for a constant $\mathbf{P}$. At the surfaces, the ends of the chains are left uncovered: positive charge where $\mathbf{P}$ points out of the surface, negative where it points in, and none where $\mathbf{P}$ is parallel to the surface. That is $\sigma_b = \mathbf{P}\cdot\hat{\mathbf{n}}$. The size comes out right, too. A slab of area A and thickness d, polarized perpendicular to its faces, has dipole moment PAd; surface charges $\pm\sigma_b A = \pm PA$ on faces a distance d apart have the same dipole moment, $(PA)d$ (Figure 9.8).

If the polarization is not uniform, the cancellation inside is not complete. In Figure 9.7b the polarization points away from the center and grows with distance from it. Each dipole has moved a little positive charge outward, and because the polarization grows outward, more positive charge leaves any small region inside than enters it from the region closer to the center. The inside is left with a net negative charge, and the positive charge piles up at the outer surface. In terms of $\mathbf{P}$, the flux of $\mathbf{P}$ out of each small region inside is positive, so $\nabla\cdot\mathbf{P} > 0$ and $\rho_b = -\nabla\cdot\mathbf{P}$ is negative.

The name "bound" distinguishes this charge from the charge on a conductor, which moves freely through the material. Bound charge stays attached to its own atom or molecule; the polarization only displaces it by a small fraction of an atomic diameter. Every other charge is called **free charge**, ρ_f or σ_f: the charge on conductors, ions that have been placed in or on a material, and charge rubbed onto an insulator, even though that charge cannot move. The total charge density is the sum,

$$\rho = \rho_b + \rho_f. \tag{9.15}$$

Since bound charge only rearranges the charge that was already there, the total bound charge of a polarized object is zero: the surface charge exactly balances the volume charge. That follows from (9.13), (9.14) and the divergence theorem (Exercise P9.6).



(a) uniform $\mathbf{P}$

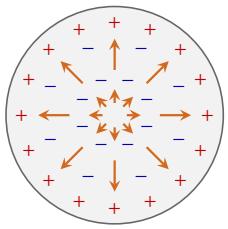
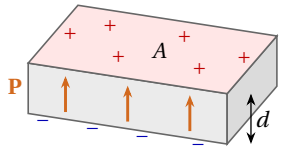
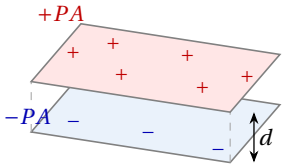

(b) $\mathbf{P} \propto \mathbf{r}$

Figure 9.7 (a) Uniform polarization. Inside, the positive end of each dipole is next to the negative end of the next one, and they cancel; only the ends at the surfaces are left over, as surface charge. (b) Polarization that points away from the center and grows with distance from it, $\mathbf{P} \propto \mathbf{r}$. The dipoles move positive charge outward and leave negative charge behind, so there is a negative volume charge inside and a positive surface charge on the outside.



(a) $p = PAd$



(b) $p = (PA)\,d$

Figure 9.8 (a) A slab of area A and thickness d with polarization $\mathbf{P}$ perpendicular to its faces has the dipole moment PAd. (b) Two sheets with charges $+PA$ and $-PA$ a distance d apart have the same dipole moment, $(PA)d$.

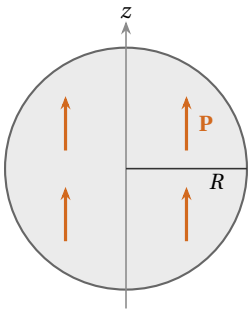


Figure 9.9 The uniformly polarized sphere of Example 9.3.

9.4 Solving for E when given P: bound-charge method

Equations (9.13) and (9.14) turn the problem of finding the field of a polarized object into a problem that earlier chapters have already solved. The **bound-charge method** has two steps:

1. Find the bound charges, $\sigma_b = \mathbf{P}\cdot\hat{\mathbf{n}}$ on every surface and $\rho_b = -\nabla\cdot\mathbf{P}$ inside.
2. Find the field or potential of those charges by any method that suits them: the script-r method of Sections 3.5 and 5.5, Gauss's law, separation of variables, the dipole approximation far away, or a numerical integration.

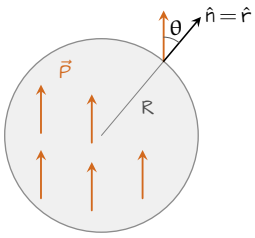
Step 2 is a problem from an earlier chapter. In that sense the method works the way the method of images does (Section 7.2): it replaces the problem by a different one that is easier to solve.

The polarization in this section is taken as given, as it would be in an electret or a ferroelectric. If the polarization was induced by an applied field, the field found by this method is only the field of the polarized object, and the applied field must be added to it to get the total field. (Sections 9.5–9.8 deal with the harder problem in which the polarization depends on the field.)

Example 9.3

A sphere of radius R is uniformly polarized, $\mathbf{P} = P\hat{\mathbf{z}}$. Find the field inside and outside.

Solution:



A cross section of the sphere shows what is given: $\mathbf{P}$ is the same everywhere inside, pointing up. For the bound surface charge we will also need the outward normal at the surface, which is $\hat{\mathbf{r}}$, at the angle θ from $\mathbf{P}$.

Step 1. The polarization is constant, so $\rho_b = -\nabla\cdot\mathbf{P} = 0$. On the surface, $\hat{\mathbf{n}} = \hat{\mathbf{r}}$, so

$$\sigma_b = P\hat{\mathbf{z}}\cdot\hat{\mathbf{r}} = P\cos\theta.$$

The bound charge is positive on the upper hemisphere and negative on the lower one.

Step 2. This is exactly the surface charge of Example 7.5, with $\sigma_0 = P$, and (7.48) gives its potential:

$$V_{\text{in}} = \frac{P}{3\varepsilon_0}\,r\cos\theta = \frac{P}{3\varepsilon_0}\,z, \qquad V_{\text{out}} = \frac{PR^3}{3\varepsilon_0}\,\frac{\cos\theta}{r^2}.$$

Inside, V depends only on z, so the field is $-\partial V/\partial z$ in the $\hat{\mathbf{z}}$ direction:

$$\mathbf{E}_{\text{in}} = -\frac{\mathbf{P}}{3\varepsilon_0}. \tag{9.16}$$

The field inside is uniform, and it points opposite to the polarization. Outside, compare V_{out} with the potential of a pure dipole, (8.14): it is exactly the potential of a dipole at the center with moment

$$\mathbf{p} = \frac{4}{3}\pi R^3 \mathbf{P},$$

which is the sphere's total dipole moment, (9.8). The field outside is therefore the dipole field (8.15), and it is exact everywhere outside the sphere, not just far away.

Reality check: At $r = R$ both potentials reduce to $(PR/3\varepsilon_0)\cos\theta$, so the potential is continuous through the surface, as it must be.

Figure 9.23b, in Section 9.8.4, shows the field lines. They begin on the positive bound charge of the upper hemisphere and end on the negative bound charge of the lower one, outside along the curves of a dipole field and inside straight down. That the field of the bound charges points against the polarization inside is typical: the bound charges sit at the "head" end of $\mathbf{P}$, which is positive, and at the "tail" end, which is negative, and the field inside points from the positive end to the negative end.

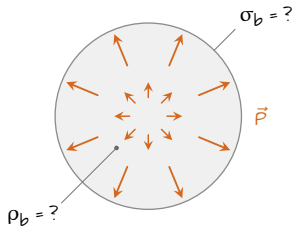
Example 9.4

A sphere of radius R has the polarization

$$\mathbf{P} = \frac{P_0}{R}\mathbf{r} = P_0\frac{r}{R}\hat{\mathbf{r}},$$

which points outward and grows linearly from zero at the center to P_0 at the surface. Find the bound charges and the field inside and outside.

Solution:



We sketch the sphere with arrows for $\mathbf{P}$, short near the center and longer near the surface, and ask what bound charge they leave behind, inside and on the surface.

Step 1. At the surface, $\hat{\mathbf{n}} = \hat{\mathbf{r}}$ and $r = R$, so $\sigma_b = P_0$. Inside, using the divergence in spherical coordinates, (1.30),

$$\rho_b = -\left[\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{P_0 r}{R}\right) + \cancel{\frac{1}{r\sin\theta}\frac{\partial(\sin\theta P_\theta)}{\partial\theta}} + \cancel{\frac{1}{r\sin\theta}\frac{\partial P_\phi}{\partial\phi}}\right]$$
$$= -\frac{3P_0}{R}.$$

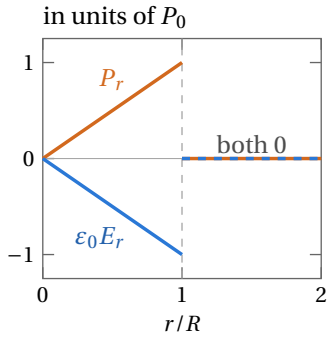


Figure 9.10 The polarization and the field of Example 9.4 along a radius, in units of P_0. Inside the sphere $\varepsilon_0\mathbf{E} = -\mathbf{P}$; outside, both are zero. At $r = R$, E_r jumps by $\sigma_b/\varepsilon_0 = P_0/\varepsilon_0$, as it must at a surface charge.

This is the situation of Figure 9.7b: a uniform negative charge density inside and a positive surface charge. Their totals are $-(3P_0/R)\,\frac{4}{3}\pi R^3 = -4\pi P_0 R^2$ and $P_0 \cdot 4\pi R^2$, which add to zero, as they must.

Step 2. The charge is spherically symmetric, so Gauss's law applies, with a concentric Gaussian sphere of radius r. Inside the sphere ($r < R$) the enclosed charge is $-(3P_0/R)\,\frac{4}{3}\pi r^3 = -4\pi P_0 r^3/R$, so

$$E \cdot 4\pi r^2 = -\frac{4\pi P_0 r^3}{\varepsilon_0 R}, \qquad \mathbf{E}_{\text{in}} = -\frac{P_0}{\varepsilon_0}\frac{r}{R}\hat{\mathbf{r}} = -\frac{\mathbf{P}}{\varepsilon_0}.$$

Outside, the enclosed charge is the total bound charge, which is zero, so $\mathbf{E}_{\text{out}} = 0$.

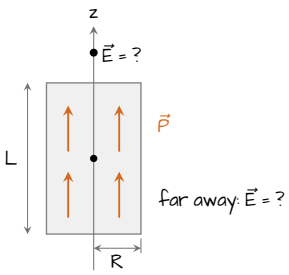
Reality check: As expected from the signs of the bound charges, the field inside points inward (Figure 9.10).

The answer in Example 9.4 came out simply because the bound charges were spherically symmetric. Usually they are not, and Step 2 takes more work.

Example 9.5

A cylinder of radius R and length L, centered on the origin with its axis along z, is uniformly polarized along its axis, $\mathbf{P} = P_0\hat{\mathbf{z}}$. Find the bound charges, the field on the axis, and the field far away.

Solution:



To see what is given and what is asked, we draw the cylinder in cross section, with $\mathbf{P}$ along its axis, and mark where the field is wanted: on the axis, and far away.

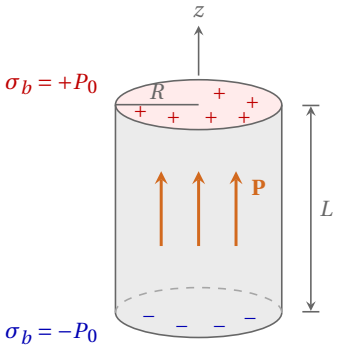


Figure 9.11 The uniformly polarized cylinder of Example 9.5. Its bound charge is $+P_0$ on the top face and $-P_0$ on the bottom face; there is none on the wrapper and none inside.

Step 1. Since $\mathbf{P}$ is constant, $\rho_b = 0$. On the top face $\hat{\mathbf{n}} = \hat{\mathbf{z}}$, so $\sigma_b = +P_0$; on the bottom face $\hat{\mathbf{n}} = -\hat{\mathbf{z}}$, so $\sigma_b = -P_0$; on the curved side (the wrapper) $\hat{\mathbf{n}} = \hat{\mathbf{s}}$ is perpendicular to $\mathbf{P}$, so $\sigma_b = 0$. The cylinder produces the field of two uniformly charged disks, $+P_0$ at $z = L/2$ and $-P_0$ at $z = -L/2$ (Figure 9.11).

Step 2. At a general point there is no symmetry to help, and the field is the sum of two double integrals over the disks, which can be set up as in Example 3.5 and done numerically. On the axis they can be done exactly. For a single disk of radius R with uniform charge density σ, at a height $u > 0$ above its center, the script-r method gives, as in Example 3.5 but with σ constant,

$$E_z = \frac{\sigma u}{2\varepsilon_0}\int_0^R \frac{s'\,ds'}{(u^2 + s'^2)^{3/2}} = \frac{\sigma}{2\varepsilon_0}\left(1 - \frac{u}{\sqrt{u^2 + R^2}}\right),$$

pointing away from the disk if $\sigma > 0$. At the center of the cylinder, each disk is a distance $L/2$ away. The field of the positive top disk points down, away from it, and the field of the negative bottom disk also points down, toward it, so the two add:

$$\mathbf{E}(0) = -\frac{P_0}{\varepsilon_0}\left(1 - \frac{L}{\sqrt{L^2+4R^2}}\right)\hat{\mathbf{z}}. \tag{9.17}$$

Again the field inside points against $\mathbf{P}$. Its size depends on the shape: for $L = 2R$ the factor in parentheses is $1 - 1/\sqrt{2} = 0.293$.

Far from the cylinder, compared with both R and L, the field is that of a dipole with the cylinder's total dipole moment, $\mathbf{p} = \pi R^2 L P_0 \hat{\mathbf{z}}$, given by (8.15).

Reality check: For a thin wafer, $L \ll R$, the two disks act like two infinite sheets of charge, and the field inside approaches $-\mathbf{P}/\varepsilon_0$, the field between the plates of a capacitor with surface charge $\pm P_0$. For a long rod, $L \gg R$, the charged ends are far from the center, and the field there goes to zero.

9.4.1 The field inside a dielectric

The bound charges are smooth densities, and the field they produce is smooth. The actual field inside a material is not: next to a nucleus it is enormous, and it changes direction over a distance of an atomic diameter. The field of the bound charges is the **macroscopic field**, the actual field averaged over a region that contains many molecules but is small compared with the object. It is the field that appears in $\mathbf{P} = \varepsilon_0 \chi_e \mathbf{E}$, and it is the field meant whenever this book speaks of the field inside a material.

The derivation of (9.12) assumed that the field point is outside the object, far from every molecule compared with the size of a molecule, so that each small piece of the material could be treated as a point dipole. Inside the object that is not true of the nearest molecules, but the bound charges still give the macroscopic field. To see why, surround the field point with a small sphere that still contains many molecules, and average the field over the sphere. The material outside the sphere is far enough away for the derivation to apply as it stands, and its field is the field of its bound charges: those of the object, and also a surface charge $\mathbf{P}\cdot\hat{\mathbf{n}}$ on the wall of the spherical hole, where $\hat{\mathbf{n}}$ points into the hole. That wall charge is the surface charge of a uniformly polarized sphere with the sign reversed, so its field in the hole is $+\mathbf{P}/3\varepsilon_0$, by (9.16). The molecules inside the sphere have a total dipole moment of $\frac{4}{3}\pi R^3 \mathbf{P}$, and the average of their field over the sphere turns out to depend only on that dipole moment:[8] it is $-\mathbf{P}/3\varepsilon_0$, which cancels the field of the wall charge. What is left is the field of the bound charges of the object, as if

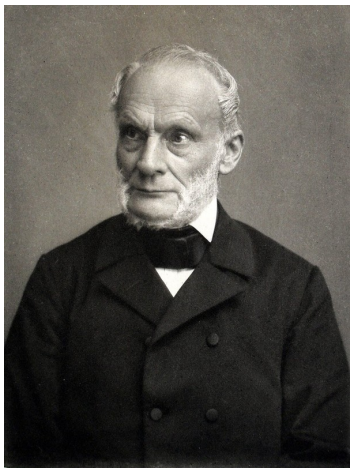
[8] For any charges inside a sphere of radius R with total dipole moment $\mathbf{p}$, the field averaged over the volume of the sphere is $-\mathbf{p}/4\pi\varepsilon_0 R^3$. For a single point charge q at $\mathbf{r}'$ inside the sphere, for example, the average is $-q\mathbf{r}'/4\pi\varepsilon_0 R^3$.



Ottaviano-Fabrizio Mossotti (1791–1863, Italian) After studying at Pavia, Mossotti worked as an astronomer at the Brera Observatory in Milan until his liberal politics drove him into exile in 1823. He spent four years in London and eight in Buenos Aires, where he worked as an astronomer and surveyor and taught physics. Back in Europe, he published in 1836 a theory of the forces between the molecules of matter that Faraday admired, taught at the Ionian University on Corfu, and in 1841 became professor of mathematical physics at Pisa. In 1846 he worked out how the polarization of the molecules of a dielectric changes the forces between charged bodies; his result, printed in 1850, is the original form of the Clausius–Mossotti relation. In 1848, at 57, he served with the Pisa university battalion of professors and students at the battle of Curtatone and Montanara, and in 1861 he was named a senator of the new Kingdom of Italy. (Wikipedia)

no hole had been cut, so the bound charges give the average field at every point, inside the object or out.

The field that polarizes a particular molecule is something else again. It is the field of everything *except* that molecule: the applied field and the fields of all the other molecules, but not the molecule's own field, which is the part of the macroscopic field that the molecule itself contributes. In a gas the molecules are far apart, and the difference is small. In a liquid or solid it is not. The tutorial below works out the difference for a simple model of the molecules and uses it to connect the polarizability α of a single molecule to the susceptibility χ_e of the material.



Rudolf Clausius (1822–1888, German) Clausius studied at the University of Berlin and took his doctorate at Halle in 1848. His paper of 1850 on the mechanical theory of heat reconciled Carnot's analysis of heat engines with the conservation of energy and gave one of the first statements of the second law of thermodynamics. In 1865 he introduced the word *entropy*, and he ended that paper with "The energy of the universe is constant. The entropy of the universe tends to a maximum." He also introduced the mean free path into the kinetic theory of gases and the virial theorem into mechanics. He was a professor at the Polytechnic in Zurich, then at Würzburg and, from 1869, at Bonn. In the Franco-Prussian War of 1870 he organized a volunteer ambulance corps and was wounded, which left him with a lasting disability. His book of 1879 on the mechanical treatment of electricity contains the relation between polarizability and dielectric constant that bears his name and Mossotti's. (Wikipedia)

Tutorial 9.1 Polarization versus polarizability

Printable version

Background

If a dielectric material is exposed to an external electric field, the atoms will polarize. For linear, isotropic materials, the polarization field is given by $\mathbf{P} = \varepsilon_0 \chi_e \mathbf{E}$, where χ_e is the electric susceptibility and $\mathbf{E}$ is the electric field inside the material.

Meanwhile, if an individual atom is exposed to an electric field, the atom will polarize to create a dipole moment. For linear, isotropic situations, the dipole moment is given by $\mathbf{p} = \alpha \mathbf{E}$, where α is the atomic polarizability and $\mathbf{E}$ is the electric field at the location of the dipole.

There is a subtlety here, which is that the two electric fields in the equations for $\mathbf{P}$ and $\mathbf{p}$ are not the same. Focusing on one representative dipole now, the electric field to be used in the $\mathbf{P}$ equation is the overall spatially averaged electric field, which includes the field from that dipole itself, but the electric field to be used in the $\mathbf{p}$ equation is the field from *all the other sources*, including any applied external electric field and the field from the other dipoles, but *not* including its own field. Written out explicitly,

$$\mathbf{P} = \varepsilon_0 \chi_e \left(\mathbf{E}_{\text{self}} + \mathbf{E}_{\text{all other}}\right), \tag{9.18}$$

$$\mathbf{p} = \alpha \mathbf{E}_{\text{all other}}. \tag{9.19}$$

Meanwhile, because $\mathbf{P}$ is the density of dipole moments, $\mathbf{P}$ and $\mathbf{p}$ are related through

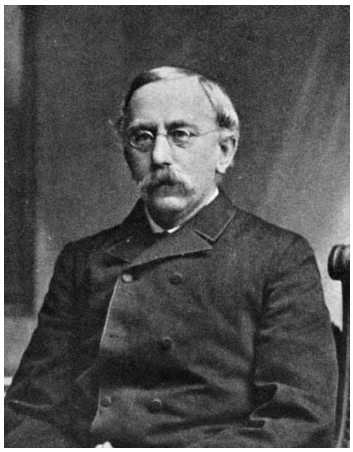
$$\mathbf{P} = N\mathbf{p}, \tag{9.20}$$

where N is the "number density" of dipole moments, namely the number of dipole moments per volume.

The goal of this tutorial is to use these concepts to figure out an equation that connects χ_e and α; it is called the **Clausius–Mossotti relation**.[a] This equation was an important development in the understanding of electric fields inside materials, since χ_e is a measurable property of bulk materials, and α is a property of isolated atoms or molecules, which can be measured in a dilute gas, or in a beam of atoms or molecules traveling through a vacuum.

[a] It is named for Ottaviano-Fabrizio Mossotti, in a memoir of 1846 (printed in 1850), and Rudolf Clausius, who arrived at it independently in a book of 1879. It is usually

written in terms of ε_r instead of χ_e, using $\varepsilon_r = \chi_e + 1$ (Section 9.5.1). When it is written in terms of the index of refraction n, using $\varepsilon_r = n^2$ (which holds for a nonmagnetic material when ε_r is measured at the frequency of the light), it is also called the Lorentz–Lorenz equation, for Ludvig Lorenz and Hendrik Lorentz, who also found it independently of each other, in 1869 and 1878. All four have their names attached to it.



Ludvig Lorenz (1829–1891, Danish) Lorenz graduated in chemistry from the Polytechnic College in Copenhagen in 1852 and taught himself physics. He never held a university post: from 1866 he taught physics at the Military High School near Copenhagen, and from 1887 the Carlsberg Foundation paid him to do research for the rest of his life. In 1867, without knowing of Maxwell's theory, he argued that light consists of electrical oscillations, and in the same paper he wrote the potentials as retarded integrals and introduced the gauge condition that now bears his name. In 1869 he found the relation between the index of refraction of a substance and its density that Hendrik Lorentz found independently in 1878; Lorentz called it "a curious case of coincidence." In 1890 he worked out the scattering of light by a sphere, anticipating much of the theory that Gustav Mie published in 1908. The Lorenz gauge was long credited to Lorentz, and the two names are still often confused. (Wikipedia)

Part 1

We will start by supposing that the field from all the other sources is uniform near the atom and points in the z direction, which causes the atom to polarize, also in the z direction. We will model the polarized atom as a small spherical shell of radius R and say that the induced surface charge density of the shell is given by $\sigma = \sigma_0 \cos\theta$.

(a) Draw the sphere in the uniform electric field with the induced positive and negative charges on it. Does the form of σ seem reasonable? Note that in the context of the above equations, the uniform electric field is $\mathbf{E}_{\text{all other}}$ and the electric field from the spherical shell is $\mathbf{E}_{\text{self}}$.

(b) From the information given, consider how you might go about relating χ_e and α. Think about what equations, quantities, and relations you would want to use or find along the way. It is fine if your ideas here don't match the method presented below.

(c) Calculate the dipole moment of the spherical shell in terms of σ_0 and R. Use symmetry to simplify the integral, and do not violate Colton's First Commandment. You should be able to use this integral: $\int_0^\pi \cos^2\theta \sin\theta \, d\theta = 2/3$.

(d) Find the field at the center of the sphere from itself, in terms of σ_0. This is $\mathbf{E}_{\text{self}}$. There are a few ways of doing this, but please do it with the script-r method for finding $\mathbf{E}$ (Section 3.5). Normally I wouldn't recommend the script-r method, but since you only want the field at a single point, and it's a pretty easy point (the center), it turns out to be straightforward. You can again use symmetry to simplify the integral, and remember again not to violate Colton's First Commandment.

(e) Combine your answers from 1(c) and 1(d) to show that $\mathbf{E}_{\text{self}} = -\mathbf{p}/4\pi\varepsilon_0 R^3$.

(f) Describe with words what you have done and found out in this part of the problem.

Part 2

(a) The number density N is the number of atoms per volume; this is also the inverse of the volume per atom, $\frac{4}{3}\pi R^3$. Use this along with (9.19) and your answer to 1(e) to deduce the relationship between $\mathbf{E}_{\text{self}}$ and $\mathbf{E}_{\text{all other}}$ in terms of N and α: $\mathbf{E}_{\text{self}} = (\text{stuff})\,\mathbf{E}_{\text{all other}}$.

(b) Describe with words what this result means, and discuss whether it makes sense in terms of its sign and its dependences on N and α.

Part 3

(a) You are now in a position to use (9.18), (9.19), and (9.20), along with your

result from Part 2, to determine the Clausius–Mossotti equation, which gives α as a function of χ_e. The equation will also depend on N and ε_0. *Hint*: Start with (9.18). Use the other equations to put everything in terms of $\mathbf{E}_{\text{all other}}$, which you can then cancel from all terms. What's left can be solved for α.

(b) Use at least one of the usual methods to consider whether the answer you obtained is reasonable.

Part 4

(a) How did the actual derivation differ from your original thoughts in 1(b) about how you might relate χ_e and α? What did you learn through this derivation that you didn't originally consider?

(b) Summarize the key differences between the electric fields used in the definitions of polarizability and susceptibility.

9.5 The D field and permittivity

In most problems with dielectrics the polarization is not given. What is given, or what we control, is the free charge: the charge we put on the plates of a capacitor, or on a conducting sphere. The polarization then depends on the field, and the field depends on the polarization, through the bound charge it produces. Gauss's law, $\nabla \cdot \mathbf{E} = \rho/\varepsilon_0$, holds as always, but the ρ in it is the total charge density, (9.15), and the bound part of it is not known until the problem is solved.

What we would like is a field whose sources are the free charges alone. It is found by moving the bound charge to the other side of Gauss's law. The **electric displacement**, or **displacement field**, or simply the **D field**, is defined as[9]

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P}. \tag{9.21}$$

Section 9.6 shows that its divergence is the free charge density. The ε_0 is there only to make the units agree: $\mathbf{P}$, and therefore $\mathbf{D}$, has the units of a surface charge density, C/m^2, while $\mathbf{E}$ has those of C/m^2 divided by ε_0. There is nothing more to the definition than that. $\mathbf{D}$ is a combination of the field and the polarization, and its usefulness comes entirely from what its divergence turns out to be. Equation (9.21) is a definition, so it holds in every material, linear or not.

[9]The name is Maxwell's. He pictured an electric field as straining the medium it occupies, displacing electricity slightly within each of its particles, "a kind of elastic yielding to the action of the force." In his 1865 paper on the electromagnetic field the displacement had the components f, g, h, and the letter D stood for the dielectric constant. The letter went to the displacement, in German type, in his *Treatise on Electricity and Magnetism* of 1873.

9.5.1 Linear dielectrics

In a linear dielectric, (9.9) gives $\mathbf{D} = \varepsilon_0\mathbf{E} + \varepsilon_0\chi_e\mathbf{E} = \varepsilon_0(1+\chi_e)\mathbf{E}$. So $\mathbf{D}$ is proportional to $\mathbf{E}$,

$$\mathbf{D} = \varepsilon\,\mathbf{E}, \tag{9.22}$$

where

$$\varepsilon = \varepsilon_0\left(1+\chi_e\right) \tag{9.23}$$

is the **permittivity** of the material. The ratio

$$\varepsilon_r = \frac{\varepsilon}{\varepsilon_0} = 1+\chi_e \tag{9.24}$$

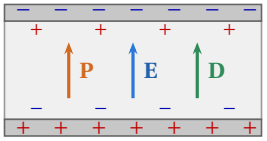
is the **relative permittivity**, more often called the **dielectric constant**. (Introductory textbooks often call it κ.) For vacuum $\chi_e = 0$ and $\varepsilon_r = 1$; for any ordinary material $\chi_e > 0$ and $\varepsilon_r > 1$. Table 9.2 lists some values.[10] The dielectric constants of gases at ordinary pressures are very close to 1, because their molecules are far apart: for air, ε_r differs from 1 by about 0.05 percent, and even for a gas of strongly polar molecules, such as ammonia, it differs by less than 1 percent. Most solid and liquid insulators have dielectric constants between 2 and 10. Water, with its strongly polar molecules (Section 9.1.4), has a dielectric constant of 80, and ferroelectric ceramics such as barium titanate have dielectric constants in the thousands.

Since χ_e and ε are positive, $\mathbf{P}$, $\mathbf{E}$, and $\mathbf{D}$ all point in the same direction in a linear dielectric. That is worth remembering, because the *field of the bound charges* usually points the other way, as in Examples 9.3–9.5. The polarization and the field it produces are different things. In a material with built-in polarization, which is not a linear dielectric, $\mathbf{P}$ and the total field can point in opposite directions (Figure 9.12); inside the sphere of Example 9.4, for instance, $\mathbf{E} = -\mathbf{P}/\varepsilon_0$.

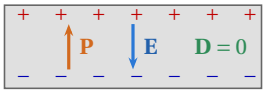
Not every equation of this chapter holds in every material. The definition (9.21), the bound charges (9.13) and (9.14), and Gauss's law for $\mathbf{D}$ in the next section hold always. Equations (9.9) and (9.22)–(9.24), and anything derived from them, hold only in linear dielectrics. The summary at the end of the chapter sorts the equations into these two groups.

Material	ε_r
Vacuum	1 (exactly)
Air	1.000536
PTFE (Teflon)	2.1
Polyethylene	2.3
Fused silica	3.8
Glass	4–10
Mica	4.5–8
Silicon	11.7
Germanium	16
Water	80.1
$BaTiO_3$ ceramic	over 1000

Table 9.2 Dielectric constants of some materials, near room temperature (sources in the footnote).



(a) linear dielectric



(b) built-in polarization

Figure 9.12 (a) In a linear dielectric, here between the plates of a charged capacitor, $\mathbf{P}$, $\mathbf{E}$, and $\mathbf{D}$ all point the same way (large signs: free charge; small signs: bound charge). (b) In a wide slab with built-in polarization and no free charge, the field inside, $\mathbf{E} = -\mathbf{P}/\varepsilon_0$, points against $\mathbf{P}$, and $\mathbf{D} = 0$.

9.6 Gauss's law for D

Gauss's law with the total charge density, (9.15), and $\rho_b = -\nabla\cdot\mathbf{P}$ give

$$\varepsilon_0\nabla\cdot\mathbf{E} = \rho_b + \rho_f = -\nabla\cdot\mathbf{P} + \rho_f, \qquad \text{so} \qquad \nabla\cdot(\varepsilon_0\mathbf{E}+\mathbf{P}) = \rho_f.$$

[10] The values for air (dry, at 20°C and 1 atmosphere) and for water (at 20°C) are measurements compiled by the U.S. National Bureau of Standards: *Table of Dielectric Constants and Electric Dipole Moments of Substances in the Gaseous State*, A. A. Maryott and F. Buckley, NBS Circular 537 (1953), https://doi.org/10.6028/NBS.CIRC.537, and "Dielectric constant of water from 0° to 100°C," C. G. Malmberg and A. A. Maryott, J. Res. Natl. Bur. Stand. 56, 1 (1956), https://nvlpubs.nist.gov/nistpubs/jres/56/jresv56n1p1_a1b.pdf. The others are typical values; they depend on composition, purity, temperature, and frequency.

By the definition (9.21), then,

$$\nabla \cdot \mathbf{D} = \rho_f. \tag{9.25}$$

The divergence of **D** is the free charge density alone. By the divergence theorem, the integral form is

$$\oint \mathbf{D} \cdot d\mathbf{a} = q_{f,\text{enc}}: \tag{9.26}$$

the flux of **D** out of a closed surface is the free charge enclosed. This is **Gauss's law for D**. Only the definitions of **D** and of the bound charge went into it, so it holds in every material.

It is used exactly the way Gauss's law for **E** was used in Section 4.3, and Colton's Second Commandment applies to it: draw the Gaussian surface. When the free charge, and everything else in the problem, has spherical, cylindrical, or planar symmetry, the flux integral reduces to D times an area, and (9.26) gives D directly from the free charge, without any need to know the bound charge. That is what makes it useful, and also what makes it easy to misuse.

9.6.1 D is not determined by the free charge alone

It is tempting to conclude from (9.25) that **D** is simply "the field of the free charge," related to ρ_f the way **E** is related to ρ. It is not. The electric field is determined by its sources because both its divergence and its curl are known: $\nabla \cdot \mathbf{E} = \rho/\varepsilon_0$ and $\nabla \times \mathbf{E} = 0$ (Section 2.3). The curl of **D** is

$$\nabla \times \mathbf{D} = \varepsilon_0 \nabla \times \mathbf{E} + \nabla \times \mathbf{P} = \nabla \times \mathbf{P}, \tag{9.27}$$

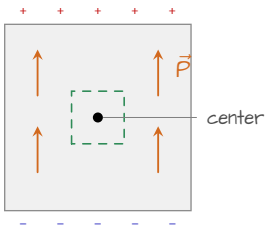
which need not be zero. Wherever the polarization has a curl, **D** has a source other than the free charge. The usual place is the surface of a polarized object, where the component of **P** parallel to the surface jumps from its value inside to zero outside; on the wrapper of the cylinder of Example 9.5, for instance. Then **D** is not zero even where there is no free charge anywhere.

Gauss's law for **D** is still true in such cases; it is just not useful, because **D** cannot be taken out of the integral. Only symmetry can tell you when it can be.

Example 9.6

A cube has uniform built-in polarization **P**, perpendicular to two of its faces, and there is no free charge anywhere. True or false: **D** is zero at the center of the cube.

Solution:



To visualize the question, we sketch the cube in cross section, with **P** pointing up, perpendicular to the top and bottom faces. That means all of the various dipole moments also point upwards, and the bound surface charges will be positive on the top surface and negative on the bottom surface (depicted). We can draw a small Gaussian surface (dashed) around the center. That surface encloses no free charge, so the flux of **D** through it is zero. Does that make **D** zero at the center?

No. The tempting argument is: draw a Gaussian surface around any region inside the cube, like the one in the sketch; it encloses no free charge, so $\oint \mathbf{D}\cdot d\mathbf{a} = 0$, so $D = 0$. The last step is wrong. The flux of **D** through every closed surface is zero, but that does not make **D** itself zero unless symmetry makes **D** constant and perpendicular over the surface, and a cube does not have that symmetry. (I have noticed that most students make this mistake when they first meet a problem like this one.)

To find the actual value, use the bound-charge method. Since **P** is constant, $\rho_b = 0$, and the bound charge is $\sigma_b = +P$ on the face that **P** points out of and $-P$ on the opposite face; the other four faces are parallel to **P** and carry none. At the center, the field of each charged face points along the normal to it (the components parallel to the face cancel by symmetry), and the normal component has a simple form. For a patch of area da' at distance $\mathscr{r}$, the normal component of its field is $(\sigma/4\pi\varepsilon_0)\,da'\cos\gamma/\mathscr{r}^2$, where γ is the angle between $\boldsymbol{\mathscr{r}}$ and the normal, and $da'\cos\gamma/\mathscr{r}^2 = d\Omega$ is the solid angle that the patch subtends (Figure 9.13a). So a uniformly charged flat face produces a normal field of $\sigma/4\pi\varepsilon_0$ times the solid angle the whole face subtends. At the center of a cube each face subtends one sixth of the full solid angle 4π. (Why does each face subtend exactly one sixth of the full solid angle? Answer in footnote.[11]) The positive face pushes and the negative face pulls in the same direction, opposite to **P**, so

$$\mathbf{E}(\text{center}) = -2\,\frac{P}{4\pi\varepsilon_0}\,\frac{4\pi}{6}\,\hat{\mathbf{P}} = -\frac{\mathbf{P}}{3\varepsilon_0}, \qquad \mathbf{D}(\text{center}) = \varepsilon_0\mathbf{E} + \mathbf{P} = \frac{2}{3}\,\mathbf{P}.$$

Reality check: A cylinder whose length and diameter both equal the cube's side fits inside the cube, with its charged ends on the cube's charged faces. Each face subtends a larger solid angle at the center than the end that lies on it, so the field at the center of the cube should be stronger than at the center of the cylinder, $0.293\,P/\varepsilon_0$ (Example 9.5 with $L = 2R$), and it is: $P/3\varepsilon_0 = 0.333\,P/\varepsilon_0$.

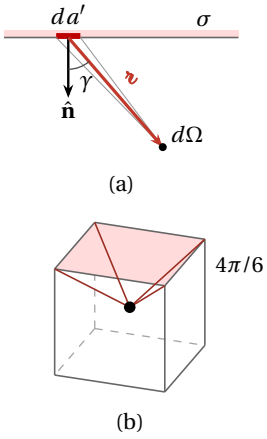
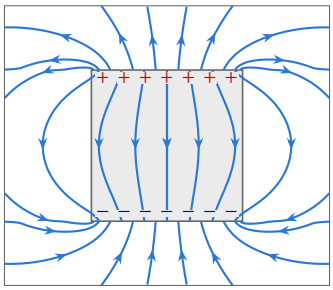
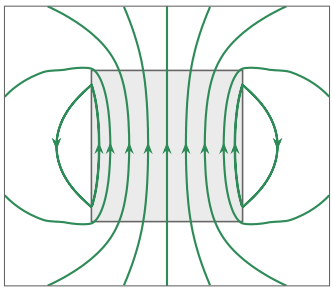


Figure 9.13 (a) A patch of area da' of a uniformly charged flat face, seen edge on, and a point at the distance $\mathscr{r}$ from it. The component of the patch's field along the normal $\hat{\mathbf{n}}$ is $(\sigma/4\pi\varepsilon_0)\,d\Omega$, where $d\Omega = da'\cos\gamma/\mathscr{r}^2$ is the solid angle that the patch subtends at the point. (b) Seen from the center of a cube, each face subtends one sixth of the full solid angle, $4\pi/6$.

[11] Every direction from the center points through exactly one of the six faces, so their solid angles add up to the full 4π; and seen from the center, the six faces all look the same, so each has the same share.



(a) **E**



(b) **D**

Figure 9.14 Field lines of (a) **E** and (b) **D** for the uniformly polarized cylinder of Example 9.5, with $L = 2R$ and no free charge. The lines of **E** begin on the positive bound charge of the top face and end on the negative bound charge of the bottom face; inside, they point against **P**. The lines of **D** never begin or end: inside they point along **P**, they pass through the faces, and outside the cylinder $\mathbf{D} = \varepsilon_0\mathbf{E}$. The spacing of the lines shows the strength of the field only roughly.

The uniformly polarized sphere of Example 9.3 has the same value everywhere inside, $\mathbf{D} = \varepsilon_0(-\mathbf{P}/3\varepsilon_0) + \mathbf{P} = \frac{2}{3}\mathbf{P}$. In the cylinder of Example 9.5, (9.17) gives $D_z = P_0 L/\sqrt{L^2+4R^2}$ at the center, which is $0.707P_0$ for $L = 2R$. None of these is zero, although none of them contains any free charge (Figure 9.14). In the cube the field is not even uniform: it is stronger near the charged faces, and weaker near the others, than at the center (Section 9.9.1).

9.6.2 Bound charge in a linear dielectric

In a linear dielectric that is homogeneous, χ_e is the same everywhere, and $\mathbf{P} = \varepsilon_0\chi_e\mathbf{E} = (\chi_e/(1+\chi_e))\,\mathbf{D}$. Then

$$\begin{aligned}\rho_b &= -\nabla\cdot\mathbf{P} \\ &= -\frac{\chi_e}{1+\chi_e}\nabla\cdot\mathbf{D} \\ &= -\frac{\chi_e}{1+\chi_e}\rho_f.\end{aligned} \tag{9.28}$$

Bound volume charge appears only where there is free charge, and it partly cancels the free charge, leaving the net charge ρ_f/ε_r. In particular, if a homogeneous linear dielectric contains no free charge, there is no bound charge inside it either, and all of its bound charge is on its surfaces. Inside, then, $\rho = 0$, and the potential satisfies Laplace's equation, (5.27). Section 9.8.4 uses that.

9.6.3 What the dielectric constant measures

Suppose a homogeneous linear dielectric fills all the space where there is any field, and the free charges, including any charge on conductors, are the same as they would be in vacuum. Then the divergence of **D** is the free charge density, and its curl is $\nabla\times(\varepsilon\mathbf{E}) = \varepsilon\nabla\times\mathbf{E} = 0$. In vacuum the field $\mathbf{E}_{\text{vac}}$ of the same free charges obeys $\nabla\cdot(\varepsilon_0\mathbf{E}_{\text{vac}}) = \rho_f$ and $\nabla\times(\varepsilon_0\mathbf{E}_{\text{vac}}) = 0$. The two fields **D** and $\varepsilon_0\mathbf{E}_{\text{vac}}$ have the same divergence and the same curl everywhere (and both go to zero far away), so they are the same field (Section 2.3). Therefore

$$\mathbf{E} = \frac{\mathbf{D}}{\varepsilon} = \frac{\mathbf{E}_{\text{vac}}}{\varepsilon_r}. \tag{9.29}$$

The dielectric reduces the field of the free charges everywhere by the factor ε_r. A point charge q embedded in a large block of dielectric, for example, has the field $q/4\pi\varepsilon r^2$: the dielectric constant measures how strongly the material screens the field. (It does not measure how good an insulator the material is. That is a different property, its conductivity.)

Equation (9.29) gives a rule of thumb that is right surprisingly often: to find the answer to a problem with a dielectric, take the answer without the dielectric and replace ε_0 by ε. The rule is exact under the conditions just stated, but it is not reliable when the dielectric fills only part of the region where the field is. For a dielectric slab in a uniform applied field $\mathbf{E}_0$, for example, the field inside is

$\mathbf{E}_0/\varepsilon_r$ if the slab is perpendicular to the field, as the rule says, but $\mathbf{E}_0$, unchanged, if it is parallel to it (Section 9.8.2), and for a dielectric sphere it is $3\mathbf{E}_0/(\varepsilon_r+2)$ (Example 9.11). As $\varepsilon_r \to \infty$, the field inside the perpendicular slab and inside the sphere goes to zero, as it does inside a conductor (Figure 9.15). In this sense a conductor behaves, as far as screening is concerned, like a dielectric with an infinite dielectric constant.

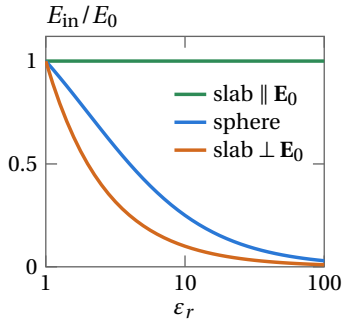


Figure 9.15 The field inside a dielectric placed in a uniform field $\mathbf{E}_0$, as a fraction of E_0. For a slab parallel to the field it is unchanged; for a sphere it is $3/(\varepsilon_r+2)$ (Example 9.11); for a slab perpendicular to the field it is $1/\varepsilon_r$.

9.7 Solving for E when given P: Gauss's law for D method (high symmetry)

Gauss's law for **D** gives a second way to find the field of a polarized object, and when it works it is much faster than the bound-charge method:

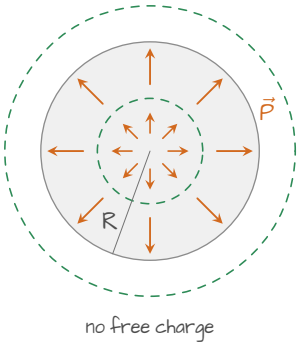
1. Find **D** from Gauss's law for **D**, (9.26), if the symmetry allows it.
2. Then $\mathbf{E} = (\mathbf{D} - \mathbf{P})/\varepsilon_0$, from the definition (9.21).

The field found this way is the total field, from all sources. The method is especially quick when there is no free charge at all, since then **D** comes out to be zero everywhere (provided the symmetry is there to justify that), and the field is simply $\mathbf{E} = -\mathbf{P}/\varepsilon_0$. The method is used even more often when it is the free charge, not the polarization, that is given (Section 9.7.1).

Example 9.7

Find the field of the sphere of Example 9.4, with $\mathbf{P} = (P_0/R)\,\mathbf{r}$, by Gauss's law for **D**.

Solution:



As usual, the first step is a picture of what we are given: the sphere, with **P** pointing outward and growing toward the surface, and no free charge anywhere. The polarization depends only on r and points along $\hat{\mathbf{r}}$, so by symmetry **D** is radial and depends only on r, and the Gaussian surfaces are concentric spheres of radius r, one inside the polarized sphere and one outside it.

There is no free charge anywhere, so

$$\oint \mathbf{D}\cdot d\mathbf{a} = D\cdot 4\pi r^2 = 0, \qquad \mathbf{D} = 0.$$

Then $\mathbf{E} = -\mathbf{P}/\varepsilon_0 = -(P_0/\varepsilon_0)(r/R)\,\hat{\mathbf{r}}$ inside, and $\mathbf{E} = 0$ outside, where $\mathbf{P} = 0$.

Reality check: That is the answer of Example 9.4, in two lines instead of a page.

The same argument does not work for the uniformly polarized sphere. Its polarization $P\hat{\mathbf{z}}$ singles out a direction, so the problem does not have spherical symmetry, $\mathbf{D}$ cannot be taken out of the flux integral, and in fact $\mathbf{D} = \frac{2}{3}\mathbf{P}$ inside, not zero (Section 9.6.1). Before concluding that $\mathbf{D}$ is zero, check that the whole problem, polarization included, has spherical, cylindrical, or planar symmetry.

9.7.1 Linear dielectrics, when the free charge is given

In most practical problems it is the free charge, not the polarization, that is known: the charge on the plates of a capacitor, say, which is filled with a dielectric of known ε_r. Then Gauss's law for $\mathbf{D}$ is the natural first step, and the rest follows in order: $\mathbf{D}$ from the free charge, then $\mathbf{E} = \mathbf{D}/\varepsilon$, then $\mathbf{P} = \varepsilon_0\chi_e\mathbf{E}$ (or $\mathbf{D} - \varepsilon_0\mathbf{E}$), and finally, if they are wanted, the bound charges.

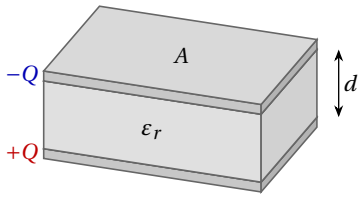
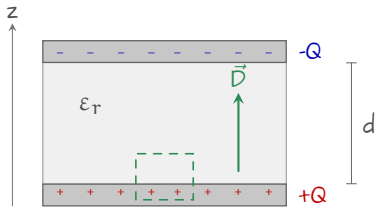


Figure 9.16 The filled capacitor of Example 9.8.

Example 9.8

A parallel-plate capacitor with plates of area A a distance d apart is completely filled with a dielectric of dielectric constant ε_r. The bottom plate carries free charge $+Q$ and the top plate $-Q$. Ignoring the fringing field at the edges, find $\mathbf{D}$, $\mathbf{E}$, the potential difference, the capacitance, $\mathbf{P}$, and the bound charges.

Solution:



We start by sketching the capacitor edge-on, with z upward, the free charge $+Q$ on the bottom plate and $-Q$ on the top one, and the dielectric filling the gap. By planar symmetry $\mathbf{D}$ is vertical and uniform between the plates, so the Gaussian surface is a pillbox with one face inside the bottom plate and the other in the dielectric.

Let the pillbox enclose an area a of the plate. Inside the metal $\mathbf{E} = 0$ and $\mathbf{P} = 0$, so $\mathbf{D} = 0$ there, and no flux passes through the face in the metal or through the sides. The free charge enclosed is $(Q/A)\,a$, so

$$D\,a = \frac{Q}{A}\,a, \qquad \mathbf{D} = \frac{Q}{A}\hat{\mathbf{z}}.$$

(Two mistakes are common here: counting flux through both faces of the pillbox, which gives $Q/2A$, and drawing the surface around both plates, which gives only $0 = 0$.) In the dielectric,

$$\mathbf{E} = \frac{\mathbf{D}}{\varepsilon_0\varepsilon_r} = \frac{Q}{\varepsilon_0\varepsilon_r A}\hat{\mathbf{z}},$$

smaller by the factor ε_r than the field of the empty capacitor. The bottom plate is at the higher potential, and

$$\Delta V = Ed = \frac{Qd}{\varepsilon_0\varepsilon_r A}, \qquad C = \frac{Q}{\Delta V} = \varepsilon_r\frac{\varepsilon_0 A}{d}.$$

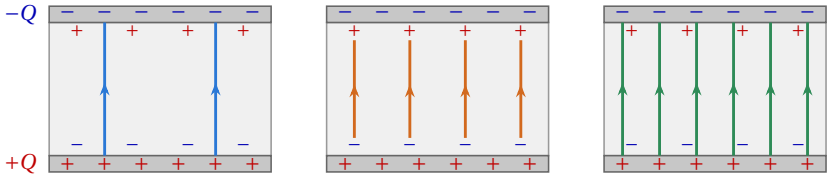
The dielectric multiplies the capacitance by ε_r. The filled capacitor has the same capacitance as an empty one whose plates are a distance d/ε_r apart.

The polarization is

$$\mathbf{P} = \varepsilon_0 \chi_e \mathbf{E} = \frac{\varepsilon_r - 1}{\varepsilon_r} \frac{Q}{A} \hat{\mathbf{z}},$$

upward, in the same direction as $\mathbf{E}$. It is uniform, so $\rho_b = 0$. On the top face of the dielectric, $\hat{\mathbf{n}} = \hat{\mathbf{z}}$ and $\sigma_b = +(1 - 1/\varepsilon_r)\, Q/A$; on the bottom face, $\sigma_b = -(1 - 1/\varepsilon_r)\, Q/A$. Each face carries bound charge of the opposite sign to the free charge on the plate next to it, as the picture of polarization in Section 9.2 suggests.

Reality check: The net charge per area on the bottom, free plus bound, is $Q/A - (1 - 1/\varepsilon_r)Q/A = Q/\varepsilon_r A$, and on the top it is $-Q/\varepsilon_r A$. Two sheets with these charges give the field $Q/\varepsilon_0 \varepsilon_r A$ between them, the field found above.



(a) **E** (b) **P** (c) **D**

Figure 9.17 The filled capacitor of Example 9.8, drawn for $\varepsilon_r = 3$. The large signs on the plates are free charge and the small signs on the faces of the dielectric are bound charge. (a) **E**: the bound charge cancels part of the free charge, and the field between the plates is reduced by ε_r. (b) **P**: only inside the dielectric, pointing along **E**; its lines run from negative to positive bound charge. (c) **D**: its lines begin and end only on free charge, and it is the same as $\varepsilon_0 \mathbf{E}$ would be in the empty capacitor.

Figure 9.17 shows the three fields. Lines of **E** begin and end on any charge, free or bound; lines of **D** begin and end only on free charge, as they must, since $\nabla \cdot \mathbf{D} = \rho_f$; and **P** exists only in the material.

Example 9.9

A spherical capacitor consists of a conducting sphere of radius a with charge $+Q$ and a concentric conducting shell of radius b with charge $-Q$. The space between them is filled with a dielectric of dielectric constant ε_r. Find the capacitance, the polarization, and the bound charges.

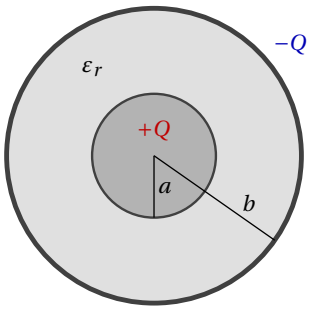
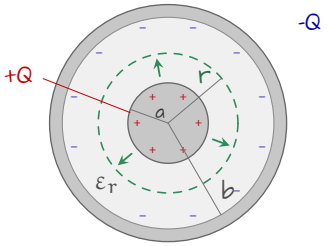


Figure 9.18 The spherical capacitor of Example 9.9, in cross section.

Solution:



The first thing to do is to draw the capacitor in cross section: the inner sphere with $+Q$, the shell with $-Q$, the dielectric between them, and a Gaussian sphere of radius r in the dielectric. By spherical symmetry $\mathbf{D}$ is radial and has the same magnitude everywhere on that sphere, which encloses the free charge Q.

Gauss's law for $\mathbf{D}$ on that sphere gives $D \cdot 4\pi r^2 = Q$, so

$$\mathbf{D} = \frac{Q}{4\pi r^2}\,\hat{\mathbf{r}}, \qquad \mathbf{E} = \frac{Q}{4\pi\varepsilon_0\varepsilon_r r^2}\,\hat{\mathbf{r}} \qquad (a < r < b),$$

and both are zero for $r < a$ and $r > b$. The potential difference is

$$\Delta V = -\int_b^a \mathbf{E}\cdot d\boldsymbol{\ell} = \frac{Q}{4\pi\varepsilon_0\varepsilon_r}\left(\frac{1}{a} - \frac{1}{b}\right), \qquad C = \frac{Q}{\Delta V} = 4\pi\varepsilon_0\varepsilon_r\,\frac{ab}{b-a},$$

ε_r times the capacitance of the empty capacitor. The polarization is

$$\mathbf{P} = \varepsilon_0\chi_e\mathbf{E} = \left(1 - \frac{1}{\varepsilon_r}\right)\frac{Q}{4\pi r^2}\,\hat{\mathbf{r}}.$$

Its divergence is zero for $r \neq 0$, (1.44), so $\rho_b = 0$, in agreement with (9.28), since there is no free charge in the dielectric. On the inner surface $\hat{\mathbf{n}} = -\hat{\mathbf{r}}$ and on the outer surface $\hat{\mathbf{n}} = +\hat{\mathbf{r}}$, so

$$\sigma_b(a) = -\left(1 - \frac{1}{\varepsilon_r}\right)\frac{Q}{4\pi a^2}, \qquad \sigma_b(b) = +\left(1 - \frac{1}{\varepsilon_r}\right)\frac{Q}{4\pi b^2}.$$

Reality check: The total bound charges on the two surfaces are $\mp(1 - 1/\varepsilon_r)Q$, which add to zero. The bound charge on the inner surface cancels part of the free charge on the inner sphere, leaving a net charge of Q/ε_r inside any sphere of radius r between a and b (Figure 9.19), which is why the field there is reduced by ε_r.

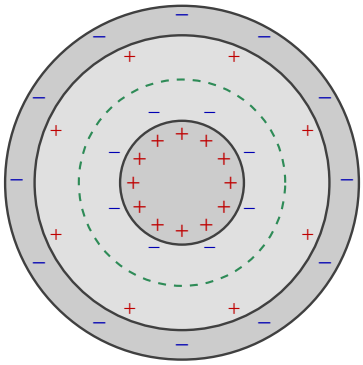


Figure 9.19 The free charge (large signs) and the bound charge (small signs) of Example 9.9, drawn for $\varepsilon_r = 3$: $-(1 - 1/\varepsilon_r)Q$ on the inner surface of the dielectric and $+(1 - 1/\varepsilon_r)Q$ on its outer surface. A sphere between a and b (dashed) encloses the net charge Q/ε_r.

9.8 Boundary conditions at dielectric interfaces

Sections 4.4 and 5.2 found the boundary conditions on the electric field at any surface:

$$E^{\perp}_{\text{above}} - E^{\perp}_{\text{below}} = \frac{\sigma}{\varepsilon_0}, \qquad \mathbf{E}^{\parallel}_{\text{above}} = \mathbf{E}^{\parallel}_{\text{below}}, \qquad V_{\text{above}} = V_{\text{below}}.$$

They still hold at the surface of a dielectric, but the σ in the first one is the total surface charge, free plus bound, and the bound part is usually not known in advance. As with Gauss's law, $\mathbf{D}$ provides a version that involves only the free charge.

9.8.1 The perpendicular component of D

Apply Gauss's law for **D** to a thin pillbox that straddles the surface, as in Section 4.4 and Figure 4.12. As the pillbox is made thinner, the flux through its sides goes to zero, and the flux through its top and bottom faces, of area A, equals the free charge enclosed: $\left(D^{\perp}_{\text{above}} - D^{\perp}_{\text{below}}\right) A = \sigma_f A$, or

$$D^{\perp}_{\text{above}} - D^{\perp}_{\text{below}} = \sigma_f. \tag{9.30}$$

At the boundary between two dielectrics there is usually no free charge, and then the perpendicular component of **D** is continuous.

9.8.2 The parallel components

The parallel component of **E** is continuous, as before. One might guess that, by analogy, $\varepsilon_r \mathbf{E}^{\parallel}$, or $\mathbf{D}^{\parallel}$, is continuous too. It is not, for two reasons. First, the condition on $\mathbf{E}^{\parallel}$ involves no charge at all, so there is nothing in it to replace by free charge. Second, it was derived from $\nabla \times \mathbf{E} = 0$, by taking the line integral of **E** around a thin rectangular loop that straddles the surface (Section 5.2), and **D** has no such property: its curl is $\nabla \times \mathbf{P}$, (9.27), which need not be zero. What the definition of **D** does give is

$$\mathbf{D}^{\parallel}_{\text{above}} - \mathbf{D}^{\parallel}_{\text{below}} = \mathbf{P}^{\parallel}_{\text{above}} - \mathbf{P}^{\parallel}_{\text{below}}, \tag{9.31}$$

since $\varepsilon_0 \mathbf{E}^{\parallel}$ is the same on both sides.

9.8.3 Interfaces between linear dielectrics

If the materials on both sides are linear dielectrics, $\mathbf{D} = \varepsilon \mathbf{E}$ on each side, and (9.30) becomes a condition on **E** that involves only the free charge,

$$\varepsilon_{\text{above}} E^{\perp}_{\text{above}} - \varepsilon_{\text{below}} E^{\perp}_{\text{below}} = \sigma_f. \tag{9.32}$$

In terms of the potential, with $E^{\perp} = -\partial V/\partial n$ and n the coordinate along the normal, pointing from below to above,

$$\varepsilon_{\text{above}} \frac{\partial V_{\text{above}}}{\partial n} - \varepsilon_{\text{below}} \frac{\partial V_{\text{below}}}{\partial n} = -\sigma_f. \tag{9.33}$$

This is the dielectric version of (7.2). With no free charge at the interface, $\varepsilon\, \partial V/\partial n$ is continuous; that is the condition used in boundary-value problems and in numerical work.

Because the parallel component of **E** is continuous and the perpendicular component jumps, field lines change direction at an interface. Let θ_1 and θ_2 be the angles that **E** makes with the normal on the two sides of an interface with no free charge. Then $\tan\theta = E^{\parallel}/E^{\perp}$ on each side, $E^{\parallel}$ is the same on both sides, and $\varepsilon_1 E^{\perp}_1 = \varepsilon_2 E^{\perp}_2$, so

$$\frac{\tan\theta_2}{\tan\theta_1} = \frac{\varepsilon_2}{\varepsilon_1}. \tag{9.34}$$

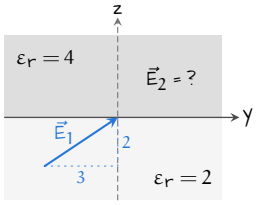
☞ *Caution*: (9.34) looks like Snell's law for the refraction of light, but these are field lines, not light rays, and the rule is different.

In the material with the larger dielectric constant, the field lines are farther from the normal. The same reasoning explains the slab orientations of Section 9.6.3. A slab perpendicular to a uniform applied field $\mathbf{E}_0$ has $D^\perp$ continuous at its faces, so $D = \varepsilon_0 E_0$ inside and $E = E_0/\varepsilon_r$. A slab parallel to the field has $\mathbf{E}^\parallel$ continuous at its faces, so the field inside is $\mathbf{E}_0$.

Example 9.10

The plane $z = 0$ separates a dielectric with $\varepsilon_r = 2$ below from one with $\varepsilon_r = 4$ above. There is no free charge on the interface. Just below it, the field is $\mathbf{E}_1 = (3\hat{\mathbf{y}} + 2\hat{\mathbf{z}})\,\mathrm{V/m}$. Find the field $\mathbf{E}_2$ just above it, the angles the field makes with the normal, and the bound charge on the interface.

Solution:



First a sketch, with the interface edge-on: the given field $\mathbf{E}_1$ just below it, split into its components parallel and perpendicular to the interface, and a question mark where $\mathbf{E}_2$ goes. The boundary conditions treat the two components differently.

The parallel component is continuous, so $E_{2y} = 3\,\mathrm{V/m}$. The perpendicular components satisfy $2\,E_{1z} = 4\,E_{2z}$, so $E_{2z} = 1\,\mathrm{V/m}$, and

$$\mathbf{E}_2 = (3\hat{\mathbf{y}} + 1\hat{\mathbf{z}})\,\mathrm{V/m}.$$

The angles from the normal are $\theta_1 = \arctan(3/2) = 56.3°$ and $\theta_2 = \arctan 3 = 71.6°$ (Figure 9.20), and $\tan\theta_2/\tan\theta_1 = 2 = \varepsilon_2/\varepsilon_1$, as (9.34) says. The jump in $E^\perp$ gives the total surface charge, which here is all bound:

$$\begin{aligned}\sigma_b &= \varepsilon_0\,(E_{2z} - E_{1z}) \\ &= -\varepsilon_0\,(1\,\mathrm{V/m}) \\ &= -8.85\times10^{-12}\,\mathrm{C/m^2}.\end{aligned}$$

Reality check: The polarization gives the same answer. On each side $\mathbf{P} = \varepsilon_0(\varepsilon_r - 1)\mathbf{E}$, so $\mathbf{P}_1 = \varepsilon_0(3\hat{\mathbf{y}} + 2\hat{\mathbf{z}})$ and $\mathbf{P}_2 = \varepsilon_0(9\hat{\mathbf{y}} + 3\hat{\mathbf{z}})$, in units of V/m. The outward normal of the lower material is $+\hat{\mathbf{z}}$ and that of the upper material is $-\hat{\mathbf{z}}$, so the two bound surface charges add to $\sigma_b = P_{1z} - P_{2z} = -\varepsilon_0\,(1\,\mathrm{V/m})$. The parallel component of $\mathbf{D}$, on the other hand, is $D_{1y} = 6\varepsilon_0$ and $D_{2y} = 12\varepsilon_0$: it jumps, by the jump in P_y, as (9.31) says.

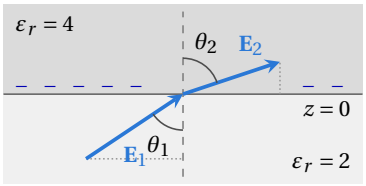


Figure 9.20 The field of Example 9.10 at an interface between $\varepsilon_r = 2$ (below) and $\varepsilon_r = 4$ (above). The parallel component is the same on both sides; the perpendicular component is halved above.

9.8.4 Boundary-value problems with dielectrics

Inside a homogeneous linear dielectric with no free charge, there is no bound volume charge either, (9.28), so the potential satisfies Laplace's equation, just as in empty space. (Where there is free charge, $\nabla \cdot \mathbf{D} = \rho_f$ with $\mathbf{D} = -\varepsilon \nabla V$ gives $\nabla^2 V = -\rho_f/\varepsilon$ in each homogeneous region, and more generally, where ε varies from place to place,

$$\nabla \cdot (\varepsilon \nabla V) = -\rho_f, \tag{9.35}$$

the version of Poisson's equation that holds in a linear dielectric.) The regions are tied together at their boundaries by the continuity of V and by (9.33). So the methods of Chapter 7 apply, with one new boundary condition.

Example 9.11

A sphere of radius R and dielectric constant ε_r is placed in an otherwise uniform field $\mathbf{E} = E_0 \hat{\mathbf{z}}$. Find the potential and the field inside and outside.

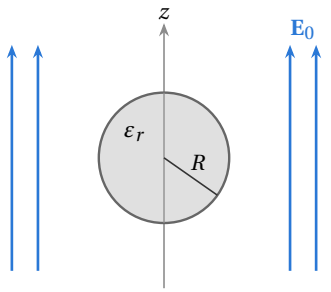
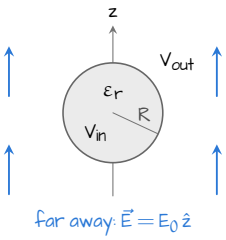


Figure 9.21 The dielectric sphere of Example 9.11. The applied field is drawn only far from the sphere, where it is known.

Solution:



We sketch the sphere in its field, which is uniform far away, and label the two regions, inside and outside. This is the problem of Example 7.4 with a dielectric sphere in place of the conductor. The sphere is no longer an equipotential, so the condition $V = 0$ on its surface is replaced by the conditions that join the inside to the outside.

Inside and outside there is no free charge, so V satisfies Laplace's equation, and the boundary conditions are

$$\begin{aligned}
&\text{BC1:} && V_{\text{in}} \text{ is finite at } r = 0,\\
&\text{BC2:} && V_{\text{in}}(R,\theta) = V_{\text{out}}(R,\theta),\\
&\text{BC3:} && \varepsilon_r \frac{\partial V_{\text{in}}}{\partial r} = \frac{\partial V_{\text{out}}}{\partial r} \text{ at } r = R,\\
&\text{BC4:} && V_{\text{out}} \to -E_0 r\cos\theta \text{ as } r \to \infty.
\end{aligned}$$

BC3 is (9.33) with $\sigma_f = 0$; outside the sphere the dielectric constant is 1. As in Example 7.4, BC4 by the matching coefficients trick leaves only the $\ell = 1$ term of the r^ℓ series outside, while all of the $1/r^{\ell+1}$ terms are kept, and BC1 removes the $1/r^{\ell+1}$ terms inside:

$$V_{\text{in}} = \sum_\ell A_\ell r^\ell P_\ell(\cos\theta), \qquad V_{\text{out}} = -E_0 r\cos\theta + \sum_\ell \frac{B_\ell}{r^{\ell+1}} P_\ell(\cos\theta).$$

Matching the coefficient of each P_ℓ in BC2 gives, for $\ell \neq 1$,

$$A_\ell R^\ell = \frac{B_\ell}{R^{\ell+1}},$$

which by itself does not make A_ℓ and B_ℓ zero. BC3 gives, for $\ell \neq 1$,

$$\varepsilon_r \ell A_\ell R^{\ell-1} = -(\ell+1)\frac{B_\ell}{R^{\ell+2}}.$$

Using the first to replace B_ℓ in the second gives $(\varepsilon_r \ell + \ell + 1)A_\ell R^{\ell-1} = 0$, so $A_\ell = 0$ and then $B_\ell = 0$: both conditions are needed to remove the terms with $\ell \neq 1$. For $\ell = 1$, BC2 and BC3 give

$$A_1 R = -E_0 R + \frac{B_1}{R^2}, \qquad \varepsilon_r A_1 = -E_0 - \frac{2B_1}{R^3},$$

two equations in two unknowns. Their solution is

$$A_1 = -\frac{3}{\varepsilon_r + 2}E_0, \qquad B_1 = \frac{\varepsilon_r - 1}{\varepsilon_r + 2}E_0 R^3.$$

So

$$V_{\text{in}} = -\frac{3E_0}{\varepsilon_r + 2} r\cos\theta, \qquad V_{\text{out}} = -E_0 r\cos\theta + \frac{\varepsilon_r - 1}{\varepsilon_r + 2}\frac{E_0 R^3 \cos\theta}{r^2}. \tag{9.36}$$

Inside, V is proportional to $z = r\cos\theta$, and the field is uniform:

$$\mathbf{E}_{\text{in}} = \frac{3}{\varepsilon_r + 2}\mathbf{E}_0. \tag{9.37}$$

Outside, the field is the applied field plus the field of a dipole at the center with moment $\mathbf{p} = 4\pi\varepsilon_0 R^3 [(\varepsilon_r - 1)/(\varepsilon_r + 2)]\,\mathbf{E}_0$.

Reality check: For $\varepsilon_r \to 1$ the sphere is not there, and $\mathbf{E}_{\text{in}} \to \mathbf{E}_0$, $B_1 \to 0$. For $\varepsilon_r \to \infty$, $\mathbf{E}_{\text{in}} \to 0$ and $B_1 \to E_0 R^3$, the conducting sphere of Example 7.4, (7.44): once again, a conductor acts like a dielectric with an infinite dielectric constant.

The field inside the dielectric sphere points along the applied field and is smaller than it, by 1/2 for $\varepsilon_r = 4$. The polarization is uniform,

$$\mathbf{P} = \varepsilon_0 \chi_e \mathbf{E}_{\text{in}} = 3\varepsilon_0 \frac{\varepsilon_r - 1}{\varepsilon_r + 2}\mathbf{E}_0,$$

with bound charge $\sigma_b = P\cos\theta$ on the surface. That connects this example to the uniformly polarized sphere of Example 9.3, and it gives a check. Inside, the bound charge produces $-\mathbf{P}/3\varepsilon_0 = -[(\varepsilon_r - 1)/(\varepsilon_r + 2)]\,\mathbf{E}_0$, and adding the applied field gives $3\mathbf{E}_0/(\varepsilon_r + 2)$, in agreement with (9.37) (Figure 9.22). Outside, the bound charge produces the field of a dipole with moment $\frac{4}{3}\pi R^3 \mathbf{P}$, which is the dipole found above. The same reasoning can be run forward, as a shortcut: *assume* that the field inside is uniform. Then $\mathbf{P}$ is uniform, its bound charge produces $-\mathbf{P}/3\varepsilon_0 = -\chi_e \mathbf{E}_{\text{in}}/3$ inside, and $\mathbf{E}_{\text{in}} = \mathbf{E}_0 - \chi_e \mathbf{E}_{\text{in}}/3$, which gives (9.37) again. The assumption is justified afterward, because the potential it leads to satisfies Laplace's equation and every boundary condition, and the solution of a problem like this is unique (Exercise P9.25).

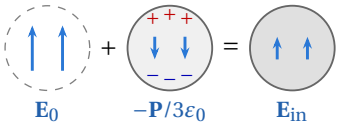


Figure 9.22 Inside the dielectric sphere of Example 9.11, drawn for $\varepsilon_r = 4$: the applied field $\mathbf{E}_0$ and the field $-\mathbf{P}/3\varepsilon_0$ of the bound charge $\sigma_b = P\cos\theta$ add to the uniform field $\mathbf{E}_{\text{in}} = 3\mathbf{E}_0/(\varepsilon_r + 2)$, here $\mathbf{E}_0/2$. Animation of the sphere's field lines as ε_r goes from 1 to 1000.

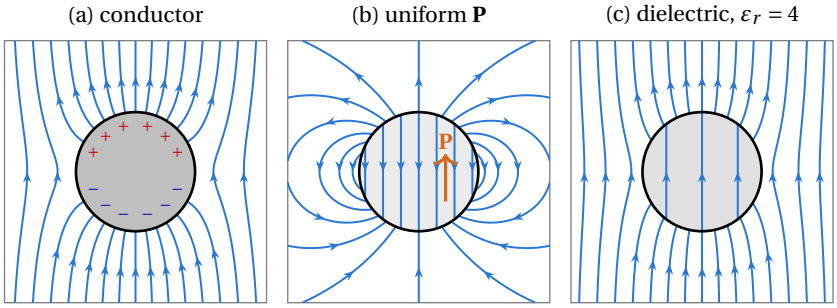


Figure 9.23 Field lines of **E** for three spheres. (a) A conducting sphere in a uniform field (Example 7.4). The lines meet the surface at right angles, and there is no field inside. (b) A uniformly polarized sphere, with no applied field (Example 9.3). The lines begin on the positive bound charge of the upper hemisphere and end on the negative bound charge of the lower one; inside, the field is uniform and points opposite to **P**. (c) A dielectric sphere with $\varepsilon_r = 4$ in a uniform field (Example 9.11). The lines meet the surface at an angle; some of them end on the bound charge, and the field inside is uniform and half the applied field. Where the field is uniform, far from the spheres and inside them, the spacing of the lines is inversely proportional to the strength of the field; elsewhere it shows the strength only roughly.

Figure 9.23 compares the dielectric sphere with the conducting sphere and the uniformly polarized sphere. Outside, the conducting and the dielectric spheres produce fields of the same form, a uniform field plus a dipole, but the conductor bends the field lines so that they meet its surface at right angles, while at the dielectric's surface they obey (9.34). Inside, the conductor has no field and the dielectric has a uniform, reduced one. The dielectric sphere is itself a uniformly polarized sphere, of the kind in panel (b), placed in the applied field; its total field is the sum of the fields in (b) and the applied field.

9.9 Numerical methods: direct integration and relaxation

The two numerical methods of earlier chapters carry over to dielectrics, one for each kind of problem. When the polarization is given, the bound charges can be summed numerically, as in Sections 3.6 and 5.6. When the free charges or the potentials on conductors are given and the dielectric is linear, the relaxation method of Section 7.3 can be adapted to a permittivity that varies from place to place.

9.9.1 Direct integration over the bound charges

The cube of Example 9.6 is a good illustration. Its bound charge is $\pm P$ on two opposite faces, and the field at any point is the sum of the fields of two uniformly charged squares, which a program can compute by dividing each square into small patches, exactly as in Section 3.6. For a cube of side $2a$ centered on the origin, with $\mathbf{P} = P\hat{\mathbf{z}}$, the results at a few points inside are listed in Table 9.3. The

Point	$\varepsilon_0 E_z / P$
$(0,0,0)$	-0.333
$(0,0,a/2)$	-0.395
$(0,0,0.9a)$	-0.525
$(a/2,0,0)$	-0.303
$(0.9a,0,0)$	-0.238

Table 9.3 The field inside a uniformly polarized cube of side $2a$, centered on the origin, with $\mathbf{P} = P\hat{\mathbf{z}}$, found by summing the fields of the bound charges on its faces numerically.

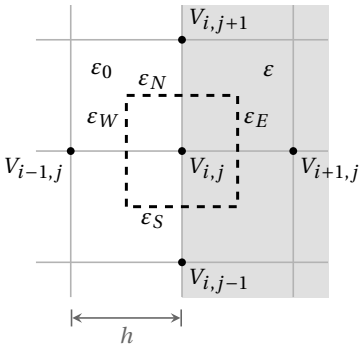


Figure 9.24 A grid point (i, j) and the small square around it, with sides halfway to the four neighboring points. Each grid cell has its own permittivity; the permittivity assigned to a side of the square, such as ε_E, is the average for the two cells that the side passes through. Here the point is on the surface of a dielectric (shaded), so $\varepsilon_E = \varepsilon$, $\varepsilon_W = \varepsilon_0$, and $\varepsilon_N = \varepsilon_S = (\varepsilon + \varepsilon_0)/2$.

value at the center is the exact $-1/3$ found in Example 9.6, which makes a good check on the program. Elsewhere the field is not uniform. It is stronger near the charged faces and weaker near the uncharged ones, and off the axes it also has x and y components. The uniformly polarized sphere, whose field inside is exactly uniform, is special in this respect.[12] Exercise P9.11 asks you to write the program.

9.9.2 Relaxation with a dielectric

The **relaxation method** of Section 7.3 replaces the potential at each grid point by the average of its four neighbors, (7.10). That rule came from Laplace's equation, which holds inside each homogeneous piece of a dielectric but not at an interface between two dielectrics, where there is bound surface charge. The rule that works everywhere comes from Gauss's law for **D** instead.

Draw a small square around the grid point (i, j), with its sides halfway to the four neighboring points (Figure 9.24). Take the problem to be two-dimensional, so that nothing depends on the third coordinate, and consider a unit length in that direction. The flux of **D** out through the right-hand side of the square, which has length h, is approximately $\varepsilon_E (V_{i,j} - V_{i+1,j})/h$ times h, where $(V_{i,j} - V_{i+1,j})/h$ is the field in the x direction between the two points and ε_E is the permittivity on that side. Adding the four sides, and setting the total flux equal to the free charge inside the square, $\rho_f h^2$ per unit length,

$$\varepsilon_E \left(V_{i,j} - V_{i+1,j}\right) + \varepsilon_W \left(V_{i,j} - V_{i-1,j}\right) + \varepsilon_N \left(V_{i,j} - V_{i,j+1}\right) + \varepsilon_S \left(V_{i,j} - V_{i,j-1}\right) = \rho_f h^2.$$

Solving for $V_{i,j}$,

$$V_{i,j} = \frac{\varepsilon_E V_{i+1,j} + \varepsilon_W V_{i-1,j} + \varepsilon_N V_{i,j+1} + \varepsilon_S V_{i,j-1} + \rho_f h^2}{\varepsilon_E + \varepsilon_W + \varepsilon_N + \varepsilon_S}. \tag{9.38}$$

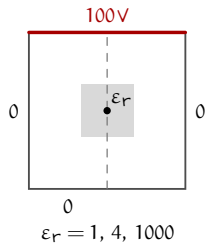
The new value is a *weighted* average of the neighbors, each weighted by the permittivity between it and the center. If the permittivity is the same on all four sides, (9.38) reduces to the ordinary average (7.10), plus the Poisson term of Section 7.3.1 with ε in place of ε_0. A convenient way to set up the weights is to give each small grid cell its own permittivity and to use, for each side of the square, the average for the two cells that the side passes through. Then no grid point needs special treatment, whether it lies inside a dielectric, outside it, on its surface, or at one of its corners. The continuity of $\varepsilon \partial V/\partial n$, (9.33), is built into the rule, since it came from Gauss's law for **D**, and the iteration itself (Jacobi or Gauss–Seidel) is the same as before.

Example 9.12

A square region has $V = 100\,\text{V}$ on its top edge and $V = 0$ on the other three. A square block of dielectric fills the middle third of the region in each direction. Use relaxation to find the potential for $\varepsilon_r = 1$, 4, and 1000.

[12] A uniformly polarized ellipsoid, of which the sphere is a special case, also has a uniform field inside.

Solution:



We start by drawing the region, with the potential on each edge and the block in the middle third. The dashed line is the vertical line through the center, along which we will compare the three cases.

With a grid of 60×60 cells and (9.38), the results along the vertical line through the center are shown in Figure 9.25. For $\varepsilon_r = 1$ the block is not really there, and the potential falls smoothly from the top edge to the bottom one. For $\varepsilon_r = 4$ the potential is flatter inside the block: along the center line it changes by 11.8 V across the block, compared with 28.7 V without it, so the field inside is smaller by a factor of 2.4. For $\varepsilon_r = 1000$ it changes by only 0.06 V, and the block is very nearly an equipotential.

Reality check: In all three cases the potential at the center is 25.000 V. That is not an accident, and it makes a useful check on a program. Imagine four problems like this one, each with 100 V on a different edge. Their sum is the problem with 100 V on all four edges, whose solution is $V = 100$ V everywhere. The block is centered in the square, so the four problems are the same problem rotated, and each contributes the same potential, 25 V, at the center.

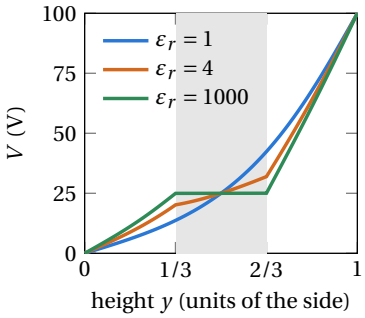


Figure 9.25 The potential along the vertical center line in Example 9.12, for $\varepsilon_r = 1$, 4, and 1000. The shaded band is the dielectric block. The larger ε_r, the flatter the potential inside the block, and the smaller the field there.

The field inside the block is reduced by a factor of 2.4, not by $\varepsilon_r = 4$, because the block fills only part of the region where the field is (Section 9.6.3). For very large ε_r the block behaves like an isolated, uncharged conductor. That gives a simple way to put such a conductor into a relaxation program: represent it as a dielectric with a very large dielectric constant. (A conductor held at a known potential is simpler still: fix the potential at its grid points, as in Section 7.3.2.)

Large dielectric constants slow the iteration down badly, however, and the warning of Section 7.3.3 about the tolerance applies with extra force. The reason is in the weights of (9.38). A grid point on the surface of the block gives its neighbor inside the block ε_r times the weight of its neighbor outside. The potentials inside the block therefore come into line with one another quickly, but the block's potential as a whole is pulled toward its correct value only a little on each pass, and the larger ε_r, the more passes that takes. On a 30×30 grid with $\varepsilon_r = 1000$, the Jacobi method changed the potential by less than 3×10^{-4} V per pass after 20,000 passes, yet the potential at the center was still 6.7 V instead of 25 V. A small change per pass does not mean that the answer is close. Compare against a value you know, such as the 25 V here, or use a faster method: successive over-relaxation, or a direct solution of the linear equations (9.38). Python's `scipy.sparse.linalg.spsolve`, or Mathematica's `LinearSolve` applied to a `SparseArray`, solves them in a fraction of a second for a grid this size.

9.10 Energy stored in the field, inside dielectrics

Section 5.9.2 found that the work needed to assemble a distribution of charge is $U = (\varepsilon_0/2)\int E^2\,d\tau$, (5.23). When dielectrics are present, the question needs to be asked more carefully. What we do when we charge a capacitor filled with a dielectric is bring in free charge; the bound charge appears by itself, as the material polarizes, and polarizing the material takes energy too: the energy of stretching its molecules, or of lining up polar molecules against their thermal motion.

So consider the work needed to bring in the free charge, a little at a time, with the dielectric in place. Adding a small amount of free charge $\delta\rho_f$ at a point where the potential is V takes the work $V\,\delta\rho_f\,d\tau$, by (5.12), so

$$\delta U = \int V\,\delta\rho_f\,d\tau.$$

By Gauss's law for $\mathbf{D}$, $\delta\rho_f = \nabla\cdot(\delta\mathbf{D})$, where $\delta\mathbf{D}$ is the change in $\mathbf{D}$. Then, by the product rule and the divergence theorem, as in Section 5.9.2,

$$\begin{aligned}\delta U &= \int V\,\nabla\cdot(\delta\mathbf{D})\,d\tau \\ &= \oint V\,\delta\mathbf{D}\cdot d\mathbf{a} - \int (\nabla V)\cdot\delta\mathbf{D}\,d\tau \\ &= \int \mathbf{E}\cdot\delta\mathbf{D}\,d\tau,\end{aligned}$$

where the surface integral has been taken over a sphere so large that it vanishes. If the dielectric is linear, $\mathbf{D} = \varepsilon\mathbf{E}$, and

$$\begin{aligned}\mathbf{E}\cdot\delta\mathbf{D} &= \varepsilon\,\mathbf{E}\cdot\delta\mathbf{E} \\ &= \frac{1}{2}\,\delta(\varepsilon E^2) \\ &= \frac{1}{2}\,\delta(\mathbf{D}\cdot\mathbf{E}).\end{aligned}$$

Adding up the steps from no charge at all to the final charge,

$$U = \frac{1}{2}\int \mathbf{D}\cdot\mathbf{E}\,d\tau. \tag{9.39}$$

In vacuum, $\mathbf{D} = \varepsilon_0\mathbf{E}$ and this is (5.23) again. The derivation used $\mathbf{D} = \varepsilon\mathbf{E}$, so (9.39) holds only for linear dielectrics, not, for example, for a ferroelectric. Written in terms of the charge instead of the field, it is $U = \frac{1}{2}\int \rho_f V\,d\tau$, the form (5.22) with only the free charge.

For the filled capacitor of Example 9.8, $\mathbf{D}$ and $\mathbf{E}$ are uniform between the

plates, and

$$\begin{aligned} U &= \frac{1}{2}\frac{Q}{A}\frac{Q}{\varepsilon_0\varepsilon_r A}Ad \\ &= \frac{Q^2 d}{2\varepsilon_0\varepsilon_r A} \\ &= \frac{Q^2}{2C}, \end{aligned}$$

with $C = \varepsilon_r\varepsilon_0 A/d$, in agreement with (5.25).

Why is the answer not simply $(\varepsilon_0/2)\int E^2\,d\tau$? That integral is the work needed to assemble all the charges, free and bound, in their final positions, treating the bound charges as if they were charges like any others, brought in from infinity. It leaves out the energy that goes into polarizing the molecules themselves. For a linear dielectric the difference is

$$\frac{1}{2}\int \mathbf{D}\cdot\mathbf{E}\,d\tau - \frac{\varepsilon_0}{2}\int E^2\,d\tau = \frac{1}{2}\int \mathbf{P}\cdot\mathbf{E}\,d\tau, \tag{9.40}$$

which is positive, since **P** and **E** are parallel (Figure 9.26). The energy we actually have to supply, by charging the capacitor, is (9.39).

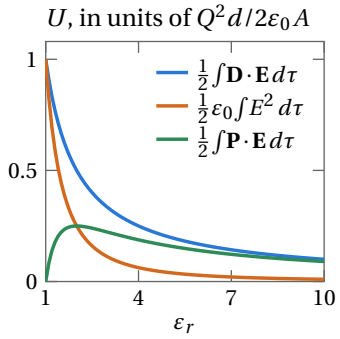


Figure 9.26 The energies of the filled capacitor of Example 9.8 with a fixed charge Q, in units of the energy $Q^2d/2\varepsilon_0A$ it has when empty: the work needed to charge it, $\frac{1}{2}\int \mathbf{D}\cdot\mathbf{E}\,d\tau = Q^2/2C$; the field energy $\frac{1}{2}\varepsilon_0\int E^2\,d\tau$; and their difference, $\frac{1}{2}\int \mathbf{P}\cdot\mathbf{E}\,d\tau$, the energy that goes into polarizing the dielectric.

9.10.1 Forces on dielectrics

A dielectric is pulled into the region of strong field between the plates of a charged capacitor. The force can be found from the energy. Consider plates of length ℓ and width w, a distance d apart, with a slab of dielectric of thickness d and width w that fills the gap over a length $\ell - x$ and leaves a length x empty (Figure 9.27). Ignoring the fringing field, the two parts act like two capacitors in parallel, so their capacitances add:

$$C(x) = \frac{\varepsilon_0 wx}{d} + \frac{\varepsilon_r\varepsilon_0 w(\ell - x)}{d} = \frac{\varepsilon_0 w}{d}\left[x + \varepsilon_r(\ell - x)\right].$$

Suppose first that the capacitor is isolated, so its charge Q is fixed. If the slab moves a distance dx, the force on it does work $F\,dx$, and that work comes out of the stored energy, $F\,dx = -dU$. With $U = Q^2/2C$,

$$\begin{aligned} F &= -\frac{dU}{dx} \\ &= \frac{Q^2}{2C^2}\frac{dC}{dx} \\ &= -\frac{Q^2 d}{2\varepsilon_0 w}\frac{\varepsilon_r - 1}{\left[x + \varepsilon_r(\ell - x)\right]^2}. \end{aligned}$$

In terms of the potential difference $\Delta V = Q/C$, this is

$$F = -\frac{\varepsilon_0\chi_e w}{2d}\Delta V^2. \tag{9.41}$$

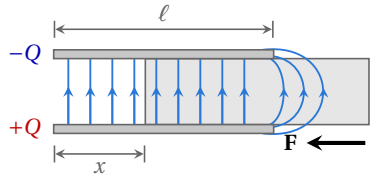


Figure 9.27 A dielectric slab partly inside a parallel-plate capacitor. Near the edge of the plates the field fringes outward. The fringing field acts on the polarized slab and pulls it in, to the left.

The force is negative, toward smaller x: the slab is drawn into the capacitor.

Suppose instead that the capacitor stays connected to a battery, so that ΔV is fixed. Then the stored energy $U = \frac{1}{2}C\,\Delta V^2$ *increases* as the slab moves in, and it is tempting to take $F = -dU/dx$ again, which gives the right magnitude with the wrong sign. The mistake is to forget the battery. As C increases by dC, the battery moves a charge $dQ = \Delta V\,dC$ through the potential difference ΔV and does the work $\Delta V^2\,dC$, which is twice the increase $\frac{1}{2}\Delta V^2\,dC$ in the stored energy. The other half does the mechanical work, $F\,dx = \frac{1}{2}\Delta V^2\,dC$, so

$$F = \frac{1}{2}\,\Delta V^2\,\frac{dC}{dx} = -\frac{\varepsilon_0 \chi_e w}{2d}\,\Delta V^2,$$

the same as (9.41). The force on the slab depends only on the charges and fields that are present at that moment, not on whether a battery is attached. So even when a battery is attached, the simplest way to find the force is to pretend that the charge is held fixed.

The energy method avoids a hard calculation, but it hides where the force comes from. In the uniform field between the plates the field is perpendicular to the faces of the slab, and it pulls up on the bound charge on one face and down on the bound charge on the other; nothing pushes sideways. The sideways force comes from the fringing field near the edge of the plates. There the field lines bulge outward, the field has a component along the slab, and it is nonuniform. Each induced dipole in the slab near the edge is pulled toward the stronger field between the plates, as in Section 9.1.3. The energy method gives the total of all those forces without our having to find the fringing field.

9.11 Summary

A dielectric in a field polarizes: its atoms and molecules become dipoles, by having their charges shifted (induced dipoles, $\mathbf{p} = \alpha\mathbf{E}$) or by being turned (polar molecules), or the polarization is built in (ferroelectrics and electrets). The polarization $\mathbf{P}$, the dipole moment per unit volume, produces exactly the field of the bound charges $\sigma_b = \mathbf{P}\cdot\hat{\mathbf{n}}$ and $\rho_b = -\nabla\cdot\mathbf{P}$, and given $\mathbf{P}$ the field can be found from them by any of the methods of earlier chapters. The displacement $\mathbf{D} = \varepsilon_0\mathbf{E} + \mathbf{P}$ has the free charge as its only divergence, so with enough symmetry Gauss's law for $\mathbf{D}$ gives $\mathbf{D}$ directly from the free charge; but $\mathbf{D}$ also has a curl wherever $\mathbf{P}$ does, so it is not determined by the free charge alone. Some equations hold in any material, and others only in linear dielectrics. (The magnetic counterparts of most of them appear in Chapter 13, and Chapter 14 sets the two side by side.)

In any material

$\boldsymbol{\tau} = \mathbf{p} \times \mathbf{E}, \quad U = -\mathbf{p} \cdot \mathbf{E}$	torque and energy of a dipole of fixed moment in a field
$\mathbf{F} = (\mathbf{p} \cdot \nabla)\mathbf{E} = \nabla(\mathbf{p} \cdot \mathbf{E})$	force on a dipole; zero in a uniform field
$\mathbf{P} = N\mathbf{p}, \quad \mathbf{p}_{\text{total}} = \int \mathbf{P}\, d\tau$	polarization: dipole moment per unit volume
$\sigma_b = \mathbf{P} \cdot \hat{\mathbf{n}}, \quad \rho_b = -\nabla \cdot \mathbf{P}$	the bound charges, whose field is the field of the polarized object
$\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P}$	the definition of the displacement
$\nabla \cdot \mathbf{D} = \rho_f, \quad \oint \mathbf{D} \cdot d\mathbf{a} = q_{f,\text{enc}}$	Gauss's law for **D**; useful when symmetry allows
$\nabla \times \mathbf{D} = \nabla \times \mathbf{P}$	why **D** is not determined by the free charge alone
$D^{\perp}_{\text{above}} - D^{\perp}_{\text{below}} = \sigma_f,$ $\mathbf{E}^{\parallel}_{\text{above}} = \mathbf{E}^{\parallel}_{\text{below}}$	boundary conditions at an interface

In linear dielectrics only

$\mathbf{P} = \varepsilon_0 \chi_e \mathbf{E}$	**E** is the total field in the material
$\mathbf{D} = \varepsilon \mathbf{E}, \quad \varepsilon = \varepsilon_0 \varepsilon_r, \quad \varepsilon_r = 1 + \chi_e$	permittivity and dielectric constant
$\rho_b = -\dfrac{\chi_e}{1+\chi_e}\rho_f$	in a homogeneous dielectric: if no free charge inside the material ($\rho_f = 0$), then all the bound charge is on the surface ($\rho_b = 0$)
$\mathbf{E} = \mathbf{E}_{\text{vac}}/\varepsilon_r$	when the dielectric fills all the space where the field is, with the same free charge
$\varepsilon_{\text{above}} \dfrac{\partial V_{\text{above}}}{\partial n} - \varepsilon_{\text{below}} \dfrac{\partial V_{\text{below}}}{\partial n} = -\sigma_f$	the boundary condition for boundary-value problems and relaxation
$U = \dfrac{1}{2}\int \mathbf{D} \cdot \mathbf{E}\, d\tau$	energy stored, the work done in bringing in the free charge
$\alpha = \dfrac{3\varepsilon_0}{N}\dfrac{\varepsilon_r - 1}{\varepsilon_r + 2}$	the Clausius–Mossotti relation (Tutorial 9.1); a good approximation for nonpolar molecules

Which method? Section 14.3 gathers the methods of electrostatics into one chart.

Exercises

Exercises for 9.1 Dipoles again

Warm-up questions Read Section 9.1, then answer before class

W9.1 A dipole in a uniform field feels no torque when $\mathbf{p}$ points along $\mathbf{E}$, and none when it points opposite to $\mathbf{E}$. Tilt each one slightly. Does the torque turn it back, or farther away? Which of the two orientations is stable, and how does (9.2) show the same thing?

W9.2 Inside a thick slab with uniform charge density $\rho > 0$, filling $-a < x < a$, Gauss's law gives the field $\mathbf{E} = (\rho x/\varepsilon_0)\hat{\mathbf{x}}$. A small dipole inside the slab points along $+y$. Is there a force on it, although the field is not uniform? Answer with both (9.3) and (9.4). If the dipole points along $+x$ instead, does the force depend on which side of the plane $x = 0$ it is on?

W9.3 Section 9.1.4 describes a competition between the field, which lines up polar molecules, and thermal motion, which keeps knocking them into random orientations. Debye's biography in the margin says that substances made of such molecules have large dielectric constants that fall as the temperature rises. Explain the connection. Would you expect the polarization of argon gas, kept at the same density in the same field, to change with the temperature? Why or why not?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P9.1 A collection of point charges q_i at positions $\mathbf{r}_i$ is placed in a uniform field $\mathbf{E}$.

(a) Show that the net force is $Q\mathbf{E}$, where Q is the total charge, and that the torque about the origin is $\mathbf{p} \times \mathbf{E}$, where $\mathbf{p}$ is the dipole moment of the collection.

(b) Show that for a neutral collection the torque is the same about every point.

(c) What is the torque on a quadrupole, a neutral collection whose dipole moment is zero, in a uniform field? Does a quadrupole feel a torque in a *nonuniform* field? Give an example to support your answer.

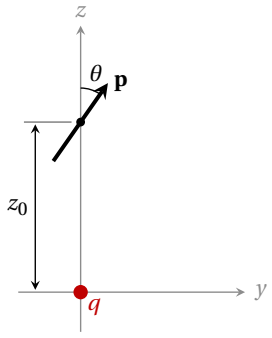


Figure 9.28 The point charge and the dipole of Exercise P9.2, in the yz plane.

P9.2 A point charge $q > 0$ is at the origin (Figure 9.28). A pure dipole of moment p is at the point $(0, 0, z_0)$, with $z_0 > 0$. The dipole lies in the yz plane and makes an angle θ with the $+z$ axis, so $\mathbf{p} = p\sin\theta\,\hat{\mathbf{y}} + p\cos\theta\,\hat{\mathbf{z}}$.

(a) Use $\mathbf{F} = (\mathbf{p}\cdot\nabla)\mathbf{E}$ to find the force on the dipole. Because $\mathbf{p}$ has both y and z components, both $\partial\mathbf{E}/\partial y$ and $\partial\mathbf{E}/\partial z$ contribute. You cannot ignore the y derivative just because the dipole is at $y = 0$: take the derivatives at a general point near the dipole, and *then* set $y = 0$. (Since everything happens in the yz plane, you may set $x = 0$ from the start.)

(b) Find the torque on the dipole about its own center.

(c) The field of the dipole can be written in a form that does not refer to any coordinate system, $\mathbf{E}_{\text{dip}} = \left[3(\mathbf{p}\cdot\hat{\mathbf{r}})\hat{\mathbf{r}} - \mathbf{p}\right]/4\pi\varepsilon_0 r^3$, where $\mathbf{r}$ is measured from the dipole. Use it to find the force that the dipole exerts on the point charge, and check that it is equal and opposite to the force in part (a).

(d) Show that the total torque on the dipole about the *origin*, which is the torque from part (b) plus $z_0\hat{\mathbf{z}}\times\mathbf{F}$, is zero. Explain why it must be.

(e) Describe the force in words for $\theta = 0$ and for $\theta = 90°$.

P9.3 A neutral atom with polarizability α is a distance s from a long, straight wire with a uniform charge λ per unit length (Figure 9.29).

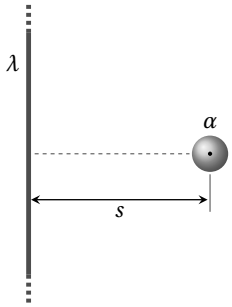


Figure 9.29 The atom and the charged wire of Exercise P9.3.

(a) Find the induced dipole moment of the atom.

(b) Use $\mathbf{F} = (\mathbf{p}\cdot\nabla)\mathbf{E}$ to find the force on the atom. Is it attractive or repulsive? Does the answer depend on the sign of λ?

(c) Show that your force is $-\nabla U$ with $U = -\frac{1}{2}\alpha E^2$, and explain the factor of $\frac{1}{2}$, given that a dipole of fixed moment has $U = -\mathbf{p}\cdot\mathbf{E}$. (Hint: as the atom is brought in, its dipole moment grows in proportion to the field.)

(d) An argon atom, $\alpha/4\pi\varepsilon_0 = 1.64\times10^{-30}\,\text{m}^3$ (Section 9.1.5), is $1.0\,\mu$m from the wire, and λ is chosen so that the field at the atom is $1.0\times10^6\,\text{V/m}$. Find λ, the dipole moment, and the force, and compare the force with the atom's weight.

P9.4 A molecule with polarizability α is at the point $(\ell, 2\ell, 3\ell)$ in the field

$$\mathbf{E} = E_0\left[5\frac{x}{\ell}\hat{\mathbf{x}} + 2\left(\frac{z}{\ell}\right)^3\hat{\mathbf{z}}\right],$$

where E_0 and ℓ are constants.

(a) Check that this field has zero curl, so that it can be an electrostatic field, and find the charge density that produces it.

(b) Find the induced dipole moment of the molecule, and the force and the torque on it.

(c) Show that the force is equal to $\nabla\left(\frac{1}{2}\alpha E^2\right)$ at the molecule.

P9.5 An isolated water molecule has the dipole moment $6.19\times10^{-30}\,\mathrm{C\,m}$.

(a) How strong a field would be needed for the energy pE to equal the thermal energy kT at 20°C? Compare it with the field at which air breaks down, about $3\times10^{6}\,\mathrm{V/m}$.

(b) The static dielectric constant of water at 20°C is 80.1, while its index of refraction for visible light is 1.333. Compare ε_r with n^2, and explain the difference. (The field of visible light oscillates about 5×10^{14} times per second.)

(c) Use the Clausius–Mossotti relation of Tutorial 9.1, with $\varepsilon_r = n^2$, to estimate the part of the polarizability α of a water molecule that comes from shifting its electrons, as $\alpha/4\pi\varepsilon_0$. Take the density of water to be $1.00\,\mathrm{g/cm^3}$. Compare your answer with the polarizabilities listed in Section 9.1.5.

Exercises for 9.2 Dielectric materials: polarization and susceptibility

Warm-up questions Read Section 9.2, then answer before class

W9.4 Section 9.2.2 says that the field outside a polarized electret bar looks like that of a bar magnet until "stray charge from the air collects on its faces and cancels most of that field." Use the layers of uncanceled dipole ends described at the start of this section to explain why that charge collects, and which sign goes to which face. Why can a bar magnet not lose its field in this way?

W9.5 The relation $\mathbf{P} = \varepsilon_0\chi_e\mathbf{E}$ assumes a material that is linear, isotropic, and homogeneous. Which assumption fails in each of these, and what goes wrong with the relation: (a) a crystal that is easier to polarize along one axis than across it; (b) a plastic whose composition changes gradually from one face to the other; (c) glass in the intense field of a focused laser pulse?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

Exercises for 9.3 Bound charges

Warm-up questions Read Section 9.3, then answer before class

W9.6 A classmate says that bound charge is only a mathematical device, like an image charge (Section 7.2): the real things are the molecules' dipoles. Is that right? Where, physically, is the bound charge on the faces of a uniformly polarized slab, and what is it made of?

W9.7 In this chapter, free charge need not be free to move. What does "free" mean here? Classify each of these as free or bound charge, with a reason: (a) the extra electrons on a negatively charged metal sphere; (b) the charge left on a plastic comb that has been run through dry hair; (c) the layer of uncanceled dipole ends at the face of a polarized glass plate; (d) the charged impurity atoms, fixed in place in the crystal, near the junction of a silicon diode.

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P9.6 An object occupies a volume $\mathcal{V}$ with surface $\mathcal{S}$ and has polarization $\mathbf{P}(\mathbf{r})$.

(a) Show that its total bound charge, $\int_{\mathcal{V}} \rho_b \, d\tau + \oint_{\mathcal{S}} \sigma_b \, da$, is zero.

(b) Show that the dipole moment of its bound charges is equal to the integral of its polarization:

$$\int_{\mathcal{V}} \mathbf{r}\rho_b \, d\tau + \oint_{\mathcal{S}} \mathbf{r}\sigma_b \, da = \int_{\mathcal{V}} \mathbf{P} \, d\tau.$$

(Work with the x component, and use the product rule $\nabla \cdot (x\mathbf{P}) = x\nabla \cdot \mathbf{P} + P_x$ together with the divergence theorem.)

(c) Check part (b) for the uniformly polarized sphere of Example 9.3.

P9.7 A cylindrical rod of length L and cross-sectional area A lies along the x axis from $x = 0$ to $x = L$ (Figure 9.30). It has the polarization $\mathbf{P} = (P_0 + P_1 x^2/L^2)\,\hat{\mathbf{x}}$, where P_0 and P_1 are constants.

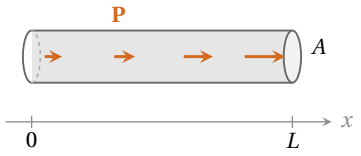


Figure 9.30 The rod of Exercise P9.7. The arrows show its polarization, drawn for $P_1 = 2P_0$.

(a) Find ρ_b inside the rod, and σ_b on its two ends and on its wrapper.

(b) Show explicitly that the total bound charge is zero.

(c) Find the dipole moment of the rod, once from $\int \mathbf{P} \, d\tau$ and once from the bound charges.

P9.8 Gallium nitride (GaN), the semiconductor in blue light-emitting diodes, has a built-in polarization along one of its crystal axes, and so does the alloy $Al_xGa_{1-x}N$, in which a fraction x of the gallium atoms is replaced by aluminum. Take z along the direction in which such crystals are

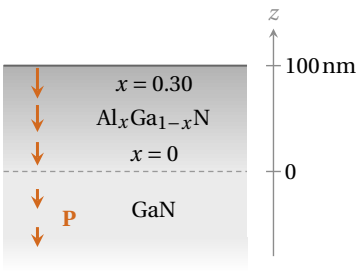


Figure 9.31 The film of Exercise P9.8, edge on, on the GaN it was grown on. The shading darkens as the aluminum fraction x rises, and the arrows show the built-in polarization. The layers are not to scale.

usually grown. Then calculations give $P_z = -0.029\,\text{C/m}^2$ for GaN and $P_z = -0.081\,\text{C/m}^2$ for aluminum nitride, AlN.[13] For the alloy, use the linear interpolation $P_z = -(0.029 + 0.052\,x)\,\text{C/m}^2$. Debdeep Jena and his colleagues grew a film on GaN in which x rises linearly from 0 at $z = 0$ to 0.30 at $z = 100\,\text{nm}$, the top surface of the crystal (Figure 9.31).[14]

(a) Find ρ_b in the film. Show that it is uniform, and find its sign. How many elementary charges per cubic centimeter is it? Compare it with the 10^{18} silicon atoms per cubic centimeter in a GaN film that the same group doped in the usual way, with donor atoms.

(b) Find the film's bound charge per unit area, and show that it equals the bound surface charge at an abrupt interface between GaN and $Al_{0.3}Ga_{0.7}N$. Grading the composition spreads that sheet of charge through the film.

(c) The bound charge draws mobile electrons into the film from states at the crystal's surface, and the film then conducts like a semiconductor doped with donor atoms, although no donor atoms were added to it. This is called *polarization doping*. The film held about 9×10^{12} conduction electrons per square centimeter. Because the alloy is grown on GaN, it is also strained, and the strain adds to P_z a piezoelectric part, which a later calculation gives as $-[0.0525\,x - 0.0282\,x(1-x)]\,\text{C/m}^2$.[15] Find the film's total bound charge per unit area. Why do you need only the values of P_z at the bottom and the top of the film? What fraction of the bound charge do the conduction electrons neutralize?

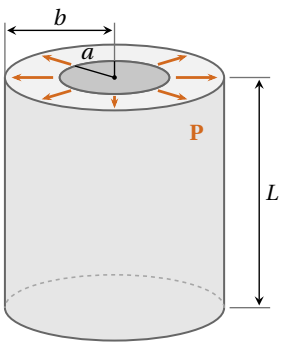


Figure 9.32 The polarized shell of Exercise P9.9.

P9.9 A cylindrical shell of length L, with inner radius a and outer radius b, has the polarization $\mathbf{P} = ks^2\,\hat{\mathbf{s}}$, where k is a constant (Figure 9.32).

(a) Find the bound charge densities on all four surfaces and inside the material.

(b) Find the total bound charge. Explain why it had to come out that way.

(c) If the shell were infinitely long, what would **E** be everywhere? Find it in two ways: from the bound charges with Gauss's law, and with Gauss's law for **D**.

[13] "Spontaneous polarization and piezoelectric constants of III–V nitrides," F. Bernardini, V. Fiorentini, and D. Vanderbilt, Phys. Rev. B 56, R10024–R10027 (1997). https://doi.org/10.1103/PhysRevB.56.R10024

[14] "Realization of wide electron slabs by polarization bulk doping in graded III–V nitride semiconductor alloys," D. Jena et al., Appl. Phys. Lett. 81, 4395–4397 (2002). https://doi.org/10.1063/1.1526161

[15] "Pyroelectric properties of Al(In)GaN/GaN hetero- and quantum well structures," O. Ambacher et al., J. Phys.: Condens. Matter 14, 3399–3434 (2002). https://doi.org/10.1088/0953-8984/14/13/302

(d) For the finite shell there is still no free charge anywhere. Is **D** zero everywhere? (Hint: find the monopole, dipole, and quadrupole moments of the bound charge, using the quadrupole moment Q_{zz} of Section 8.5.)

Exercises for 9.4 Solving for E when given P: bound-charge method

Warm-up questions Read Section 9.4, then answer before class

W9.8 Example 9.5 found that the field at the center of a uniformly polarized cylinder depends on its shape: for a thin wafer, $L \ll R$, it approaches $-\mathbf{P}/\varepsilon_0$, and for a long rod, $L \gg R$, it goes to zero. Explain both limits from where the bound charge sits. Is the field at the center of a long rod polarized *across* its axis also nearly zero? Why or why not?

W9.9 Sketch the field lines of the polarized cylinder of Example 9.5, inside and outside. On the axis just below the top face, does **E** point along **P** or against it? On the axis just above the top face? Explain each from the fields of the two charged faces.

W9.10 In the argument of Section 9.4.1, the field is averaged over a sphere around the field point, a sphere that is small compared with the object but still contains many molecules. What would go wrong with the argument if the sphere were only as large as one molecule? What would go wrong if it were as large as the object?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P9.10 A sphere of radius R is made of a ferroelectric material with the built-in polarization $\mathbf{P} = kr^2\,\hat{\mathbf{r}}$, where k is a constant (Figure 9.33).

(a) What are the units of k? Find σ_b and ρ_b, and show that the total bound charge is zero.

(b) Use Gauss's law with the bound charges to find **E** inside and outside the sphere.

(c) Find **E** again with Gauss's law for **D**.

(d) Inside the sphere, **E** and **P** point in opposite directions. Could that happen in a linear dielectric?

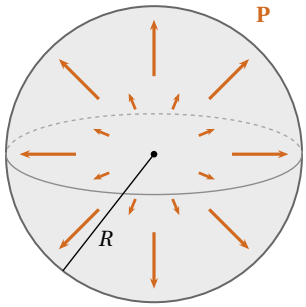


Figure 9.33 The sphere of Exercise P9.10. The arrows show its built-in polarization.

P9.11 A cube of side $2a$, centered on the origin with its faces perpendicular to the axes, has the uniform polarization $\mathbf{P} = P\,\hat{\mathbf{z}}$ (Figure 9.34).

(a) Find the bound charge on all six faces and inside.

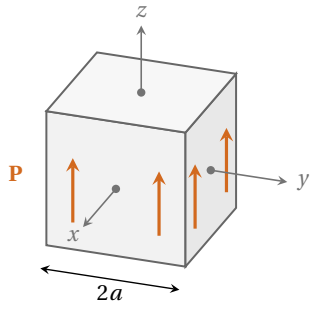


Figure 9.34 The uniformly polarized cube of Exercise P9.11.

(b) Set up the integrals that give the potential $V(x,y,z)$ at any point.

(c) Write a program that finds the field anywhere by dividing the charged faces into small patches and adding their fields. Use it to find **E**, in units of P/ε_0, at the points $(0,0,0)$, $(0,0,a/2)$, $(0,0,0.9a)$, $(a/2,0,0)$, $(0.9a,0,0)$, and $(0.9a,0.9a,0)$. Check your program against the exact value at the center found in Example 9.6.

(d) Find **E** at $(a/2,a/2,a/2)$. What do you notice about E_z there? (For an explanation, consider the cube polarized along x, along y, and along z, and add the three results.)

P9.12 The polarizability of an argon atom is $\alpha/4\pi\varepsilon_0 = 1.641\times10^{-30}\,\mathrm{m}^3$.

(a) For argon gas at 0°C and 1 atmosphere, with $N = 2.687\times10^{25}$ atoms per cubic meter, find χ_e two ways: from $\chi_e = N\alpha/\varepsilon_0$, which ignores the difference between the field at an atom and the average field, and from the Clausius–Mossotti relation of Tutorial 9.1.

(b) Liquid argon at its boiling point, 87.3 K, has a density of $1.40\,\mathrm{g/cm}^3$, and its measured dielectric constant is about 1.50. Find ε_r both ways, and compare with the measured value.

(c) In the liquid, by what factor does the field that polarizes an atom, $\mathbf{E}_{\text{all other}}$, exceed the macroscopic field **E**?

(d) The index of refraction of liquid argon for visible light is about 1.23. Compare n^2 with ε_r, and explain why the two agree for argon but not for water (Exercise P9.5).

P9.13 The Clausius–Mossotti relation of Tutorial 9.1, $\alpha = (3\varepsilon_0/N)(\varepsilon_r-1)/(\varepsilon_r+2)$, connects the polarizability of a single atom to the dielectric constant of a material made of such atoms.

(a) Solve it for ε_r in terms of $x = N\alpha/3\varepsilon_0$.

(b) What happens as x approaches 1?

(c) In Tutorial 9.1, $\frac{4}{3}\pi R^3 = 1/N$. Show that $x = 1$ means $\alpha = 4\pi\varepsilon_0 R^3$, and compare this with the polarizability of a conducting sphere of radius R, which you can find from Example 7.4. Interpret the result using the simple model of an atom in Example 9.1.

(d) With $\mathbf{E}_{\text{all other}} = \mathbf{E} + \mathbf{P}/3\varepsilon_0$ and $\mathbf{P} = N\alpha\mathbf{E}_{\text{all other}}$, show that $\mathbf{P} = N\alpha\mathbf{E}/(1-x)$. Explain why a material with $x \geq 1$ might polarize by itself, with no applied field.

Exercises for 9.5 The D field and permittivity

Warm-up questions Read Section 9.5, then answer before class

W9.11 This section says that **D** "is a combination of the field and the polarization, and its usefulness comes entirely from what its divergence turns out to be." Say in your own words what that means. A classmate says that inside a dielectric **D** is the real electric field, and **E** only a step on the way to it. What would you tell the classmate?

W9.12 An electret's polarization was frozen in when it was made. Which of these equations hold inside it: (a) $\mathbf{D} = \varepsilon_0\mathbf{E} + \mathbf{P}$; (b) $\rho_b = -\nabla\cdot\mathbf{P}$; (c) $\mathbf{P} = \varepsilon_0\chi_e\mathbf{E}$; (d) $\mathbf{D} = \varepsilon\mathbf{E}$? Give a one-sentence reason for each.

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

Exercises for 9.6 Gauss's law for D

Warm-up questions Read Section 9.6, then answer before class

W9.13 Water has a dielectric constant of 80, and PTFE (Teflon) one of 2.1 (Table 9.2). A classmate concludes that water is the better insulator. What does the dielectric constant measure, according to this section? Does it tell you which material is the better insulator?

W9.14 An infinite slab has a uniform built-in polarization **P** parallel to its faces, and there is no free charge anywhere. Find the bound charges, and from them **E** and **D** inside the slab. A classmate draws a pillbox across one face and argues from planar symmetry that $\mathbf{D} = 0$ everywhere. Where does the argument go wrong? (Which component of **D** does a pillbox tell you about?)

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P9.14 Each object below has a built-in polarization and no free charge. For each, say whether Gauss's law for **D**, together with symmetry, shows that $\mathbf{D} = 0$ everywhere, and explain why or why not.

(a) An infinitely long cylindrical shell with $\mathbf{P} = ks^2\,\hat{\mathbf{s}}$.

(b) The same shell, but of finite length.

(c) A sphere with $\mathbf{P} = (P_0/R)\,\mathbf{r}$.

(d) A sphere with $\mathbf{P} = P_0\,\hat{\mathbf{z}}$.

(e) For the sphere of part (d), find **D** inside, given that $\mathbf{E} = -\mathbf{P}/3\varepsilon_0$ there. Do **D** and **E** point the same way?

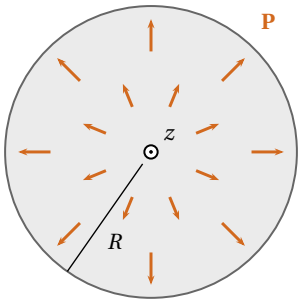


Figure 9.35 A cross section of the cylinder of Exercise P9.15. The arrows show its built-in polarization.

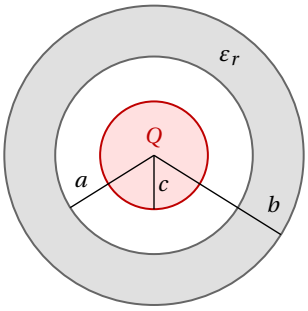


Figure 9.36 The charged ball and the dielectric shell of Exercise P9.16, in cross section.

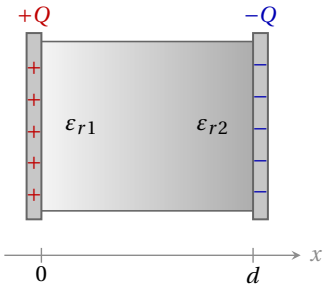


Figure 9.37 The capacitor of Exercise P9.17, edge on. The shading darkens as the dielectric constant rises.

Exercises for 9.7 Solving for E when given P: Gauss's law for D method (high symmetry)

Warm-up questions Read Section 9.7, then answer before class

W9.15 Lines of **E** begin and end on any charge, free or bound, and lines of **D** only on free charge (Figure 9.17). Use $\rho_b = -\nabla \cdot \mathbf{P}$ to explain why lines of **P** begin on negative bound charge and end on positive bound charge, the reverse of lines of **E**. Why, then, do the lines of **P** and **E** in the filled capacitor point the same way?

W9.16 Example 9.8 mentions two common mistakes in using Gauss's law for **D**. Why does counting flux through both faces of the pillbox give $Q/2A$, and why is that wrong? Why does a Gaussian surface drawn around both plates give only $0 = 0$?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P9.15 An infinitely long cylinder of radius R has the built-in polarization $\mathbf{P} = P_0\,(s/R)\,e^{-s/R}\,\hat{\mathbf{s}}$ (Figure 9.35), and there is no free charge.

(a) Use Gauss's law for **D** to find **E** inside and outside the cylinder. Draw your Gaussian surface.

(b) Find ρ_b and σ_b.

(c) Find **E** inside again, from the bound charges and Gauss's law for **E**, and check that the total bound charge per unit length is zero.

P9.16 A ball of radius c carries free charge Q, spread uniformly through its volume; the ball itself is not polarizable. It is surrounded by a concentric shell of a dielectric with dielectric constant ε_r, with inner radius $a > c$ and outer radius b (Figure 9.36). Everywhere else is vacuum.

(a) Find **D** in the four regions $r < c$, $c < r < a$, $a < r < b$, and $r > b$.

(b) Find **E** and **P** in the same regions.

(c) Find the bound charge densities on both surfaces of the shell and inside it, and show that the total bound charge is zero.

(d) Check that the jump in $E^{\perp}$ at $r = a$ is σ_b/ε_0.

P9.17 A parallel-plate capacitor has plates of area A at $x = 0$ and $x = d$, with charges $+Q$ and $-Q$ (Figure 9.37); ignore fringing. The space between them is filled with a dielectric whose dielectric constant rises linearly from ε_{r1} at $x = 0$ to ε_{r2} at $x = d$.

(a) Find **D**, **E**, and **P** between the plates.

(b) Find the capacitance, and check that it has the right limit when $\varepsilon_{r2} \to \varepsilon_{r1}$.

(c) Find ρ_b in the dielectric and σ_b on its two faces, and show that the total bound charge is zero. Why is ρ_b not zero, although there is no free charge in the dielectric?

(d) For $\varepsilon_{r1} = 2$ and $\varepsilon_{r2} = 6$, compare the capacitance with $\varepsilon_0 A/d$ times the average dielectric constant, 4.

P9.18 In a coaxial cable with uniform insulation the field is strongest at the surface of the inner conductor (Exercise P5.13), so in a high-voltage cable that is where the insulation is most likely to break down. One remedy, called *capacitance grading*, is to use insulation whose dielectric constant is largest next to the inner conductor. Consider a long cable with an inner conductor of radius a, an outer conductor of radius b, and insulation with the dielectric constant $\varepsilon_r(s) = \varepsilon_{ra}\, a/s$ between them (Figure 9.38). The inner conductor is at the potential ΔV above the outer one.

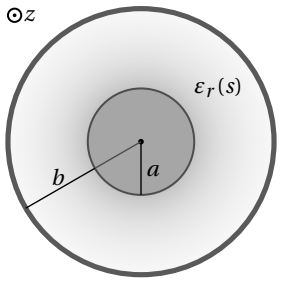


Figure 9.38 The cable of Exercise P9.18, end on. The shading darkens as the dielectric constant rises toward the inner conductor.

(a) Show that in any linear dielectric whose dielectric constant varies from place to place,

$$\rho_b = -\left(1 - \frac{1}{\varepsilon_r}\right)\rho_f - \frac{\varepsilon_0}{\varepsilon_r}\,\mathbf{E}\cdot\nabla\varepsilon_r.$$

(Write **P** in terms of **D**, and use the product rule (1.52).) Check that this reduces to (9.28) when ε_r is constant, and that it gives your ρ_b of Exercise P9.17.

(b) With free charge λ per unit length on the inner conductor, find **D**, **E**, and **P** in the insulation. Show that E is the same at every radius, and find λ in terms of ΔV.

(c) Find ρ_b in the insulation, once from $-\nabla\cdot\mathbf{P}$ and once from part (a), and σ_b on the insulation's inner and outer surfaces. Show that the total bound charge per unit length is zero.

(d) Explain why E comes out the same at every radius, in terms of how the net charge per unit length, free and bound, inside a cylinder of radius s depends on s.

(e) For $b/a = 2.5$, find the ratio of the field in the graded cable to the largest field in a cable with uniform insulation and the same ΔV. By what factor can ΔV be raised in the graded cable before its field reaches that largest value? What is the smallest ε_{ra} for which $\varepsilon_r \geq 1$ throughout the insulation?

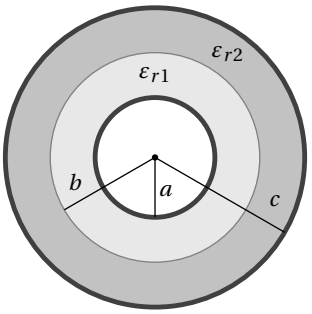


Figure 9.39 The spherical capacitor of Exercise P9.19, in cross section.

P9.19 Two concentric conducting shells, of radii a and c, are separated by two layers of dielectric (Figure 9.39): dielectric constant ε_{r1} for $a < r < b$ and ε_{r2} for $b < r < c$.

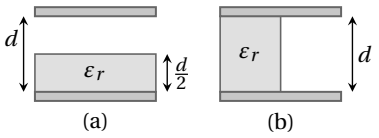


Figure 9.40 The two arrangements of Exercise P9.20, edge on: (a) a slab on the lower plate; (b) a slab that fills half of the gap.

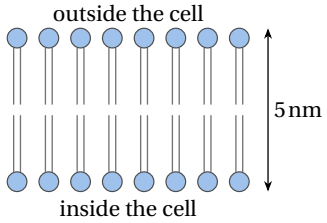


Figure 9.41 The membrane of Exercise P9.21, edge on. Each lipid molecule has a head (circle) that faces the water on one side and two oily tails that point toward the middle.

(a) Find the capacitance.

(b) With charge $+Q$ on the inner shell and $-Q$ on the outer one, find the bound surface charge density at $r = b$, where the two dielectrics meet, and check it against the jump in $E^{\perp}$ there.

(c) Under what condition is that bound charge zero?

P9.20 A parallel-plate capacitor has plates of area A a distance d apart, and capacitance $C_0 = \varepsilon_0 A/d$ when empty.

(a) A slab of dielectric constant ε_r and thickness $d/2$, with the same area as the plates, is laid flat on the lower plate (Figure 9.40a). Find the capacitance.

(b) Instead, a slab of thickness d fills the gap over half of the plates' area (Figure 9.40b). Find the capacitance.

(c) In each case, compare your answer with $\varepsilon_r C_0$, which is what the rule of replacing ε_0 by ε would give. Which condition of Section 9.6.3 is not met?

(d) In part (a), which field is the same in the dielectric and in the empty part of the gap, **E** or **D**? In part (b)?

P9.21 The membrane around a living cell is a double layer of lipid molecules about 5 nm thick (Figure 9.41), and its measured capacitance per unit area is about $1\,\mu\text{F/cm}^2$.

(a) Treat the membrane as a parallel-plate capacitor filled with a dielectric of dielectric constant 2, the value for its oily interior. What thickness does the measured capacitance imply? What capacitance per unit area would a layer 5 nm thick give? Suggest a reason why the two do not agree.

(b) The potential difference across the membrane is typically 70 mV. Find the free charge per unit area on each face, and the number of singly charged ions per square micrometer that it represents.

(c) Using the thickness found in part (a), estimate the field in the membrane, and compare it with the field at which air breaks down, about $3 \times 10^6\,\text{V/m}$.

(d) Find the energy stored per unit area of membrane.

Exercises for 9.8 Boundary conditions at dielectric interfaces

Warm-up questions Read Section 9.8, then answer before class

W9.17 The field inside a large block of linear dielectric is uniform, **E**, and there is no free charge. A small, thin cavity is cut in the block. Use the boundary conditions of this section to find the field in the cavity, in terms of **E** and ε_r, when the cavity is (a) a thin needle along **E**; (b) a thin disk perpendicular to **E**.

W9.18 Use (9.34) to explain why, in the vacuum just outside a dielectric with a very large dielectric constant, field lines meet the surface nearly at right angles, as they do at the surface of a conductor. Which panel of Figure 9.23 would the field lines of a dielectric sphere with $\varepsilon_r = 1000$ resemble?

W9.19 Suppose that in Example 9.11 you wrote BC3 without its ε_r, as $\partial V_{\text{in}}/\partial r = \partial V_{\text{out}}/\partial r$ at $r = R$. What physical condition at the surface would that impose, and what would it say about the bound charge there? Without solving, what field inside the sphere would it lead to?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P9.22 Field lines change direction where they cross an interface between two dielectrics.

(a) At an interface between two linear dielectrics with no free surface charge, let θ_1 and θ_2 be the angles that **E** makes with the normal on the two sides. Derive (9.34).

(b) Below a flat horizontal interface is a material with $\varepsilon_r = 4.0$, and above it is air (Figure 9.42). Just below the interface, the field is 1000 V/m, at 45° to the normal, pointing up toward the interface. Find the magnitude and direction of the field just above it.

(c) Find the bound surface charge on the interface, once from the jump in $E^{\perp}$ and once from $\mathbf{P}\cdot\hat{\mathbf{n}}$.

(d) Find the parallel component of **D** on both sides, and check that its jump is equal to the jump in the parallel component of **P**.

(e) In a drawing of the field near an interface, the field lines bend toward the normal as they pass from material 1 into material 2. Which material has the larger dielectric constant?

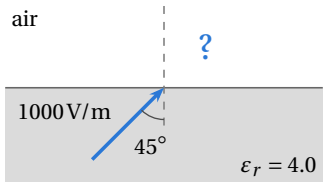


Figure 9.42 Part (b) of Exercise P9.22: the field just below the interface, in the material with $\varepsilon_r = 4.0$, and the unknown field just above it, in air.

P9.23 A long cylinder of radius R and dielectric constant ε_r is placed in a uniform field $E_0\hat{\mathbf{x}}$, perpendicular to its axis (Figure 9.43).

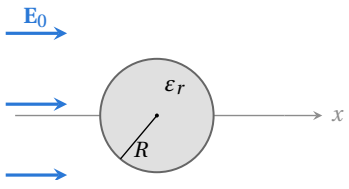


Figure 9.43 The dielectric cylinder of Exercise P9.23, in cross section, and the applied field, shown only far from the cylinder.

(a) Use separation of variables in cylindrical coordinates (Section 7.4.4) to find the potential inside and outside the cylinder.

(b) Show that the field inside is uniform and equal to $2\mathbf{E}_0/(\varepsilon_r+1)$. Compare it with the field $3\mathbf{E}_0/(\varepsilon_r+2)$ inside a sphere, for $\varepsilon_r = 4$.

(c) Find the bound surface charge, and the dipole moment per unit length of the cylinder. Check that the dipole moment per unit length is $\pi R^2 P$.

(d) What happens as $\varepsilon_r \to \infty$?

(e) A long cylinder that is uniformly polarized perpendicular to its axis has the uniform field $-\mathbf{P}/2\varepsilon_0$ inside. Use that fact to rederive the result of part (b) by the self-consistent argument used for the sphere in Section 9.8.4.

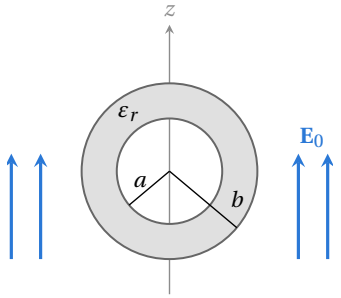


Figure 9.44 The dielectric shell of Exercise P9.24, in cross section, and the applied field, shown only far from the shell.

P9.24 A spherical shell of dielectric, with dielectric constant ε_r, inner radius a, and outer radius b, is placed in a uniform field $E_0\hat{\mathbf{z}}$ (Figure 9.44).

(a) Write the potential in the three regions as sums of Legendre terms, and show that the boundary conditions leave only the $\ell = 1$ terms.

(b) Show that the field in the cavity is uniform, with magnitude

$$E_{\text{cavity}} = \frac{9\varepsilon_r E_0}{(2\varepsilon_r+1)(\varepsilon_r+2) - 2(a/b)^3(\varepsilon_r-1)^2}.$$

(c) Check the limits $\varepsilon_r \to 1$ and $a \to 0$.

(d) Evaluate E_{cavity}/E_0 for $\varepsilon_r = 4$, 10, 100, and 1000 with $a/b = 0.9$, and for $\varepsilon_r = 10$ with $a/b = 0.5$. How good a shield is a dielectric shell, compared with a conducting one?

P9.25 The proof of the second uniqueness theorem in Section 7.1.4 carries over to linear dielectrics. Suppose that two potentials V_1 and V_2 both satisfy (9.35), $\nabla\cdot(\varepsilon\nabla V) = -\rho_f$, with the same free charge, in a region filled with pieces of linear dielectric; that both are continuous, with $\varepsilon\,\partial V/\partial n$ continuous, at every interface between dielectrics, none of which carries free charge; and that both take the same values on the outer boundary of the region. Let $V_3 = V_1 - V_2$ and $\mathbf{D}_3 = -\varepsilon\nabla V_3$.

(a) Show that $\nabla\cdot(V_3\mathbf{D}_3) = -\varepsilon|\nabla V_3|^2$ everywhere in the region except at the interfaces.

(b) Integrate over each piece of dielectric separately and use the divergence theorem on each. Show that the surface terms at each interface cancel and that those on the outer boundary are zero, so that $\int \varepsilon|\nabla V_3|^2\,d\tau = 0$.

(c) Conclude that $V_1 = V_2$ everywhere in the region. Where did you use the fact that $\varepsilon > 0$?

(d) In Example 9.11 the region is all of space. Show that the argument still works if V_3 falls off at least as fast as $1/r$ far away.

P9.26 Near the junction between the p-type and n-type regions of a silicon diode, the mobile charge carriers are swept away, leaving the fixed charge of the ionized impurity atoms: $\rho = -eN_A$ for $-x_p < x < 0$ and $\rho = +eN_D$ for $0 < x < x_n$, where N_A and N_D are the densities of the two kinds of impurity (Figure 9.45). Outside this depletion region the field is zero. Silicon is a linear dielectric with $\varepsilon = 11.7\varepsilon_0$, and the impurity charge plays the role of free charge, so (9.35) applies; with uniform ε it is $d^2V/dx^2 = -\rho/\varepsilon$.

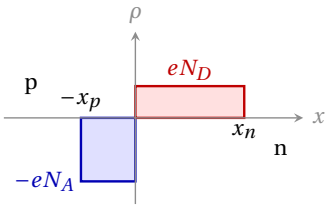


Figure 9.45 The charge density near the junction of Exercise P9.26.

(a) Find $E(x)$ and $V(x)$.

(b) Show that $N_A x_p = N_D x_n$, and that the potential difference across the junction is $V_{\text{bi}} = (e/2\varepsilon)(N_A x_p^2 + N_D x_n^2)$.

(c) For $N_A = N_D = 10^{16}\,\text{cm}^{-3}$ and $V_{\text{bi}} = 0.70\,\text{V}$, find the width $W = x_p + x_n$ of the depletion region and the largest field, with its direction.

(d) Repeat for $N_A = 10^{18}\,\text{cm}^{-3}$ and $N_D = 10^{16}\,\text{cm}^{-3}$, for which $V_{\text{bi}} = 0.83\,\text{V}$. What fraction of the depletion region is on the n side?

(e) A reverse bias V_R adds to V_{bi}, and the depletion region acts like a parallel-plate capacitor of thickness W. Find the capacitance per unit area of the junction of part (c) at zero bias and at a reverse bias of 5.0 V.

(f) What width would you get in part (c) if you forgot the dielectric constant of silicon?

Exercises for 9.9 Numerical methods: direct integration and relaxation

Warm-up questions Read Section 9.9, then answer before class

W9.20 Figure 9.24 shows a grid point on the surface of a dielectric, where the neighbor inside the dielectric gets the weight $\varepsilon_E = \varepsilon$ in (9.38) and the neighbor outside gets $\varepsilon_W = \varepsilon_0$. Why should the neighbor inside count more? Explain with the continuity of $\varepsilon\,\partial V/\partial n$: on which side of the surface does V change more slowly?

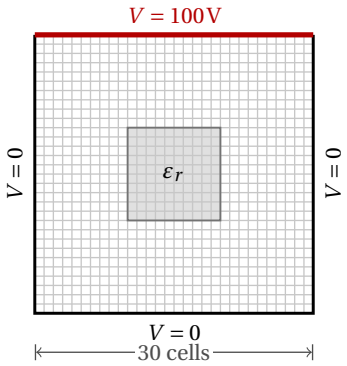


Figure 9.46 The region of Exercise P9.27 and of Example 9.12: $V = 100\,\text{V}$ on the top edge and 0 on the other three, with a block of dielectric (shaded) filling the middle third, on a grid of 30×30 cells.

W9.21 Example 9.12 explains why the potential at the center is exactly 25 V for every ε_r. Would that still be true if (a) the block were moved off center; (b) the square block were replaced by a centered disk; (c) it were replaced by a centered rectangle twice as wide as it is tall? Explain each with the example's argument.

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P9.27 Write a relaxation program based on (9.38) for the region of Example 9.12 (Figure 9.46): a square with $V = 100\,\text{V}$ on its top edge and $V = 0$ on the other three, and a square block of dielectric filling the middle third in each direction. Use a grid of 30×30 cells. *Caution:* the larger ε_r, the more passes the program needs (Section 9.9.2 explains why), so do not stop after a fixed number of passes.

(a) The program has to handle the grid points at the corners of the block. Apply Gauss's law for $\mathbf{D}$ to the small square around such a point, as in Section 9.9.2, and find the four weights ε_E, ε_W, ε_N, and ε_S. Check that they agree with the rule of giving each side of the small square the average permittivity of the two cells it passes through.

(b) Find the potential for $\varepsilon_r = 1$, 4, and 10, and plot each along the vertical line through the center. Find the potential difference across the block along that line, and the factor by which the block reduces the field inside it there.

(c) Start each run with $V = 0$ at every grid point not on the edges, and stop when the potential at the center is within 0.01 V of its exact value, 25 V. How many passes does each value of ε_r need? How large was the largest change in V during the last pass? If you had stopped instead when the largest change in one pass fell below $10^{-3}\,\text{V}$, how far from 25 V would the center have been?

P9.28 Adapt your program from Exercise P9.27 to model a long dielectric cylinder of radius R and $\varepsilon_r = 4$ in a uniform field $E_0\hat{\mathbf{x}}$, as in Exercise P9.23. Take a square region of half-width H centered on the cylinder, with $V = -E_0 x$ on its edges, and give each grid cell the permittivity of the cylinder if its center lies inside the cylinder and ε_0 otherwise.

(a) Find the field at the center for $H = 3R$, $5R$, and $10R$, with grid spacings $R/10$ and $R/20$.

(b) Compare with $2E_0/(\varepsilon_r + 1)$, the result for a cylinder in an infinite region, and with $E_{\text{in}}/E_0 = 2H^2/[(\varepsilon_r + 1)H^2 - (\varepsilon_r - 1)R^2]$, the exact result when $V = -E_0 x$ is imposed on a *circle* of radius H instead of a square. What limits the accuracy of your results?

(c) Plot V and E_x along the x axis. Is the field inside uniform?

P9.29 A long bar of linear dielectric with $\varepsilon_r = 4$ has a square cross section 2.0 cm on a side (Figure 9.47). It sits in the middle of a square region 8.0 cm on a side, on whose edges $V = -E_0 x$ with $E_0 = 1.00\,\text{kV/m}$, so the applied field is uniform and points along x. The origin is at the center of the bar, whose faces are at $x = \pm 1.0\,\text{cm}$ and $y = \pm 1.0\,\text{cm}$. The file `ch09-square-bar.csv`, on the book's web site at https://em.byu.edu/data, gives the potential found with (9.38) on a grid with spacing $h = 1.0\,\text{mm}$: 81 rows of 81 values, in volts. The value in row j and column i (both counted from 0) is $V_{i,j}$, the potential at $x = (i - 40)\,\text{mm}$ and $y = (j - 40)\,\text{mm}$. (You can also make the file yourself with your program from Exercise P9.27.)

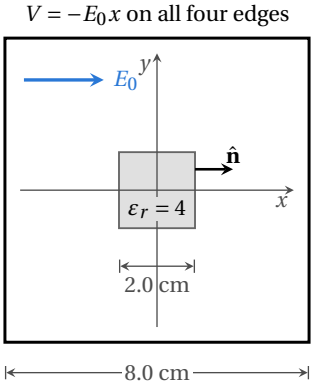


Figure 9.47 The dielectric bar of Exercise P9.29, 2.0 cm square with $\varepsilon_r = 4$, at the center of a region 8.0 cm square whose edges are held at $V = -E_0 x$, so that the applied field points along $+x$. The outward normal $\hat{\mathbf{n}}$ is shown on one face.

(a) At every grid point on the surface of the bar except the four corners, find the slope $\partial V/\partial n$ just inside and just outside the bar, with n measured along the outward normal. At a point (i, j) on the face at $x = +1.0\,\text{cm}$, for example, use

$$\left(\frac{\partial V}{\partial n}\right)_{\text{inside}} \approx \frac{3V_{i,j} - 4V_{i-1,j} + V_{i-2,j}}{2h},$$
$$\left(\frac{\partial V}{\partial n}\right)_{\text{outside}} \approx \frac{-3V_{i,j} + 4V_{i+1,j} - V_{i+2,j}}{2h}.$$

These one-sided differences use only points on one side of the surface, where V is smooth, and, like the central difference (5.18), their errors are proportional to h^2. Show that $\varepsilon_r\,(\partial V/\partial n)_{\text{inside}} = (\partial V/\partial n)_{\text{outside}}$, as (9.33) requires, except close to the corners. Where is the agreement worst, and why?

(b) Repeat part (a) with the one-cell differences

$$\left(\frac{\partial V}{\partial n}\right)_{\text{inside}} \approx \frac{V_{i,j} - V_{i-1,j}}{h}, \qquad \left(\frac{\partial V}{\partial n}\right)_{\text{outside}} \approx \frac{V_{i+1,j} - V_{i,j}}{h}.$$

How well do the two sides agree now? Why is the agreement worse? (Hint: at what point does a one-cell difference best approximate the slope?)

(c) The surface of the bar carries only bound charge. Use the boundary condition (7.2) and your slopes from part (a) to find σ_b at each of the points of part (a), being careful about which side is side 1. Plot σ_b around the perimeter of the bar. Check that the bound charge adds up to zero, and compare σ_b with $\mathbf{P}\cdot\hat{\mathbf{n}}$, where $\mathbf{P} = \varepsilon_0\chi_e\mathbf{E}$ just inside the surface.

Exercises for 9.10 Energy stored in the field, inside dielectrics

Warm-up questions Read Section 9.10, then answer before class

W9.22 The derivation of (9.39) uses $\mathbf{D} = \varepsilon\mathbf{E}$ in one step. Which step, and what does it accomplish? A ferroelectric's P, plotted against E, traces out a hysteresis loop (Section 9.2.2), so the work needed to charge a capacitor filled with one depends on the path, not just on the final charge. What is the significance of the area of the P–E loop?

W9.23 Between the plates of a capacitor the field is uniform, and its forces on the bound charges of a dielectric slab are perpendicular to the plates. Suppose the field were exactly uniform between the plates and exactly zero outside them. Would anything then pull a partly inserted slab sideways into the capacitor? Why or why not?

W9.24 Use $\oint \mathbf{E}\cdot d\boldsymbol{\ell} = 0$, around a rectangular loop that runs up through the gap between a capacitor's plates and back down just outside their edge, to show that the field cannot be exactly uniform between the plates and exactly zero outside them. (What would the side in the gap contribute, and what would the other three sides?) What must the field be like near the edges of the plates?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

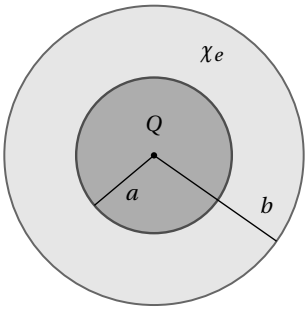


Figure 9.48 The coated sphere of Exercise P9.30, in cross section.

P9.30 A conducting sphere of radius a carries free charge Q (Figure 9.48). It is coated with a dielectric of susceptibility χ_e out to radius b; beyond that is vacuum.

(a) Find $\mathbf{D}$, $\mathbf{E}$, and $\mathbf{P}$ everywhere, and σ_b at $r = a$ and $r = b$.

(b) Find the energy $U = \frac{1}{2}\int \mathbf{D}\cdot\mathbf{E}\,d\tau$.

(c) Find the capacitance of the sphere from its potential, and check that $U = Q^2/2C$. Is the capacitance ε_r times that of the bare sphere? Why or why not?

(d) Find $(\varepsilon_0/2)\int E^2\,d\tau$, and show that it is smaller than U by $\frac{1}{2}\int \mathbf{P}\cdot\mathbf{E}\,d\tau$. Which of the two is the work needed to put the charge Q on the sphere, and what does the difference represent?

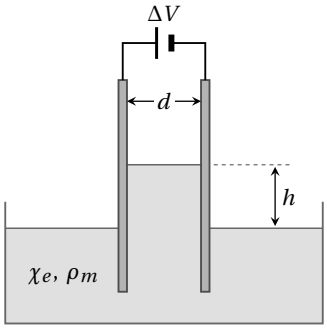


Figure 9.49 The plates and the liquid of Exercise P9.31.

P9.31 Two vertical parallel plates, a distance d apart, dip into a liquid dielectric with susceptibility χ_e and density ρ_m (Figure 9.49). When a potential difference ΔV is applied between the plates, the liquid between them rises to a height h above the level outside.

(a) Explain why the liquid rises, and find h. (Treat the liquid between the plates like the slab of Section 9.10.1, with ΔV held fixed.)

(b) How much work does the battery do as the liquid rises, compared with the increase in the energy stored in the capacitor? Where does the rest go? (Hint: compare the work done on the liquid by the electric force as it rises with the gravitational potential energy that it gains.)

(c) Suppose the plates are 2.0 mm apart, the liquid has $\varepsilon_r = 2.2$ and density 880 kg/m^3, and $\Delta V = 4.0$ kV. How high does the liquid rise? Is the field below the breakdown field of the air above the liquid, about 3×10^6 V/m?

Exercises for 9.11 Summary

P9.32 True or false? Explain your reasoning.

(a) A positive point charge is embedded in a large block of linear dielectric. The negative bound charge that collects around it outweighs the point charge, so near the charge the total field points toward it.

(b) Bound volume charge can exist only at points where $\mathbf{P} \neq 0$.

(c) When a dielectric is placed in an electric field, bound charge appears on every one of its surfaces.

(d) In a homogeneous linear dielectric with no free charge in it, there is no bound volume charge, even where **P** is not uniform.

(e) Since $\nabla \cdot \mathbf{D} = \rho_f$, the field **D** is determined by the free charge alone, just as **E** is determined by the total charge.

(f) In a linear isotropic dielectric, **D** points in the same direction as **E**.

(g) Inside a uniformly polarized sphere with no free charge, **D** and **E** point in opposite directions.

(h) At a boundary between two dielectrics, the jump in $E^\perp$ depends only on the bound surface charge.

(i) Bringing an uncharged block of linear dielectric up close to an isolated metal sphere that carries a charge Q lowers the stored energy.

(j) If a battery holds the sphere at a fixed potential instead, bringing the same block up close to it also lowers the stored energy.

(k) At a boundary between two linear dielectrics with no free charge, $E^\parallel$ and $D^\perp$ are continuous, but $E^\perp$ and $D^\parallel$ need not be.

(l) The field inside any uniformly polarized object is uniform.

P9.33 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as the force on a polar molecule near a charged object, the field in and around a polarized object of any shape found from its bound charges, the field lines in and around a dielectric sphere in a uniform field, the bending of field lines at the boundary between two dielectrics, the dielectric constant of a liquid predicted from the polarizability of its molecules, or the force that pulls a dielectric slab into a capacitor.

Part III

Magnetostatics

From the preceding facts we may likewise infer that this conflict performs circles; for without this condition it seems impossible that the one part of the uniting wire, when placed below the magnetic pole, should drive it towards the east, and when placed above it towards the west; for it is the nature of a circle that the motions in opposite parts should have an opposite direction.

— Hans Christian Ørsted, "Experiments on the Effect of a Current of Electricity on the Magnetic Needle" (1820)*

*"Experiments on the effect of a current of electricity on the magnetic needle," H. C. Ørsted, Ann. Philos. 16, 273–276 (1820), translated from the Latin pamphlet *Experimenta circa effectum conflictus electrici in acum magneticam* (Copenhagen, 1820).

Chapter 10

The Lorentz Force and the Biot–Savart Law

Chapters 3 through 9 were about charges at rest. When charges move, a second kind of force appears. Two long wires carrying currents in the same direction attract each other, even though each wire is electrically neutral; a compass needle near a wire turns when a current is switched on; a beam of electrons bends when a magnet is brought near it. These are **magnetic forces**. As with the electric force, it helps to split the interaction into two steps: moving charges produce a **magnetic field B**, and the magnetic field exerts a force on other moving charges. That gives this chapter two questions to answer. How do charges and currents behave in a magnetic field? And how do charges and currents produce one? We take the first question first, since it is probably the more familiar; its answer is the Lorentz force law. The answer to the second is the Biot–Savart law, the magnetic counterpart of Coulomb's law. From it, Chapter 11 derives Ampère's law, the counterpart of Gauss's law.

This chapter and the next two are about **magnetostatics**, the magnetic fields of steady currents. Their equations will look a great deal like those of Chapters 3 through 5: similar, but not identical, and the differences matter as much as the similarities. Permanent magnets wait until Chapter 13, and fields that change in time until Chapter 15.



Hans Christian Ørsted (1777–1851, Danish) Ørsted was the son of an apothecary on the island of Langeland. He qualified as a pharmacist in 1797, and in 1799 the University of Copenhagen gave him a doctorate for a thesis on Kant's philosophy of nature. After several years of travel in Germany and France, he became a professor at Copenhagen in 1806. In the spring of 1820, during a lecture, he first saw a current in a wire deflect a compass needle. He studied the effect more carefully that July and announced it in a four-page Latin pamphlet, dated 21 July 1820, which he sent to scientists across Europe. In 1825 he announced the first production of metallic aluminum, though not in pure form, and in 1829 he founded the school that is now the Technical University of Denmark. The oersted, the CGS unit of the field **H** of Chapter 13, is named for him. (Wikipedia)

10.1 Forces on moving charges: the Lorentz force law

10.1.1 What the experiments show

Magnetic forces were known long before anyone connected them with electricity, from lodestones and compass needles. The connection came in 1820, when Hans Christian Ørsted found that a current in a wire deflects a nearby compass needle, and within weeks of hearing about it André-Marie Ampère showed that two wires carrying currents exert forces on each other: parallel currents attract, and opposite currents repel. A wire carrying a current is electrically neutral,

because the electrons that carry the current move through a lattice of positive ions with the same total charge, so this force is not the electric force of the earlier chapters. And it acts only on charges that are moving.

Experiments on charged particles moving through a magnetic field, such as electrons in a vacuum tube near a magnet, show that the force on a particle with charge q and velocity $\mathbf{v}$ has these properties:

1. It is proportional to q, and it reverses if the sign of q is reversed.
2. It is proportional to the speed, and it reverses if the velocity is reversed.
3. It is always perpendicular to the velocity.
4. At each point there is one line along which the particle can move without feeling any force. If the particle moves at an angle α to that line, the force is proportional to $\sin\alpha$.
5. The force is also perpendicular to that line.

The special line through each point gives the direction of the magnetic field there, up to a sign. The field of a long straight wire, for example, circles the wire: a compass needle near the wire turns until it points around the wire, not toward it or along it.

These observations lead us to describe magnetism with vector cross products, the directions of which are given by the **right-hand rule**. It pays to be fluent with the rules for finding those directions.

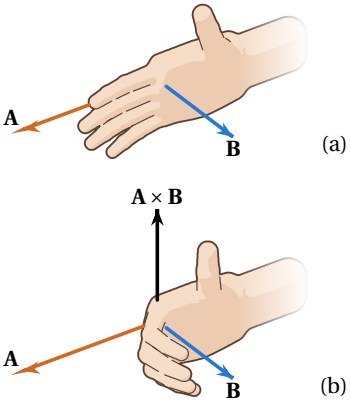


Figure 10.1 The right-hand rule for $\mathbf{A}\times\mathbf{B}$: (a) point your right hand, with the fingers flat, along $\mathbf{A}$; (b) bend your fingers toward $\mathbf{B}$. Your thumb then points along $\mathbf{A}\times\mathbf{B}$.

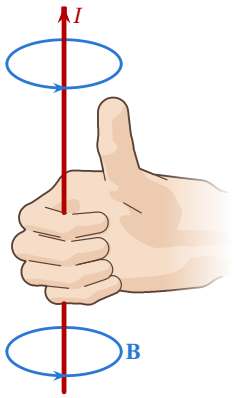


Figure 10.2 The right-hand rule for the magnetic field of a straight current: with the thumb of your right hand along the current I, your fingers curl around the wire in the direction of $\mathbf{B}$.

- For $\mathbf{A}\times\mathbf{B}$, point your whole right hand, with the fingers flat, in the direction of $\mathbf{A}$, then bend your fingers toward $\mathbf{B}$, through the smaller angle between the two; your thumb then points in the direction of $\mathbf{A}\times\mathbf{B}$ (Figure 10.1).
- For the field of a current, point your right thumb along the current; your fingers curl in the direction of $\mathbf{B}$ (Figure 10.2). The field of a straight current, including its direction, comes mathematically from the Biot–Savart law (Section 10.5).
- For unit vectors, write each coordinate system's unit vectors in its usual order, (x,y,z), (s,ϕ,z), or (r,θ,ϕ). The cross product of two of them, taken from left to right in that order, is the third, and taken from right to left it is the negative of the third, with the list wrapping around at the end:

$$\hat{\mathbf{x}}\times\hat{\mathbf{y}}=\hat{\mathbf{z}},\quad \hat{\mathbf{y}}\times\hat{\mathbf{z}}=\hat{\mathbf{x}},\quad \hat{\mathbf{z}}\times\hat{\mathbf{x}}=\hat{\mathbf{y}};\qquad \hat{\mathbf{s}}\times\hat{\boldsymbol{\phi}}=\hat{\mathbf{z}},\quad \hat{\boldsymbol{\phi}}\times\hat{\mathbf{z}}=\hat{\mathbf{s}},\quad \hat{\mathbf{z}}\times\hat{\mathbf{s}}=\hat{\boldsymbol{\phi}};$$

$$\hat{\mathbf{r}}\times\hat{\boldsymbol{\theta}}=\hat{\boldsymbol{\phi}},\quad \hat{\boldsymbol{\theta}}\times\hat{\boldsymbol{\phi}}=\hat{\mathbf{r}},\quad \hat{\boldsymbol{\phi}}\times\hat{\mathbf{r}}=\hat{\boldsymbol{\theta}};$$

and reversing the order of any of them changes the sign, so that $\hat{\mathbf{y}}\times\hat{\mathbf{x}}=-\hat{\mathbf{z}}$ and $\hat{\boldsymbol{\phi}}\times\hat{\mathbf{s}}=-\hat{\mathbf{z}}$. These hold for two unit vectors at the same point, wherever that point is, even though $\hat{\mathbf{s}}$, $\hat{\boldsymbol{\phi}}$, $\hat{\mathbf{r}}$, and $\hat{\boldsymbol{\theta}}$ point in different directions at different points. They do not hold for unit vectors at two different points, such as $\hat{\boldsymbol{\phi}}'$ at a source point and $\hat{\mathbf{s}}$ at a field point.

Figures in magnetism often need vectors that point into or out of the page. The symbol ⊙ means *out of* the page: picture an arrow coming toward you, and you see its point. The symbol ⊗ means *into* the page: you see the feathers at the back of an arrow going away.

10.1.2 The Lorentz force law

Everything in the list of five properties is summarized by one equation. The magnetic force on a charge q moving with velocity $\mathbf{v}$ through a magnetic field $\mathbf{B}$ is

$$\mathbf{F} = q\,\mathbf{v}\times\mathbf{B}. \tag{10.1}$$

The cross product supplies all five properties at once. It is proportional to q and to $\mathbf{v}$; it is perpendicular to $\mathbf{v}$ and to $\mathbf{B}$; and its magnitude is $qvB\sin\alpha$, which is zero when $\mathbf{v}$ is along $\mathbf{B}$. Its direction is given by the right-hand rule for cross products (Figure 10.1); for the magnetic force on a negative charge, reverse the result. Section 10.2 shows that the cross product is not just a convenient way to write these facts: for a force that is linear in the velocity, the first three properties already require it. If an electric field is present too, the total force is

$$\mathbf{F} = q\,(\mathbf{E}+\mathbf{v}\times\mathbf{B})\,. \qquad \text{(Lorentz force law)} \tag{10.2}$$

With words, a charge feels an electric force $q\mathbf{E}$ whether or not it is moving, and a magnetic force $q\mathbf{v}\times\mathbf{B}$, perpendicular to its velocity, only if it is moving. Equation (10.2) is called the **Lorentz force law**. (Some authors use the name for the magnetic part, (10.1), alone; there is no general agreement.) Oliver Heaviside wrote the magnetic force on a moving charge in this form in 1889, and Hendrik Lorentz made it the foundation of his theory of electrons, beginning in 1892; a $\mathbf{v}\times\mathbf{B}$ term already appears in the electromotive force that Maxwell wrote in 1865 for a conductor moving through a magnetic field.

The Lorentz force law is an experimental law, but it is also, in a sense, the definition of the magnetic field, in the same way that $\mathbf{F} = q\mathbf{E}$ defines the electric field. It does not say how to calculate $\mathbf{B}$; that is the subject of most of the rest of this chapter. It says what $\mathbf{B}$ is: the field that makes (10.1) true for every moving test charge. Its SI unit follows from (10.1): a field that exerts a force of one newton on a charge of one coulomb moving at one meter per second, perpendicular to the field, is one **tesla**,

$$1\,\mathrm{T} = 1\,\frac{\mathrm{N}}{\mathrm{C}\cdot\mathrm{m/s}} = 1\,\frac{\mathrm{N}}{\mathrm{A}\cdot\mathrm{m}}.$$

A tesla is a large field. Earth's field at its surface is about $50\,\mu$T, and the CGS unit, the **gauss** ($1\,\mathrm{G} = 10^{-4}\,\mathrm{T}$), is often more convenient for fields of everyday size (Section 17.1.7).



Hendrik Antoon Lorentz (1853–1928, Dutch) Lorentz studied at the University of Leiden and received his doctorate in 1875 for a thesis on the reflection and refraction of light. In 1878 he took up Leiden's new chair of theoretical physics, one of the first such chairs in Europe. From 1892 he developed a theory in which matter contains small charged particles that respond to electric and magnetic fields, and in it he gave the force on such a particle in the form used today. The theory explained the splitting of spectral lines in a magnetic field discovered in 1896 by his former student Pieter Zeeman, and the two shared the 1902 Nobel Prize in Physics. His equations for moving bodies, including the transformation that bears his name, prepared the way for special relativity. He presided over the Solvay conferences from 1911 to 1927, and from 1918 to 1926 he chaired the state committee that worked out how closing off the Zuiderzee would change the tides. (Wikipedia)

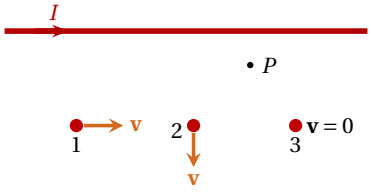


Figure 10.3 Example 10.1: a long straight wire carrying a current I to the right, the point P below it, and three positive charges: (1) moving to the right, (2) moving straight down, (3) at rest.

Example 10.1

A long straight wire carries a current I to the right (Figure 10.3). (a) Which way

does its magnetic field point at the point P, below the wire? (b) Find the direction of the magnetic force on a positive charge below the wire if the charge is (1) moving to the right, parallel to the current; (2) moving straight down, away from the wire; (3) at rest. (c) What if the charge moves straight into the page?

Solution:



We start by drawing the field of the wire, which the right-hand rule gives, and the axes we will use for the cross products.

(a) Put your right thumb along the current, pointing to the right. Your fingers curl out of the page above the wire and into the page below it, so at P the field points into the page. Take the page to be the xy plane, with $\hat{\mathbf{x}}$ to the right, $\hat{\mathbf{y}}$ up the page, and $\hat{\mathbf{z}}$ out of the page. Then below the wire $\mathbf{B} = -B\hat{\mathbf{z}}$, with $B > 0$.

(b) Use (10.1) for each charge, with the cross products of the unit vectors ($\hat{\mathbf{x}} \times \hat{\mathbf{y}} = \hat{\mathbf{z}}$ and its cyclic relatives, Section 10.1.1). For charge 1, $\mathbf{v} = v\hat{\mathbf{x}}$ and

$$\begin{aligned}\mathbf{F} &= q\,(v\hat{\mathbf{x}}) \times (-B\hat{\mathbf{z}}) \\ &= -qvB\,(\hat{\mathbf{x}} \times \hat{\mathbf{z}}) \\ &= qvB\hat{\mathbf{y}},\end{aligned}$$

up the page, toward the wire. For charge 2, $\mathbf{v} = -v\hat{\mathbf{y}}$ and

$$\begin{aligned}\mathbf{F} &= q\,(-v\hat{\mathbf{y}}) \times (-B\hat{\mathbf{z}}) \\ &= qvB\,(\hat{\mathbf{y}} \times \hat{\mathbf{z}}) \\ &= qvB\hat{\mathbf{x}},\end{aligned}$$

to the right, parallel to the current. Charge 3 has $\mathbf{v} = 0$, so it feels no magnetic force at all, however strong the field.

(c) A velocity into the page is parallel to $\mathbf{B}$, which also points into the page below the wire, and the cross product of parallel vectors is zero. There is no magnetic force.

Reality check: The right-hand rule for cross products (Figure 10.1) gives the directions of (b) without any axes. Point the fingers of your right hand along $\mathbf{v}$ and bend them toward $\mathbf{B}$, into the page: your thumb points up the page for charge 1, and to the right for charge 2.

Notice what (b) says about charge 1. A positive charge moving parallel to the current is attracted toward the wire. A current in a second wire, parallel to the first, is just such a stream of moving charges, which is why parallel currents attract.

The force on a wire that carries a current follows from (10.1) in the same way, because a current is a stream of moving charges. Take a straight piece of wire of length L in a uniform field $\mathbf{B}$, and let λ be the charge per unit length of the charges that move in it, with drift velocity $\mathbf{v}$ along the wire. (The wire as a whole is neutral; λ counts only the charges that move.) Those charges carry a total charge λL, so the magnetic force on them is $\lambda L\mathbf{v}\times\mathbf{B}$. The current is $I=\lambda v$ (Section 10.4), so if $\mathbf{L}$ is a vector of length L that points along the current, $\lambda L\mathbf{v}=I\mathbf{L}$, and

$$\mathbf{F}=I\mathbf{L}\times\mathbf{B}. \qquad \text{(force on a straight wire)} \quad (10.3)$$

Section 10.4.2 gives the force on a wire of any shape, in a field that need not be uniform.

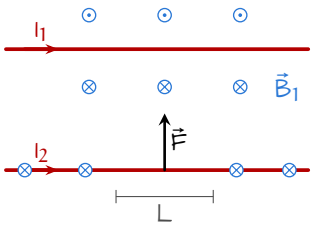
Example 10.2

Two long straight parallel wires, a distance d apart, carry currents I_1 and I_2 in the same direction (Figure 10.4). Find the magnetic force per unit length that each wire exerts on the other. The field of a long straight wire carrying a current I has the magnitude $B=\mu_0 I/2\pi s$ at a distance s from the wire, as you may remember from your introductory course; Example 10.5 derives it. Here μ_0 is the permeability of free space, very nearly $4\pi\times10^{-7}\,\mathrm{N/A^2}$ (Section 10.5).



Figure 10.4 Example 10.2: two long parallel wires, a distance d apart, carrying currents I_1 and I_2 in the same direction.

Solution:



To start out, we draw the field of wire 1 where wire 2 is. By the right-hand rule it circles wire 1: out of the page above it, and into the page below it, so at wire 2 it points into the page. Its magnitude there is $B_1=\mu_0 I_1/2\pi d$, and it is perpendicular to wire 2. Then we mark a length L of wire 2 and find the force on that length with (10.3).

Take $\hat{\mathbf{x}}$ along the currents, to the right, $\hat{\mathbf{y}}$ up the page from wire 2 toward wire 1, and $\hat{\mathbf{z}}$ out of the page. At wire 2, $\mathbf{B}_1=-(\mu_0 I_1/2\pi d)\,\hat{\mathbf{z}}$, and a length L of wire 2 has $\mathbf{L}=L\hat{\mathbf{x}}$, so

$$\begin{aligned}\mathbf{F}&=I_2\mathbf{L}\times\mathbf{B}_1\\&=I_2L\hat{\mathbf{x}}\times\left(-\frac{\mu_0 I_1}{2\pi d}\hat{\mathbf{z}}\right)\\&=\frac{\mu_0 I_1 I_2 L}{2\pi d}\,(-\hat{\mathbf{x}}\times\hat{\mathbf{z}})\\&=\frac{\mu_0 I_1 I_2 L}{2\pi d}\hat{\mathbf{y}},\end{aligned}$$

toward wire 1. The force per unit length has the magnitude

$$\frac{F}{L}=\frac{\mu_0 I_1 I_2}{2\pi d}. \qquad (10.4)$$

The same reasoning for the force of wire 2 on wire 1 gives the same magnitude and the opposite direction, as Newton's third law requires, so the wires attract. If one of the currents is reversed, both forces reverse, and the wires repel.

Two wires one meter apart, each carrying one ampere, attract with a force of 2×10^{-7} newtons per meter of length. From 1948 until 2019, that was the definition of the ampere: the current that, in two long parallel wires one meter apart, produces exactly this force (Sections 17.1.4 and 17.4).

Reality check: The units are right: with μ_0 in $\mathrm{N/A^2}$ and the currents in amperes, $\mu_0 I_1 I_2/2\pi d$ is in newtons per meter, a force per unit length.

10.1.3 Magnetic forces do no work

Because the magnetic force is always perpendicular to the velocity of a moving charge, magnetic fields do no work on moving charges. In a time dt the charge moves $d\boldsymbol{\ell} = \mathbf{v}\,dt$, and the work done on it is

$$\begin{aligned} dW &= \mathbf{F}\cdot d\boldsymbol{\ell} \\ &= q(\mathbf{v}\times\mathbf{B})\cdot\mathbf{v}\,dt \\ &= 0, \end{aligned}$$

since $\mathbf{v}\times\mathbf{B}$ is perpendicular to $\mathbf{v}$. A magnetic force can change the direction of a charge's motion, but never its speed or its kinetic energy. This holds whether or not the field is uniform, and even if it changes in time; a field that changes in time does produce changes in kinetic energy, but through the *electric* field that the changing magnetic field induces (Section 15.2), not through $q\mathbf{v}\times\mathbf{B}$.

This can seem to contradict everyday experience, in which magnets lift things and electric motors do work. Example 10.2 makes the puzzle sharper. We have just calculated a magnetic force between two wires that attracts them to each other, and if nothing stops them, it will move them closer together, faster and faster. How can it be that magnetic forces do no work? The magnetic force does take part in such processes, but as an agent that redirects work done by something else. A box pushed up a ramp is a useful comparison: the normal force of the ramp is perpendicular to the box's motion and does no work, but it turns a horizontal push into motion up the ramp. When a current in a wire experiences a perpendicular magnetic force, that force pushes the moving charges toward one side of the wire, where they pile up, and *electrical* forces from the piled-up charge cause the wire to be pushed to the side (Figure 10.5). Section 10.4.2 shows where the energy comes from when a magnetic force lifts a wire that carries a current: the battery that keeps the current flowing. The same is true of the wires of Example 10.2. (A permanent magnet lifting a nail is a subtler case, because the magnetism of iron comes mostly from the spin of its electrons rather than from currents in the ordinary sense; see Section 13.11.2.)

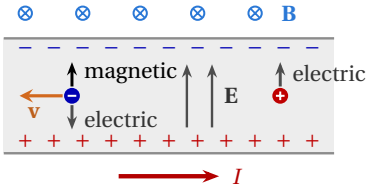


Figure 10.5 A wire carrying a current I to the right in a magnetic field $\mathbf{B}$ that points into the page. The magnetic force pushes the drifting electrons (velocity $\mathbf{v}$) toward the top of the wire. The negative charge that piles up there, and the positive charge left behind at the bottom, make an electric field $\mathbf{E}$ inside the wire. It holds the electrons to their path along the wire and pushes the positive ions, and with them the whole wire, upward.

10.1.4 Motion in a uniform field

In a uniform magnetic field, with no electric field, a charge moving perpendicular to the field moves in a circle. The force has the constant magnitude qvB and is always perpendicular to the velocity, which is exactly the force needed for uniform circular motion. Setting qvB equal to mv^2/R gives the radius,

$$R = \frac{mv}{qB} = \frac{p}{qB}, \tag{10.5}$$

where p is the momentum. (Here q means the magnitude of the charge.) The angular frequency of the motion is $\omega = v/R$, or

$$\omega_c = \frac{qB}{m}, \qquad \nu_c = \frac{qB}{2\pi m}, \tag{10.6}$$

the **cyclotron frequency**. Faster particles move on larger circles, but they take exactly the same time to go around. Looking along the field in the direction it points, a positive charge circles counterclockwise and a negative charge clockwise (Figure 10.6a). You do not need to memorize the radius: it is a two-line derivation, and it is worth being able to do in your head.

If the velocity has a component $v_\parallel$ along the field, that component feels no force and stays constant, while the perpendicular component $v_\perp$ circles as before. The path is a helix wound around the field direction (Figure 10.6b), with radius $mv_\perp/qB$; in each period $2\pi m/qB$ the particle advances a distance $2\pi m v_\parallel/qB$, the **pitch** of the helix. Charged particles spiral along magnetic field lines this way in Earth's field. Those trapped in it make up the Van Allen radiation belts, discovered in 1958 with Geiger counters carried by the first American satellites, and electrons guided down along the field lines near the poles make the air glow in auroras.

The speed independence of (10.6) is the principle of the **cyclotron**, the particle accelerator built by Ernest Lawrence and his student M. Stanley Livingston in 1931 and 1932. Ions circle in a magnetic field inside two hollow D-shaped electrodes and get a push from an alternating voltage each time they cross the gap between them. Because each half circle takes the same time whatever the speed, a voltage of fixed frequency stays in step with the ions as they speed up and spiral outward.

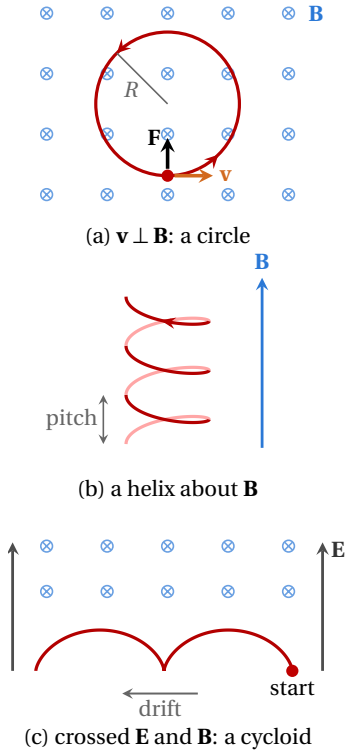
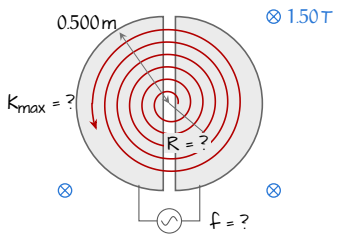


Figure 10.6 Motion of a positive charge in uniform fields. (a) With the velocity perpendicular to **B**, a circle; the force $q\mathbf{v}\times\mathbf{B}$ points toward the center. (b) With a component of velocity along **B**, a helix. (c) In crossed fields, **E** up the page and **B** into it, a charge released from rest moves on a cycloid and drifts in the direction of $\mathbf{E}\times\mathbf{B}$. Animation of (a)–(c) and of a magnetic bottle. Animation of electrons making an aurora.

Example 10.3

A cyclotron accelerates protons in a magnetic field of 1.50 T. (a) At what frequency must its voltage alternate? (b) What is the radius of the orbit of a proton with a kinetic energy of 1.00 MeV? (c) If the electrodes have a radius of 0.500 m, what is the largest kinetic energy the protons can reach?

Solution:



We label a sketch of the cyclotron with what we are given and what we want: the frequency of the voltage, the radius of the orbit at 1.00 MeV, and the largest kinetic energy the electrodes allow.

(a) By (10.6), with $q = e$ and $m = 1.673 \times 10^{-27}\,\text{kg}$,

$$\begin{aligned} \nu_c &= \frac{eB}{2\pi m} \\ &= \frac{(1.602 \times 10^{-19}\,\text{C})(1.50\,\text{T})}{2\pi\,(1.673 \times 10^{-27}\,\text{kg})} \\ &= 2.29 \times 10^{7}\,\text{Hz} \\ &= 22.9\,\text{MHz}, \end{aligned}$$

a radio frequency.

(b) A kinetic energy of $1.00\,\text{MeV} = 1.602 \times 10^{-13}\,\text{J}$ gives $v = \sqrt{2K/m} = 1.38 \times 10^{7}\,\text{m/s}$, about 5% of the speed of light, so nonrelativistic mechanics is adequate. By (10.5),

$$\begin{aligned} R &= \frac{mv}{eB} \\ &= \frac{(1.673 \times 10^{-27}\,\text{kg})(1.384 \times 10^{7}\,\text{m/s})}{(1.602 \times 10^{-19}\,\text{C})(1.50\,\text{T})} \\ &= 0.0963\,\text{m}, \end{aligned}$$

keeping one more digit of v than we quoted.

(c) The orbit radius grows with the speed, $v = eBR/m$, so the largest speed is reached at the edge of the electrodes, and

$$\begin{aligned} K_{max} &= \frac{1}{2}mv^2 \\ &= \frac{(eBR)^2}{2m} \\ &= 4.31 \times 10^{-12}\,\text{J} \\ &= 26.9\,\text{MeV}. \end{aligned}$$

At this energy the proton's speed is a quarter of the speed of light, and relativity (Example 25.7) starts to matter: the period grows with the energy, and the ions slowly fall out of step with the voltage. Larger machines change the frequency or the field as the particles speed up.

Reality check: The proton of (b) goes around its orbit in $2\pi R/v = 43.7\,\text{ns}$, which is one period of the voltage of (a), $1/(22.9\,\text{MHz}) = 43.7\,\text{ns}$, as it must be, since the period does not depend on the speed.

Adding a uniform electric field perpendicular to **B** gives two more useful results. A particle moving perpendicular to both fields with exactly the speed

$$v = \frac{E}{B}$$

feels electric and magnetic forces that cancel, and it moves in a straight line, while particles with other speeds are deflected. Crossed fields therefore make a *velocity selector*, and J. J. Thomson used them in 1897 to measure the speed of cathode rays and from it their charge-to-mass ratio. A particle released from rest in crossed fields moves on a cycloid, the curve traced by a point on the rim of a rolling wheel (Figure 10.6c): it speeds up along **E**, is turned by **B**, comes to rest at a cusp, and starts again, drifting on average with the velocity $\mathbf{E}\times\mathbf{B}/B^2$. Exercise P10.3 works this out. Notice that the drift is perpendicular to the electric field, not along it.



J. J. Thomson (1856–1940, English) Joseph John Thomson entered Owens College in Manchester at fourteen, went on to Trinity College, Cambridge, in 1876, and graduated second wrangler in 1880. In 1884, at twenty-eight, he succeeded Lord Rayleigh as Cavendish Professor of Experimental Physics, and he led the Cavendish Laboratory until 1918. In 1897 he showed that cathode rays are streams of negatively charged particles, which he called corpuscles, with a charge-to-mass ratio more than a thousand times that of a hydrogen ion: the electron, the first particle found that is smaller than an atom. He received the Nobel Prize in Physics in 1906 for his work on the conduction of electricity through gases, and he was knighted in 1908. Ernest Rutherford was one of several of his students who later won Nobel Prizes, and in 1937 his son George shared one for showing that electrons are diffracted by crystals, as waves are. Thomson was Master of Trinity College from 1918 until his death. (Wikipedia)

10.1.5 Numerical method: trajectories

In fields that are not uniform, such as Earth's, the equation of motion $m\,d\mathbf{v}/dt = q(\mathbf{E}+\mathbf{v}\times\mathbf{B})$ usually has to be solved numerically. The simplest way is to take small time steps Δt. At each step, calculate the force at the particle's present position and velocity, divide by m to get the acceleration, and use it to update the velocity and then the position:

$$\mathbf{v}_{n+1} = \mathbf{v}_n + \frac{q}{m}(\mathbf{E}+\mathbf{v}_n\times\mathbf{B})\,\Delta t,$$
$$\mathbf{x}_{n+1} = \mathbf{x}_n + \mathbf{v}_{n+1}\,\Delta t.$$

As in Section 3.6, each quantity is a vector, kept in Cartesian components, and the calculation is repeated with smaller and smaller Δt until the trajectory stops changing.

Magnetic forces show a weakness of this simple method very clearly. With $\mathbf{E}=0$ and the velocity perpendicular to **B**, the change $(q/m)\,\mathbf{v}_n\times\mathbf{B}\,\Delta t$ in one step is perpendicular to $\mathbf{v}_n$ and has magnitude $\omega_c\Delta t\,v_n$, so

$$v_{n+1}^2 = v_n^2\left[1+(\omega_c\Delta t)^2\right].$$

Every step increases the speed, and the magnetic field appears to do work. With 63 steps per turn ($\omega_c\Delta t = 0.1$) the speed grows by a factor of 23 in ten turns; with 6300 steps per turn ($\omega_c\Delta t = 0.001$) it still grows by 3%. The error does go to zero as Δt does, but slowly.

A better choice for charged particles is the **Boris method**, a standard algorithm in plasma physics, which replaces the magnetic part of the velocity update by a rotation. With $\mathbf{a} = (q\,\Delta t/2m)\,\mathbf{E}$ and $\mathbf{t} = (q\,\Delta t/2m)\,\mathbf{B}$, each step is

$$\mathbf{v}^- = \mathbf{v}_n + \mathbf{a}, \qquad \mathbf{v}' = \mathbf{v}^- + \mathbf{v}^-\times\mathbf{t}, \qquad \mathbf{v}^+ = \mathbf{v}^- + \frac{2\mathbf{v}'\times\mathbf{t}}{1+t^2},$$
$$\mathbf{v}_{n+1} = \mathbf{v}^+ + \mathbf{a}, \qquad \mathbf{x}_{n+1} = \mathbf{x}_n + \mathbf{v}_{n+1}\,\Delta t.$$



Ernest Lawrence (1901–1958, American) Ernest Orlando Lawrence was born in Canton, South Dakota, and studied chemistry and physics at the University of South Dakota, the University of Minnesota, and the University of Chicago before receiving his PhD from Yale in 1925. He moved to the University of California, Berkeley, in 1928, conceived the cyclotron there in 1929, and directed its Radiation Laboratory from 1936 until his death. The laboratory built ever larger machines and produced radioisotopes, which his brother John, a physician, used in medicine. Lawrence received the Nobel Prize in Physics in 1939 "for the invention and development of the cyclotron and for results obtained with it, especially with regard to artificial radioactive elements." During the Second World War he developed the electromagnetic separation of uranium-235 for the Manhattan Project. Element 103, lawrencium, and the Lawrence Berkeley and Lawrence Livermore National Laboratories are named for him. (Wikipedia)

The step from $\mathbf{v}^-$ to $\mathbf{v}^+$ rotates the velocity about $\mathbf{B}$ without changing its length, so with $\mathbf{E} = 0$ the speed stays constant to within rounding error, whatever the step size. A few tips:

- General-purpose solvers, such as `NDSolve` in Mathematica and `solve_ivp` in Python's `scipy.integrate`, work well for runs of a few turns if you set tight error tolerances. Over thousands of turns they too slowly gain or lose energy, and the Boris method is the better choice.
- Check a simulation against a case you can solve by hand, such as a circle in a uniform field or a cycloid in crossed fields, and check that the speed stays constant when $\mathbf{E} = 0$.
- In Mathematica, `ParametricPlot3D` and `ListPointPlot3D` plot a trajectory in three dimensions; in Python, use the 3-D axes of `matplotlib`.

Exercises P10.4 and P10.5 use these methods, the second for a proton trapped in Earth's field.

10.2 Why the force and the field need cross products

The electric force between two charges points along the line that joins them, and nothing more elaborate than a unit vector along that line is needed to describe it. The magnetic force on a moving charge cannot be described that way, because it is perpendicular to the charge's own velocity. Any force with that property turns out to require a cross product.

Here is the argument. Properties 1 and 2 of Section 10.1.1 say that the force is proportional to q and to the speed, and reverses when either q or $\mathbf{v}$ is reversed. Experiments also show that it obeys superposition in the velocity: the force for a velocity $\mathbf{v}_1 + \mathbf{v}_2$ is the sum of the forces for $\mathbf{v}_1$ and $\mathbf{v}_2$. Together these make the force a *linear* function of $\mathbf{v}$. The most general linear function of a vector is a matrix acting on it, $F_i = q\sum_j M_{ij} v_j$, where the nine numbers M_{ij} can depend on position but not on $\mathbf{v}$. Property 3 says that the force is perpendicular to $\mathbf{v}$ for *every* velocity, so $\mathbf{v}\cdot\mathbf{F} = q\sum_{i,j} v_i M_{ij} v_j = 0$ for all $\mathbf{v}$. Choosing $\mathbf{v}$ along each axis in turn shows that the diagonal elements are zero, and choosing it along $\hat{\mathbf{x}} + \hat{\mathbf{y}}$, and so on, shows that $M_{ij} = -M_{ji}$: the matrix is antisymmetric. An antisymmetric 3×3 matrix has only three independent elements, and written out it is

$$\begin{pmatrix} F_x \\ F_y \\ F_z \end{pmatrix} = q \begin{pmatrix} 0 & B_z & -B_y \\ -B_z & 0 & B_x \\ B_y & -B_x & 0 \end{pmatrix} \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} = q\,\mathbf{v} \times \mathbf{B},$$

where (B_x, B_y, B_z) are names for the three independent elements. So any force that is linear in the velocity and always perpendicular to it has the form $q\mathbf{v} \times \mathbf{B}$ for some vector $\mathbf{B}$ (Exercise P10.6 fills in the details). The direction of motion that produces no force, property 4, is the direction of $\mathbf{B}$, since $\mathbf{v} \times \mathbf{B} = 0$ when $\mathbf{v}$ is

parallel to **B**. The field is a vector, but it enters the force in a way that no electric field does.

The field produced by a current involves a cross product for a similar reason. Experiments on long straight wires (Section 10.5) show that the field circles the wire: at every point it is perpendicular both to the wire and to the line from the wire to the point. Suppose the field of each short piece of the wire, $I\,d\boldsymbol{\ell}'$, has the same property, so that it is perpendicular both to $d\boldsymbol{\ell}'$ and to the separation vector $\boldsymbol{\imath}$ from the piece to the field point. The only vector that can be built from $d\boldsymbol{\ell}'$ and $\boldsymbol{\imath}$ and is perpendicular to both is a multiple of $d\boldsymbol{\ell}' \times \boldsymbol{\imath}$, and that is the form the Biot–Savart law takes. The magnetic field therefore enters physics through two cross products, one in the way currents produce it and one in the way it acts on moving charges.

One consequence is that the force between two short pieces of current, which involves both cross products, does not in general point along the line joining them, as the force between two charges does.[1] There is no simple magnetic counterpart of Coulomb's law for the force between two currents, and that is the main reason the field is so useful in magnetism: it splits a complicated problem into two simpler steps.

10.3 The magnetic field and field lines

The magnetic field, like the electric field, is a vector field: a vector **B** at every point in space. Table 10.1 lists the size of the field in some familiar and some less familiar places.[2]

Magnetic fields are pictured with field lines, drawn by the same rules as electric field lines (Section 3.3): a field line is tangent to **B** at every point, and the lines are drawn closer together where the field is stronger. Iron filings sprinkled near a magnet or a wire carrying a strong current line up along the field, which makes the lines visible. Figure 10.7 shows three examples, calculated with the methods of this chapter. The field of a long straight wire circles the wire, and its strength falls off as $1/s$ with the distance s from the wire (Section 10.6.1), so the circles get farther apart going out. The field of a circular loop of wire passes up through the loop and returns around the outside; far from the loop it has the

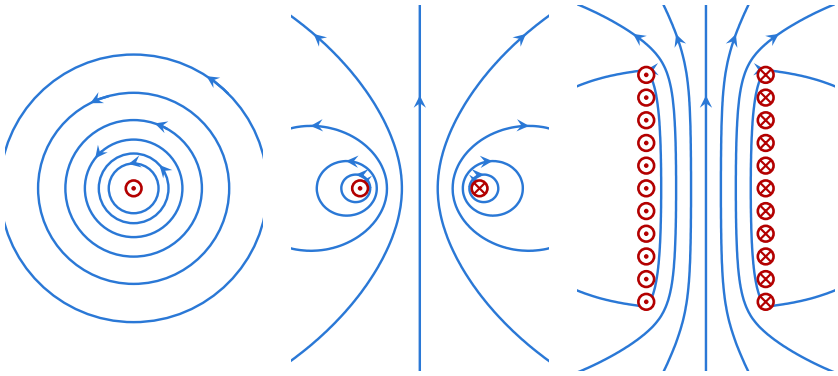
	B
Earth's surface	22–67 μT
Provo, Utah	50 μT
1 cm from a 10 A wire	0.2 mT
Refrigerator magnet	1–10 mT
Surface of a strong magnet	0.5–1 T
Clinical MRI scanner	1.5–3 T
Strongest steady laboratory field	45 T
Strongest nondestructive pulsed field	100 T
Surface of a neutron star	10^8 T
Magnetars, the most strongly magnetized neutron stars, up to	10^{11} T

Table 10.1 Some magnetic field strengths; the last four are approximate (sources in the footnote).

[1] Ampère wrote a force law for pieces of current that does point along the line joining them. The form implied by the Biot–Savart and Lorentz force laws was given by Hermann Grassmann in 1845. For complete circuits, the only kind that carry steady currents, the two give exactly the same total force.

[2] Earth's field: the smallest value, in the South Atlantic Anomaly, is from *December 2025 State of the Geomagnetic Field*, NOAA National Centers for Environmental Information and British Geological Survey, https://repository.library.noaa.gov/view/noaa/72269; the largest value and the value at Provo are from the World Magnetic Model WMM2025, which NOAA's online Magnetic Field Calculator evaluates, https://www.ngdc.noaa.gov/geomag/calculators/magcalc.shtml. Laboratory fields: National High Magnetic Field Laboratory, https://nationalmaglab.org/about-the-maglab/organization/history/timeline/. Neutron stars and magnetars: NASA High Energy Astrophysics Science Archive Research Center, https://heasarc.gsfc.nasa.gov/docs/asca/science/magnetar.html. The values for magnets and MRI scanners are typical ones.

same shape as the field of an electric dipole. A long coil of wire, a **solenoid**, has a nearly uniform field inside, and the lines that leave one end spread out through all of space before they return to the other, so the field outside is weak.



(a) long straight wire (b) circular loop (c) finite solenoid

Figure 10.7 Magnetic field lines of (a) a long straight wire carrying current out of the page, (b) a circular loop, and (c) a solenoid of length four times its radius. In (b) and (c) the lines are drawn in a plane containing the axis, the current comes out of the page on the left and goes into it on the right, and the axis is a field line too. In a plane like this the spacing of the lines shows the strength of the field only roughly, except inside the solenoid, where the field is nearly uniform.

Magnetic field lines differ from electric field lines in two important ways.

- **Magnetic field lines never begin or end.** Every line in Figure 10.7 is a closed loop or, like the axis of the loop, runs off to infinity in both directions. There are no magnetic charges for field lines to begin and end on, the way electric field lines begin and end on electric charges. This is the statement $\nabla\cdot\mathbf{B} = 0$, derived in Section 11.1.1, and it is the oldest of Maxwell's equations (Section 2.4). (In complicated fields a line can wind around forever without closing, but it still never ends.)

☞ *Caution*: Magnetic field lines are not lines of force. The magnetic force on a moving charge is perpendicular to **B**.

- **Magnetic field lines are not lines of force.** The electric force on a charge points along the electric field line through it. The magnetic force on a moving charge is perpendicular to **B**, so it points *across* the field lines, and a charged particle does not move along them, although in a strong field it spirals around them (Section 10.1.4).

10.4 Currents and current densities

A **current** is a flow of charge. The current in a wire is the charge per unit time that passes any point along it,

$$I = \frac{dq}{dt},$$

measured in amperes (1 A = 1 C/s). If a line of charge with line charge density λ moves along its length with speed v, then in a time dt a length $v\,dt$ of it passes a given point, carrying a charge $\lambda v\,dt$, so

$$I = \lambda v. \tag{10.7}$$

In a metal wire the moving charges are electrons, and they move opposite to the direction of the current, while the positive ions stay put; the wire as a whole is neutral. By convention, the direction of a current is the direction in which positive charge flows, or would flow. For the magnetic force on a wire and the magnetic field it produces, it makes no difference: a negative charge moving one way has the same $q\mathbf{v}$, and so feels the same magnetic force and produces the same magnetic field, as a positive charge moving the other way. (The Hall effect, at the end of this section, is the exception.) For now we care only about the overall flow, not about the individual charges that carry it. (The fields of individual moving charges are the subject of Chapter 27.)

A current has a direction as well as a size, and the combination that appears in magnetism is $I\,d\boldsymbol{\ell}$, the current times a short piece of the wire, with $d\boldsymbol{\ell}$ pointing along the current. It is the moving-charge counterpart of dq: the piece carries a charge $dq = \lambda\,d\ell$ moving with velocity $\mathbf{v}$ parallel to $d\boldsymbol{\ell}$, so

$$dq\,\mathbf{v} = \lambda v\,d\boldsymbol{\ell} = I\,d\boldsymbol{\ell}.$$

When charge flows over a surface, or through a volume, the current is described by a density, just as charge is (Figure 10.8). A surface charge density σ moving with velocity $\mathbf{v}$ gives a **surface current density**

$$\mathbf{K} = \sigma\mathbf{v}, \tag{10.8}$$

in amperes per meter. Its magnitude is the current per unit length measured *across* the flow. The current that crosses a line drawn on the surface is

$$I = \int K_\perp\,d\ell, \tag{10.9}$$

where $K_\perp$ is the component of $\mathbf{K}$ perpendicular to the line. Similarly, a volume charge density ρ moving with velocity $\mathbf{v}$ gives a **volume current density**

$$\mathbf{J} = \rho\mathbf{v}, \tag{10.10}$$

in amperes per square meter: the current per unit area measured perpendicular to the flow. The current through a surface $\mathcal{S}$ is the flux of $\mathbf{J}$ through it,

$$I = \int_{\mathcal{S}} \mathbf{J}\cdot d\mathbf{a}, \tag{10.11}$$

since only the component of $\mathbf{J}$ perpendicular to the surface carries charge through it. To find a current, then, integrate over a line or a surface that the flow crosses, so as to "catch the current." Integrating over a surface that the current runs along

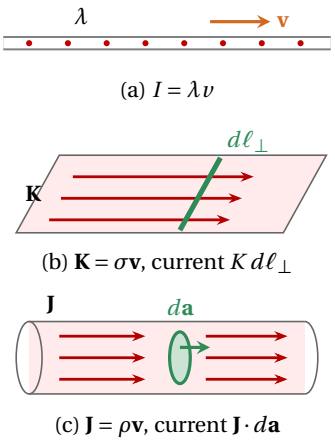


Figure 10.8 (a) A line current. (b) A surface current, with surface current density $\mathbf{K}$; the current crossing a short segment $d\ell_\perp$ drawn across the flow is $K\,d\ell_\perp$. (c) A volume current, with volume current density $\mathbf{J}$; the current through an area element $d\mathbf{a}$ is $\mathbf{J}\cdot d\mathbf{a}$.

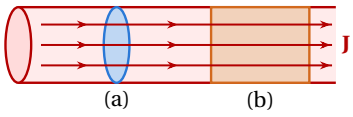


Figure 10.9 Catching the current in a wire. All of the current crosses the cross section (a); the rectangle (b), in a plane through the axis, lies along the flow and catches none of it.

instead of through catches none of it (Figure 10.9). Notice also that the units are not quite parallel to those of the charge densities: σ is in C/m^2 but $\mathbf{K}$ is in A/m, and ρ is in C/m^3 but $\mathbf{J}$ is in A/m^2. A current already involves one dimension, the direction of flow, so a current density needs one fewer power of length than the corresponding charge density.

The short pieces of current that appear in the magnetic formulas of this chapter are

$$I\,d\boldsymbol{\ell}, \qquad \mathbf{K}\,da, \qquad \mathbf{J}\,d\tau,$$

the counterparts of $\lambda\,d\ell$, $\sigma\,da$, and $\rho\,d\tau$. In each, the vector describes the flow, and the length, area, or volume element describes the piece.

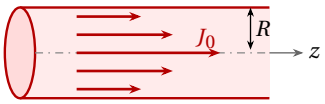
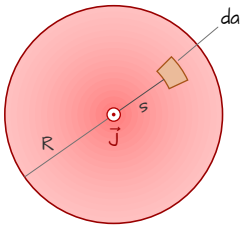


Figure 10.10 The wire of Example 10.4. The current density is largest on the axis and falls off toward the surface.

Example 10.4

A long wire of radius R carries a current whose density is greatest on its axis, $\mathbf{J} = J_0\,e^{-s/R}\,\hat{\mathbf{z}}$, where s is the distance from the axis. What is the total current?

Solution:



We sketch the cross section of the wire, shaded darker where the current density is larger, near the axis. The current flows along $\hat{\mathbf{z}}$, so the surface that catches it is a cross section of the wire, a disk of radius R with $d\mathbf{a} = s\,ds\,d\phi\,\hat{\mathbf{z}}$.

By (10.11),

$$\begin{aligned} I &= \int \mathbf{J}\cdot d\mathbf{a} \\ &= \int_0^{2\pi}\int_0^R J_0\,e^{-s/R}\,s\,ds\,d\phi \\ &= 2\pi J_0 \int_0^R s\,e^{-s/R}\,ds. \end{aligned}$$

The integral can be done by parts, or by computer:

$$\int_0^R s\,e^{-s/R}\,ds = \left[-Rs\,e^{-s/R} - R^2 e^{-s/R}\right]_0^R = R^2\left(1-\frac{2}{e}\right).$$

So

$$I = 2\pi J_0 R^2\left(1-\frac{2}{e}\right) = 1.66\,J_0 R^2.$$

Reality check: A uniform current density J_0 over the whole cross section would carry $\pi R^2 J_0 = 3.14\,J_0 R^2$. Our density equals J_0 only on the axis and falls to J_0/e at the surface, so the current should be smaller, and it is, by about half. The units are also right: amperes per square meter times square meters.

10.4.1 Steady currents

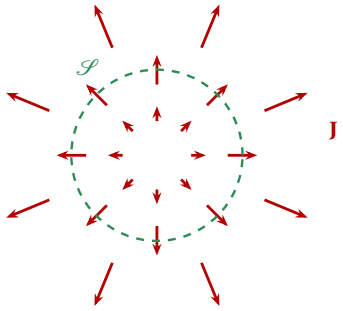
Charge is conserved: it cannot be created or destroyed, only moved from place to place. The local form of this law is the **continuity equation** (Section 20.1),

$$\nabla \cdot \mathbf{J} = -\frac{\partial \rho}{\partial t}. \qquad \text{(continuity equation)} \quad (10.12)$$

With words, wherever current spreads out from a point, so that $\nabla \cdot \mathbf{J} > 0$, the charge density there decreases, because the charge that leaves has to come from somewhere (Figure 10.11). Magnetostatics deals with **steady currents**, currents whose pattern of flow does not change in time. The charges keep moving, but the charge density stays the same everywhere, so $\partial\rho/\partial t = 0$ and

$$\nabla \cdot \mathbf{J} = 0. \qquad \text{(steady currents)} \quad (10.13)$$

As much current flows into any region as flows out of it. In particular, the same current flows through every cross section of a wire, and steady currents flow around closed circuits. Where a wire narrows, the same current has to pass through a smaller cross section, so $\mathbf{J}$ is larger there and its lines crowd together. That does *not* make $\nabla \cdot \mathbf{J}$ nonzero: $\mathbf{J}$ grows along the narrowing part of the wire, but every line of $\mathbf{J}$ that enters a small box there also leaves it (Figure 10.12). Lines that crowd together, like lines that spread apart, are not a sign of divergence (Section 1.4.2). A current that charges a capacitor is not steady, because charge piles up on the plates; neither is the "current" of a single moving charge, whose charge density moves along with it. Such currents are taken up in Chapters 16 and 27.



$\nabla \cdot \mathbf{J} > 0$, so $\partial\rho/\partial t < 0$

Figure 10.11 Current spreading out from a region, so that $\nabla \cdot \mathbf{J} > 0$ there. Current flows out through the closed surface $\mathscr{S}$ (dashed) all the way around, and none flows in, so the charge inside must decrease.

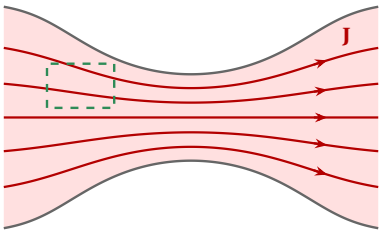


Figure 10.12 The lines of $\mathbf{J}$ in a flat conducting strip with a neck, seen from above, drawn with the same current between each pair of neighboring lines, so that they crowd together where $\mathbf{J}$ is larger. One line enters the small box (dashed) through its left side and one through its top, and two leave through its right side: $\mathbf{J}$ is larger on the right, but as much current leaves the box as enters it, and $\nabla \cdot \mathbf{J} = 0$ because there are no sources of current inside the box.

10.4.2 Forces on currents

The magnetic force on a piece of a current-carrying wire is the force on the charges moving in it. Replacing $q\mathbf{v}$ in (10.1) by $dq\,\mathbf{v} = I\,d\boldsymbol{\ell}$ and adding up the pieces gives

$$\mathbf{F} = \int I\,d\boldsymbol{\ell} \times \mathbf{B}, \qquad (10.14)$$

and for surface and volume currents

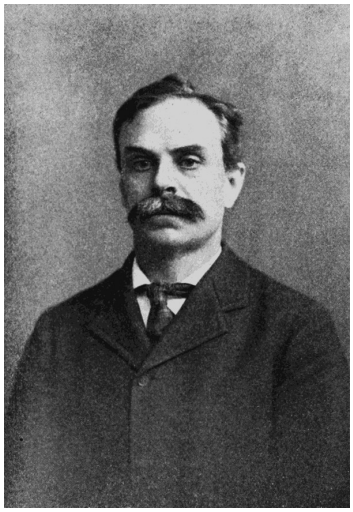
$$\mathbf{F} = \int \mathbf{K} \times \mathbf{B}\,da, \qquad \mathbf{F} = \int \mathbf{J} \times \mathbf{B}\,d\tau. \qquad (10.15)$$

For a straight wire carrying a current I in a uniform field, $d\boldsymbol{\ell}$ always points the same way, and (10.14) becomes $\mathbf{F} = I\mathbf{L} \times \mathbf{B}$, (10.3) of Section 10.1.2. In fact, in a uniform field $\mathbf{B}$ comes out of the integral for a wire of any shape, $\mathbf{F} = I\left(\int d\boldsymbol{\ell}\right) \times \mathbf{B}$, and $\int d\boldsymbol{\ell}$ is simply the vector from one end of the wire to the other. A closed loop of any shape in a uniform field therefore feels no net force at all,

$$\mathbf{F} = I\left(\oint d\boldsymbol{\ell}\right) \times \mathbf{B} = 0,$$

because the pieces $d\ell$ of a closed loop add up to zero. (Such a loop does feel a torque, which is what turns an electric motor; Section 13.1.1 takes it up.)

The magnetic force on a wire acts on the moving charges inside it, and those charges are pushed toward one side of the wire. They pile up there until the electric field of the piled-up charge pushes back just as hard on the moving charges, and it is through this electric field that the force is passed on to the rest of the wire (Figure 10.5). The field is perpendicular to the current, so it produces a small voltage across the wire, the **Hall voltage**, discovered by Edwin Hall in 1879. Its sign reveals the sign of the charges that carry the current, and its size their density. Hall probes, which measure magnetic fields this way, are a common laboratory instrument (Exercise P10.10).



Edwin Hall (1855–1938, American) Edwin Herbert Hall was born in Gorham, Maine, graduated from Bowdoin College in 1875, and in 1877 began graduate work in physics under Henry Rowland at the new Johns Hopkins University. Maxwell had written in his *Treatise* that the magnetic force on a wire acts on the conductor, not on the current itself, and Hall set out to test this. Thicker metal plates showed nothing, but on Rowland's suggestion he sent the current through gold leaf, and in October 1879 he found the transverse voltage now named for him. He received his doctorate in 1880 and taught at Harvard from 1881 until 1921, from 1914 as Rumford Professor of Physics. He worked for decades on conduction and thermoelectricity in metals, and at the request of Harvard's president he wrote a list of forty laboratory exercises (1886) that let secondary schools meet Harvard's new laboratory requirement for admission. (Wikipedia)

Forces on currents also show how magnetic forces take part in doing work without doing any themselves (Section 10.1.3). Take a straight horizontal piece of wire of length ℓ, carrying a current I to the right through a uniform field $\mathbf{B}$ that points into the page, with the rest of the circuit outside the field. The magnetic force on the piece is $I\ell B$, straight up, and if it is larger than the weight of the wire, the wire rises and gains energy. Suppose the wire moves up with speed u. Each charge q in it then has two velocities, its drift velocity w along the wire and the velocity u of the wire, and the magnetic force on it has two parts: qwB up, which lifts the wire, and quB along the wire, *opposing* the current. The total magnetic force on the charge is perpendicular to its total velocity, and does no work on it. But to keep the current flowing, the battery that drives it must push each charge along the wire with an extra force quB. The moving charges in the piece have a total charge $\lambda\ell$, where λ counts only the moving charges (the wire as a whole is neutral), and they all move at the drift speed w, so the battery does extra work at the rate

$$(\lambda\ell)\,uB\,w = (\lambda w)\,\ell B u = I\ell B\,u,$$

exactly the rate Fu at which the magnetic force on the wire does work lifting it. The energy comes from the battery; the magnetic force only redirects it. The same argument applies to a wire of Example 10.2 as it moves toward the other: the source of its current supplies the kinetic energy it gains. (The source of the other wire's current must supply energy too, an equal amount, which goes into the magnetic field. That takes Faraday's law, Section 15.2; Exercise P19.11 works out the energy for two loops.)

10.5 The Biot–Savart law

We turn now to the second question: how do currents produce magnetic fields? Ørsted's discovery reached Paris in September 1820, when François Arago described it to the Académie des Sciences and then repeated the experiment there. Ampère's experiments on the forces between currents followed within weeks, and on 30 October Jean-Baptiste Biot and Félix Savart reported measurements of the force that a long straight current exerts on a magnetized needle. They timed the needle's small oscillations at different distances from the wire and found that the

force on it was perpendicular both to the wire and to the line from the wire to the needle, and inversely proportional to the distance. Pierre-Simon Laplace showed that a force falling off as the inverse distance from a long wire is what you get if each short piece of the wire contributes a force that falls off as the inverse *square* of its distance, and Biot stated the law for a single piece.

In modern form, the law says that a short piece of current $I\,d\boldsymbol{\ell}'$ at the source point $\mathbf{r}'$ produces at the field point $\mathbf{r}$ the field

$$d\mathbf{B} = \frac{\mu_0}{4\pi}\frac{I\,d\boldsymbol{\ell}' \times \hat{\boldsymbol{\imath}}}{\imath^2},$$

where, as in Chapter 3, $\boldsymbol{\imath} = \mathbf{r} - \mathbf{r}'$ is the separation vector from the source point to the field point (Figure 10.13). Adding up the contributions of all the pieces of a wire gives the **Biot–Savart law**,

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi}\int \frac{I\,d\boldsymbol{\ell}' \times \hat{\boldsymbol{\imath}}}{\imath^2} = \frac{\mu_0 I}{4\pi}\int \frac{d\boldsymbol{\ell}' \times \boldsymbol{\imath}}{\imath^3}, \tag{10.16}$$

where the integral runs along the wire in the direction of the current. For surface and volume currents, $I\,d\boldsymbol{\ell}'$ is replaced by $\mathbf{K}\,da'$ or $\mathbf{J}\,d\tau'$:

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi}\int \frac{\mathbf{K}(\mathbf{r}') \times \hat{\boldsymbol{\imath}}}{\imath^2}\,da', \tag{10.17}$$

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi}\int \frac{\mathbf{J}(\mathbf{r}') \times \hat{\boldsymbol{\imath}}}{\imath^2}\,d\tau'. \tag{10.18}$$

The constant μ_0 is the **permeability of free space**, also called the magnetic constant. Its value is

$$\mu_0 = 4\pi \times 10^{-7}\,\mathrm{N/A^2} = 1.2566 \times 10^{-6}\,\mathrm{N/A^2}; \tag{10.19}$$

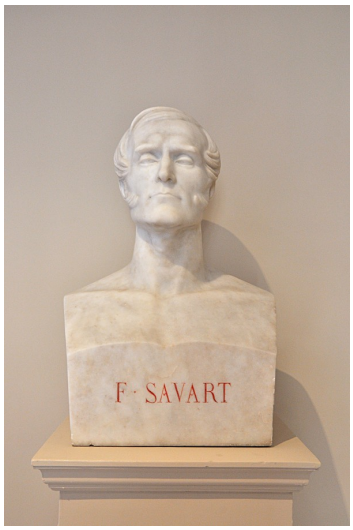
until 2019 this was its exact value by definition, and it is now a measured quantity that agrees with $4\pi \times 10^{-7}$ to about one part in 10^{10} (Section 17.4). Its unit can also be written $\mathrm{T \cdot m/A}$, which makes it easy to check that (10.16) gives a field in teslas. The two constants ε_0 and μ_0 combine to give a speed, $1/\sqrt{\varepsilon_0\mu_0} = 3.00 \times 10^8\,\mathrm{m/s}$, the speed of light. That is not a coincidence, as Section 16.5 will show.

The Biot–Savart law is the magnetic counterpart of Coulomb's law in the form (3.11). Both are inverse-square laws, and both build the field of an extended source by superposition, adding up the fields of its small pieces. The difference is that the source of a magnetic field is a vector, $I\,d\boldsymbol{\ell}'$, and it enters through a cross product. With words, each short piece of current produces a field that circles the line along which the piece points, perpendicular both to the piece and to the line from the piece to the field point, with a strength proportional to $\sin\alpha/\imath^2$, where α is the angle between $d\boldsymbol{\ell}'$ and $\boldsymbol{\imath}$. A field point straight ahead of a piece, or straight behind it, gets no field from that piece at all.

A few cautions about using the law:



Jean-Baptiste Biot (1774–1862, French) Biot studied at the École Polytechnique, and in 1800 he became professor of mathematical physics at the Collège de France. His report on a fall of stones near L'Aigle, in Normandy, in 1803 helped convince scientists that meteorites really do fall from the sky, and in 1804 he and Joseph-Louis Gay-Lussac rose to about 4000 m in a balloon to measure Earth's magnetic field at altitude. From 1815 he studied the way some liquids rotate the plane of polarization of light, the basis of the technique called polarimetry. In 1820, with Félix Savart, he showed that the magnetic force of a long straight current is inversely proportional to the distance from it, and with the help of a calculation by Laplace he went on to state the law for the magnetic action of a short piece of current that now carries their names. (Wikipedia)



Félix Savart (1791–1841, French) Savart trained as a physician, at the military hospital in Metz and at the University of Strasbourg, where he received his medical degree in 1816. His interest turned to acoustics: he built an experimental violin, and in 1819 he presented a memoir on stringed instruments to the Académie des Sciences. In Paris, Jean-Baptiste Biot helped him find a post teaching physics, and in 1820 the two measured how the magnetic force of a current-carrying wire depends on the distance from it. Savart was elected to the Académie in 1827 and taught at the Collège de France. Around 1830 he built a rotating toothed wheel that produced tones of known frequency, which he used to find the lowest frequency that can be heard. (Wikipedia)

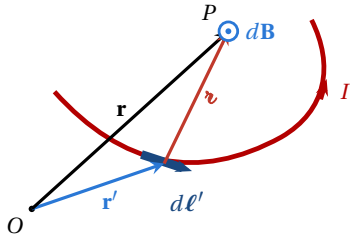


Figure 10.13 The Biot–Savart law. A short piece $d\boldsymbol{\ell}'$ of a wire carrying a steady current I, at the source point $\mathbf{r}'$, contributes $d\mathbf{B}$ at the field point P. The contribution is perpendicular to both $d\boldsymbol{\ell}'$ and $\boldsymbol{\mathcal{r}}$; here it points out of the page.

- It holds exactly for steady currents, just as Coulomb's law holds exactly for charges at rest. For currents that change slowly it is still an excellent approximation (Section 16.4), even when they are not steady, as for a charging capacitor (Section 16.3), and Example 24.3 shows when it is exact. A single moving point charge is not a steady current, and the formula $(\mu_0/4\pi)\, q\mathbf{v}\times\hat{\boldsymbol{\mathcal{r}}}/\mathcal{r}^2$ that (10.16) suggests for it is only an approximation, good for speeds much less than the speed of light. The exact fields of a charge moving at constant velocity are found in Section 26.7, and those of a charge moving in any way in Chapter 27.

- The integral runs over *all* of the current. For a surface or volume current, integrate over the whole area or volume that carries it, not over a line or surface that the current crosses, which is what you use to find the current itself with (10.9) or (10.11).

- For steady currents a wire must form a closed circuit, and the field of a single straight segment or a single wire running off to infinity is really the contribution of that part of a circuit whose other parts are far away or are accounted for separately. The fields of such pieces are still useful building blocks, as the next section shows.

The law always works, at least in principle. The difficulty is in doing the integrals, which are integrals of vectors with cross products in them, and the next section shows how to set them up.

10.6 Solving for B when given J: script-r method

The Biot–Savart law is a blueprint, like Coulomb's law in Section 3.5: it says how to construct the integral for a particular current, and the construction follows nearly the same steps.

1. Draw a picture, mark a piece of the current at an arbitrary location (not at a special point such as an end or a corner), and draw $\mathbf{r}$, $\mathbf{r}'$, and $\boldsymbol{\mathcal{r}}$. Use symmetry and the right-hand rule to find the direction of $\mathbf{B}$ before calculating anything, if you can.

2. Write down $\mathbf{r}$, $\mathbf{r}'$, $\boldsymbol{\mathcal{r}} = \mathbf{r}-\mathbf{r}'$, and $\mathcal{r}$, each on a line of its own, and the element $d\boldsymbol{\ell}'$, da', or $d\tau'$. The vector $d\boldsymbol{\ell}'$ points along the current. The element is the size of one piece of the current (a piece of a ring of radius a has length $a\,d\phi'$, not $d\phi'$), and $I\,d\boldsymbol{\ell}'$, $\mathbf{K}\,da'$, or $\mathbf{J}\,d\tau'$ is the moving charge in that piece times its velocity, $dq\,\mathbf{v}$ (Section 10.4).

3. Write the current density in terms of the primed variables, and check that its integral over a cross section of the current is the total current I. Then set the limits: integrate over the source, counting each piece of it once.

4. Form the cross product, $d\boldsymbol{\ell}'\times\boldsymbol{\mathcal{r}}$, $\mathbf{K}\times\boldsymbol{\mathcal{r}}$, or $\mathbf{J}\times\boldsymbol{\mathcal{r}}$.

5. Do the integrals, one for each Cartesian component.
6. Check the result: its direction, its units (a combination like $\mu_0 I$/length is a field in teslas), and its limits.

Getting $\mathbf{r}$, $\mathbf{r}'$, $\boldsymbol{\mathscr{r}}$, $\mathscr{r}$, and $d\boldsymbol{\ell}'$ right is at least half the battle. Once the integral is set up correctly, the rest is mathematics.

Colton's First Commandment (Section 3.5.1) applies with full force, since the cross products are full of unit vectors. Any unit vector that depends on the variables of integration, such as $\hat{\mathbf{s}}'$ or $\hat{\boldsymbol{\phi}}'$, must be written in Cartesian components before it is integrated. There are two ways to obey the Commandment. You can convert every unit vector to Cartesian components first and then take the cross product, or you can take the cross product first, using the rules for cylindrical or spherical unit vectors in Section 10.1.1, and then convert whatever variable unit vectors are left. The second way is often shorter, as the ring example below shows.

10.6.1 Worked examples

Example 10.5

A long straight wire on the y axis carries a current I in the $+y$ direction. Find the magnetic field at a point on the z axis, a distance z from the wire.

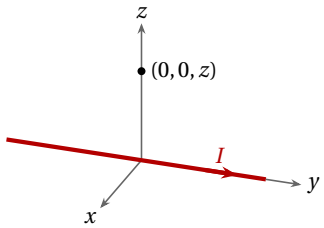
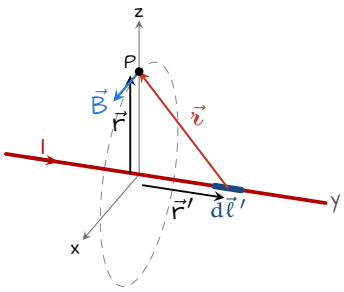


Figure 10.14 Example 10.5: the wire on the y axis and the field point on the z axis.

Solution:



We draw the wire, a piece of it at an arbitrary place, the field point, and the vectors $\mathbf{r}$ (to the field point), $\mathbf{r}'$ (to the piece) and $\boldsymbol{\mathscr{r}}$. First the direction. With your right thumb along $+\hat{\mathbf{y}}$, your fingers curl through the point $(0,0,z)$ in the $+x$ direction, so we expect $\mathbf{B}$ to point along $\hat{\mathbf{x}}$ for $z>0$.

The pieces of the blueprint are

$$\begin{aligned}\mathbf{r} &= z\hat{\mathbf{z}}\\ \mathbf{r}' &= y'\hat{\mathbf{y}}\\ \boldsymbol{\mathscr{r}} &= \mathbf{r}-\mathbf{r}' = z\hat{\mathbf{z}} - y'\hat{\mathbf{y}}\\ \mathscr{r} &= \sqrt{z^2+y'^2},\end{aligned}$$

and $d\boldsymbol{\ell}' = dy'\,\hat{\mathbf{y}}$. The cross product is

$$\begin{aligned}d\boldsymbol{\ell}'\times\boldsymbol{\mathscr{r}} &= dy'\,\hat{\mathbf{y}}\times\left(z\hat{\mathbf{z}} - y'\hat{\mathbf{y}}\right)\\ &= z\,dy'\,(\hat{\mathbf{y}}\times\hat{\mathbf{z}}) - y'\,dy'\,(\hat{\mathbf{y}}\times\hat{\mathbf{y}})\\ &= z\,dy'\,\hat{\mathbf{x}},\end{aligned}$$

since the cross product of a vector with itself is zero. Every piece of the wire contributes a field in the same direction, so nothing cancels, and by (10.16)

$$\mathbf{B} = \frac{\mu_0 I}{4\pi}\int_{-\infty}^{\infty}\frac{z\,dy'\,\hat{\mathbf{x}}}{\left(z^2+y'^2\right)^{3/2}} = \frac{\mu_0 I z\hat{\mathbf{x}}}{4\pi}\int_{-\infty}^{\infty}\frac{dy'}{\left(z^2+y'^2\right)^{3/2}}.$$

This is the integral of Example 3.4, in the limit of an infinitely long line, and it equals $2/z^2$, so

$$\mathbf{B} = \frac{\mu_0 I}{2\pi z}\hat{\mathbf{x}}.$$

Reality check: The field points along $\hat{\mathbf{x}}$, as the right-hand rule predicted in the sketch, and $\mu_0 I/2\pi z$ has the units of $\mu_0 I$ divided by a length, a field in teslas.

The integral runs from $-\infty$ to ∞ because that is the direction of the current. As in Section 5.4, write $d\boldsymbol{\ell}'$ in its standard form, $dy'\,\hat{\mathbf{y}}$, whichever way the current flows, and let the limits set the direction: for a current in the $-y$ direction they would run from ∞ to $-\infty$, with $d\boldsymbol{\ell}'$ still $dy'\,\hat{\mathbf{y}}$. Putting a minus sign into $d\boldsymbol{\ell}'$ as well would count the direction twice and give the wrong sign.

For a wire on the z axis, carrying I in the $+z$ direction, the same calculation at a distance s from the wire gives a field that circles it:

$$\mathbf{B} = \frac{\mu_0 I}{2\pi s}\hat{\boldsymbol{\phi}}. \tag{10.20}$$

The field of the infinite wire falls off as $1/s$, not $1/s^2$, for the same reason that the field of an infinite line of charge does (Example 3.4): the farther you are from the wire, the longer the part of it that is not much farther away than the nearest point. Equation (10.20) is the magnetic field you will use most often, and it is worth remembering, along with its direction. It is the field that Example 10.2 used to find the force between two long parallel wires, (10.4).

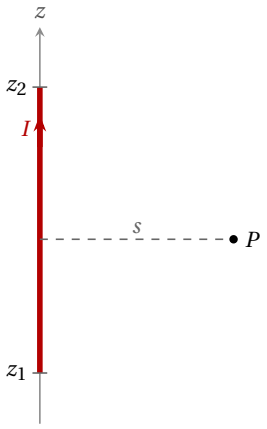
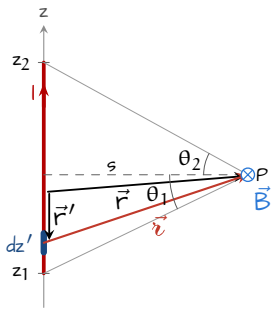


Figure 10.15 Example 10.6: a straight segment of current on the z axis, from z_1 to z_2, and the field point P, a distance s from the axis.

Example 10.6

A straight wire runs along the z axis from z_1 to z_2 and carries a current I in the $+z$ direction. (The rest of its circuit is far away.) Find the magnetic field at a distance s from the axis and height z (Figure 10.15).

Solution:



Following the blueprint, we draw the segment, a piece dz' of it at height z', the vectors $\mathbf{r}$ (to P) and $\mathbf{r}'$ (to the piece), the separation vector from the piece to P, and the angles θ_1 and θ_2 at which P sees the two ends. By the right-hand rule, the field at P points into the page.

In cylindrical coordinates centered on the wire,

$$\mathbf{r} = s\hat{\mathbf{s}} + z\hat{\mathbf{z}}$$
$$\mathbf{r}' = z'\hat{\mathbf{z}}$$
$$\boldsymbol{\mathcal{r}} = \mathbf{r} - \mathbf{r}' = s\hat{\mathbf{s}} + (z - z')\hat{\mathbf{z}}$$
$$\mathcal{r} = \sqrt{s^2 + (z - z')^2},$$

and $d\boldsymbol{\ell}' = dz'\hat{\mathbf{z}}$. The cross product is $d\boldsymbol{\ell}' \times \boldsymbol{\mathcal{r}} = s\,dz'(\hat{\mathbf{z}} \times \hat{\mathbf{s}}) = s\,dz'\hat{\boldsymbol{\phi}}$, and $\hat{\boldsymbol{\phi}}$ can be taken outside the integral. (Why does taking $\hat{\boldsymbol{\phi}}$ outside the integral not break Colton's First Commandment? Answer in footnote.[3]) So

$$\mathbf{B} = \frac{\mu_0 I}{4\pi}\hat{\boldsymbol{\phi}}\int_{z_1}^{z_2}\frac{s\,dz'}{\left[s^2 + (z - z')^2\right]^{3/2}} = \frac{\mu_0 I}{4\pi}\hat{\boldsymbol{\phi}}\left[\frac{z' - z}{s\sqrt{s^2 + (z - z')^2}}\right]_{z_1}^{z_2}.$$

The two terms are the sines of the angles θ_2 and θ_1 in the sketch, measured from the perpendicular dropped from P to the wire, positive in the direction of the current (θ_1 in the sketch is negative), so

$$\mathbf{B} = \frac{\mu_0 I}{4\pi s}(\sin\theta_2 - \sin\theta_1)\hat{\boldsymbol{\phi}}. \tag{10.21}$$

Reality check: For an infinite wire, $\theta_2 = 90°$ and $\theta_1 = -90°$, and (10.21) becomes (10.20), the field of Example 10.5.

At a point opposite the end of a wire that runs to infinity in one direction, the field is half of that. And at the center of a square loop of side w, each side is a distance $w/2$ away and is seen at angles of $\pm 45°$, so each side contributes $(\mu_0 I/2\pi w)(2\sin 45°)$, and the four sides together give

$$B = \frac{2\sqrt{2}\,\mu_0 I}{\pi w} = 0.90\,\frac{\mu_0 I}{w},$$

perpendicular to the loop.

[3] Here $\hat{\mathbf{s}}$ and $\hat{\boldsymbol{\phi}}$ belong to the *field* point, not to the piece of the wire. They do not depend on the variable of integration, z', so they are not the variable unit vectors that the Commandment is about.

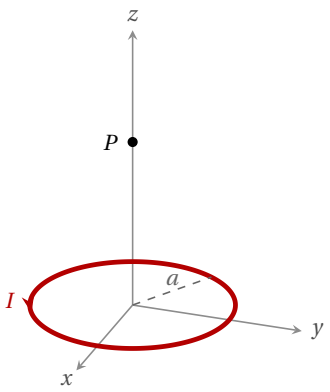
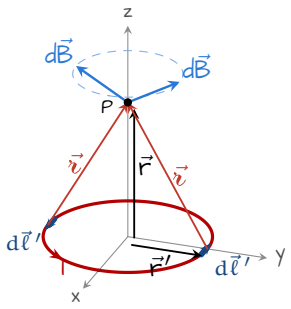


Figure 10.16 Example 10.7: a ring of radius a in the xy plane, carrying a current I counterclockwise as seen from above, and a field point P on the z axis.

Example 10.7

A ring of radius a in the xy plane, centered on the origin, carries a current I, counterclockwise as seen from above. Find the magnetic field at a point on the z axis.

Solution:



We copy the ring of Figure 10.16 and add two pieces of it on opposite sides, with the vectors $\mathbf{r}$, $\mathbf{r}'$ and $\boldsymbol{\mathscr{r}}$ for one of them and the contributions of both at P. First, the direction. Each piece of the ring contributes a field perpendicular both to the piece and to $\boldsymbol{\mathscr{r}}$, which tilts up and toward the side of the ring where the piece is, as drawn. The piece on the opposite side of the ring contributes a field tilted the other way, and their sideways parts cancel, so the total field is along $\hat{\mathbf{z}}$.

The pieces of the blueprint are

$$\begin{aligned}
\mathbf{r} &= z\hat{\mathbf{z}} \\
\mathbf{r}' &= a\hat{\mathbf{s}}' = a\left(\cos\phi'\,\hat{\mathbf{x}} + \sin\phi'\,\hat{\mathbf{y}}\right) \\
\boldsymbol{\mathscr{r}} &= \mathbf{r} - \mathbf{r}' = z\hat{\mathbf{z}} - a\hat{\mathbf{s}}' \\
&= -a\cos\phi'\,\hat{\mathbf{x}} - a\sin\phi'\,\hat{\mathbf{y}} + z\hat{\mathbf{z}} \\
\mathscr{r} &= \sqrt{z^2 + a^2},
\end{aligned}$$

and $d\boldsymbol{\ell}' = a\,d\phi'\,\hat{\boldsymbol{\phi}}'$. The unit vector $\hat{\mathbf{s}}'$ is the one that points from the axis to the piece; it is primed because it varies with ϕ' as the piece goes around the ring. The magnitude $\mathscr{r}$ follows from the Pythagorean theorem, since $\hat{\mathbf{z}}$ and $\hat{\mathbf{s}}'$ are perpendicular, and it is the same for every piece. Taking the cross product first, with $\hat{\boldsymbol{\phi}} \times \hat{\mathbf{z}} = \hat{\mathbf{s}}$ and $\hat{\boldsymbol{\phi}} \times \hat{\mathbf{s}} = -\hat{\mathbf{z}}$,

$$d\boldsymbol{\ell}' \times \boldsymbol{\mathscr{r}} = a\,d\phi'\,\hat{\boldsymbol{\phi}}' \times \left(z\hat{\mathbf{z}} - a\hat{\mathbf{s}}'\right) = az\,d\phi'\,\hat{\mathbf{s}}' + a^2\,d\phi'\,\hat{\mathbf{z}}.$$

The first term still contains the variable unit vector $\hat{\mathbf{s}}'$, so by the Commandment it must be written as $\cos\phi'\,\hat{\mathbf{x}} + \sin\phi'\,\hat{\mathbf{y}}$ before it is integrated, and then it integrates to zero over a full circle. The second term contains only the constant $\hat{\mathbf{z}}$. Since $\mathscr{r}$ is constant,

$$\begin{aligned}
\mathbf{B} &= \frac{\mu_0 I}{4\pi}\,\frac{1}{\left(z^2+a^2\right)^{3/2}}\left[az\int_0^{2\pi}\left(\cos\phi'\,\hat{\mathbf{x}} + \sin\phi'\,\hat{\mathbf{y}}\right)d\phi' + a^2\,\hat{\mathbf{z}}\int_0^{2\pi} d\phi'\right] \\
&= \frac{\mu_0 I}{4\pi}\,\frac{2\pi a^2\,\hat{\mathbf{z}}}{\left(z^2+a^2\right)^{3/2}},
\end{aligned}$$

or

$$\mathbf{B} = \frac{\mu_0 I a^2}{2\left(z^2+a^2\right)^{3/2}}\,\hat{\mathbf{z}}. \tag{10.22}$$

Converting to Cartesian components first gives the same thing with more algebra: $d\boldsymbol{\ell}' = a\,d\phi'(-\sin\phi'\,\hat{\mathbf{x}} + \cos\phi'\,\hat{\mathbf{y}})$ and $\boldsymbol{\mathscr{r}} = -a\cos\phi'\,\hat{\mathbf{x}} - a\sin\phi'\,\hat{\mathbf{y}} + z\hat{\mathbf{z}}$, whose cross product, multiplied out term by term, is $a\,d\phi'\,(z\cos\phi'\,\hat{\mathbf{x}} + z\sin\phi'\,\hat{\mathbf{y}} + a\hat{\mathbf{z}})$.

Reality check: At the center of the ring, $z = 0$, (10.22) gives

$$B = \frac{\mu_0 I}{2a},$$

which is the field of an infinite wire at a distance a, $\mu_0 I/2\pi a$, without the π. Bending the wire around the point puts every piece of it at the smallest distance, a, and at right angles to $\boldsymbol{\mathscr{r}}$, while most of a straight wire is farther away and seen at a slant.

Far from the ring, $z \gg a$,

$$B \approx \frac{\mu_0 I a^2}{2z^3} = \frac{\mu_0}{2\pi}\,\frac{I\pi a^2}{z^3},$$

which falls off as $1/z^3$, not $1/z^2$ (Figure 10.17). The ring looks from far away like the electric dipole of Section 8.3.5, whose field on its axis is $2p/4\pi\varepsilon_0 z^3$; it is a *magnetic dipole*, with a dipole moment equal to the current times the area of the loop, $I\pi a^2$ (Section 12.6.2). Because there are no magnetic charges, the field of a current loop of any shape falls off at least this fast far away. Fields that fall off more slowly, such as the $1/s$ field of a long straight wire, are found only at distances much smaller than the size of the circuit.

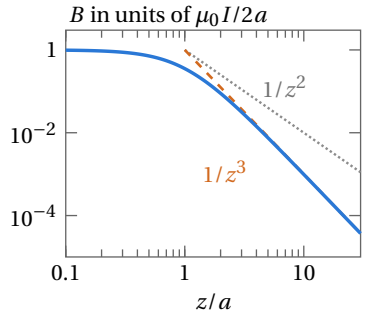
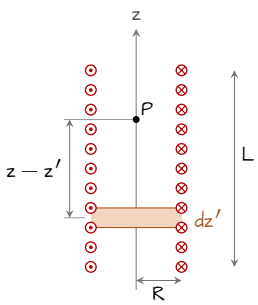


Figure 10.17 The field on the axis of a ring of radius a, (10.22), on logarithmic scales. Far from the ring it follows the line $1/z^3$ (dashed), not the line $1/z^2$ (dotted); both lines are drawn through 1 at $z = a$.

Example 10.8

A solenoid of radius R and length L is centered on the origin with its axis along z. It has n closely spaced turns per unit length, each carrying a current I counterclockwise as seen from above. Find the magnetic field on its axis.

Solution:



We sketch the solenoid from the side, with a thin slice of it at height z' and the field point at height z on the axis. Treat the solenoid as a stack of rings.

A slice of thickness dz' at height z' contains $n\,dz'$ turns, so it is a ring carrying the current $nI\,dz'$, and by (10.22), with a replaced by R and z by the distance $z - z'$ from

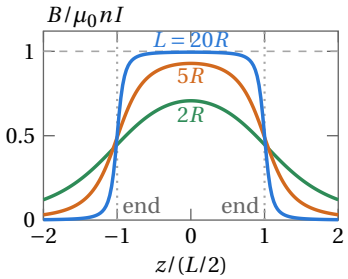


Figure 10.18 The field on the axis of a solenoid of length L and radius R, from (10.23), in units of $\mu_0 nI$. The ends are at $z = \pm L/2$. A long solenoid has a nearly uniform field inside, and half of that field at its ends.

the slice to the field point, its field is

$$dB = \frac{\mu_0 (nI\,dz')\,R^2}{2\left[R^2 + (z-z')^2\right]^{3/2}},$$

along $\hat{\mathbf{z}}$. Adding up the slices,

$$B = \frac{\mu_0 nI}{2}\int_{-L/2}^{L/2} \frac{R^2\,dz'}{\left[R^2 + (z-z')^2\right]^{3/2}} = \frac{\mu_0 nI}{2}\left[\frac{z'-z}{\sqrt{R^2+(z'-z)^2}}\right]_{-L/2}^{L/2},$$

so

$$\mathbf{B} = \frac{\mu_0 nI}{2}\left[\frac{L/2-z}{\sqrt{R^2+(L/2-z)^2}} + \frac{L/2+z}{\sqrt{R^2+(L/2+z)^2}}\right]\hat{\mathbf{z}}. \tag{10.23}$$

At the center, $z = 0$, and at either end, $z = \pm L/2$, the field is

$$B_{\text{center}} = \frac{\mu_0 nIL}{\sqrt{L^2+4R^2}}, \qquad B_{\text{end}} = \frac{\mu_0 nIL}{2\sqrt{L^2+R^2}}.$$

For a solenoid ten times as long as its radius, these are $0.981\,\mu_0 nI$ and $0.498\,\mu_0 nI$. Figure 10.18 shows the field along the axis for three lengths.

Reality check: For a very short solenoid, $L \ll R$, with $N = nL$ turns, B_{center} becomes $\mu_0 NI/2R$: N times the field at the center of a ring, $\mu_0 I/2a$ with $a = R$ (Example 10.7), as it should be, since all N turns then lie in the plane of the center.

As the solenoid gets longer, the field near its center approaches

$$B = \mu_0 nI,$$

which depends neither on the radius nor on the position along the axis, and the field at its ends approaches half that value. The field of a long solenoid is therefore nearly uniform inside, at least on the axis. Section 11.2.3 shows, with much less work, that inside an infinitely long solenoid the field is $\mu_0 nI$ everywhere, not just on the axis.

10.6.2 Tips for Mathematica and Python

The vector integrals of the script-r method can be set up and done by computer, which is a good way to check work done by hand.

- Write vectors as lists of Cartesian components and let the software do the cross product: `Cross[dl, r - rp]` in Mathematica, or `dl.cross(r - rp)` with sympy's `Matrix` vectors. Writing everything in Cartesian components keeps Colton's First Commandment automatically.
- Mathematica's `Integrate` works on each component of a list, so the integral of a vector can be done in one command. In sympy, apply `integrate` to each component.

- Tell the software what it may assume, with `Assuming[a > 0 && z > 0, ...]` in Mathematica or with `positive=True` when you define sympy symbols; otherwise results such as $\sqrt{z^2} = z$ are not simplified.
- For a vector integral that must be done numerically, `NIntegrate` accepts a list in Mathematica, and `quad_vec` in Python's `scipy.integrate` integrates all the components at once.

10.7 Numerical method: direct integration

The ring and solenoid examples of the last section had the field point on the axis. Move it off the axis, and even the field of a single circular loop has no expression in terms of elementary functions; for a real coil, wound as a helix with leads running to it, there is no closed form at all. The Biot–Savart integral can still be done numerically, and the chunk-and-sum method of Section 3.6 carries over with only one change: each chunk of current is a short vector, and its field involves a cross product (Figure 10.19).

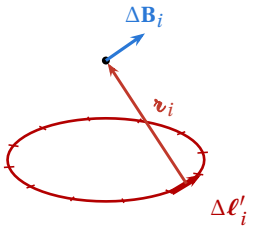


Figure 10.19 A loop cut into short pieces. Piece i is the short vector $\Delta\boldsymbol{\ell}'_i$, and its contribution $\Delta\mathbf{B}_i$ to the field at the field point (dot) is perpendicular to both $\Delta\boldsymbol{\ell}'_i$ and $\boldsymbol{\mathscr{r}}_i$.

1. Divide the current into N short pieces. For a wire, piece i is a short vector $\Delta\boldsymbol{\ell}'_i$ pointing along the current, centered at the point $\mathbf{r}'_i$. If the wire is described by a parameter u, as $\mathbf{r}'(u)$, a convenient choice is to divide the range of u into N equal steps Δu, take $\mathbf{r}'_i$ at the middle of each step, and set $\Delta\boldsymbol{\ell}'_i = (d\mathbf{r}'/du)\,\Delta u$ there.
2. For the field point $\mathbf{r}$, find $\boldsymbol{\mathscr{r}}_i = \mathbf{r} - \mathbf{r}'_i$ and $\mathscr{r}_i$ for each piece.
3. Add up the contributions:

$$\mathbf{B}(\mathbf{r}) \approx \frac{\mu_0 I}{4\pi} \sum_{i=1}^{N} \frac{\Delta\boldsymbol{\ell}'_i \times \boldsymbol{\mathscr{r}}_i}{\mathscr{r}_i^3}. \tag{10.24}$$

   For surface and volume currents, $I\,\Delta\boldsymbol{\ell}'_i$ becomes $\mathbf{K}_i\,\Delta a_i$ or $\mathbf{J}_i\,\Delta\tau_i$.
4. Increase N and repeat, until the answer stops changing at the accuracy you need.

As before, working in Cartesian components from the start takes care of Colton's First Commandment, and the method fails if the field point is on the current itself.

As an example, take a loop of radius R in the xy plane, centered on the origin and carrying a current I counterclockwise, and find the field at two points off its axis, in the xz plane. With $u = \phi'$, the pieces are centered at $\phi'_i = (i - \frac{1}{2})\,2\pi/N$, at $\mathbf{r}'_i = R\cos\phi'_i\,\hat{\mathbf{x}} + R\sin\phi'_i\,\hat{\mathbf{y}}$, with $\Delta\boldsymbol{\ell}'_i = R\,\Delta\phi'(-\sin\phi'_i\,\hat{\mathbf{x}} + \cos\phi'_i\,\hat{\mathbf{y}})$ and $\Delta\phi' = 2\pi/N$. Table 10.2 gives the components of $\mathbf{B}$ in units of $\mu_0 I/R$ (B_y is zero at these points, by symmetry), along with the exact values, which can be written in terms of complete elliptic integrals (Exercise P10.22). At the first point, which is $0.71R$ from the wire, 16 pieces are enough for five significant figures. The second point is

$s = 0.5R$, $z = 0.5R$

N	B_x	B_z
8	0.12774	0.34547
16	0.12867	0.34583
32	0.12867	0.34583
64	0.12867	0.34583
exact	0.12867	0.34583

$s = 0.8R$, $z = 0.2R$

N	B_x	B_z
8	0.25660	0.52416
16	0.38769	0.63538
32	0.40520	0.64945
64	0.40536	0.64957
exact	0.40536	0.64957

Table 10.2 The field of a ring of radius R at two points off its axis, in units of $\mu_0 I/R$, from the sum over N pieces and from the exact formula.

only $0.28R$ from the wire, and 16 pieces, each $0.39R$ long, are off by 4%; 32 pieces, each $0.20R$ long, are good to 0.04%. The lesson is the same as for electric fields: each piece must be short compared with its distance from the field point, so a field point close to the current needs many more pieces than one far away. (The errors shrink very quickly as N grows here because the loop closes on itself, so the integrand is periodic, the case in which the midpoint rule works best.)

A few tips:

- In Python, put the positions and vectors of all the pieces in `numpy` arrays, one row per piece; `numpy.cross` works on whole arrays of vectors at once, and the sum in (10.24) becomes one line with no explicit loop. In Mathematica, `Table` and `Total` do the same job.
- Test the program on a case you can do by hand, such as the field on the axis of the loop, (10.22), before trusting it anywhere else.
- To see the field, evaluate it on a grid of points and plot arrows: in Mathematica with `VectorPlot3D` or `VectorPlot`, and in Python with the `quiver` function of `matplotlib`. To draw field lines instead, use `StreamPlot` or `streamplot`. Leave out points on or very near the wire, where the field is enormous.

Exercises P10.21 and P10.22 apply the method to a short helical coil and to a Helmholtz pair, two coils arranged to give a nearly uniform field.

10.8 Summary

Magnetic fields are produced by moving charges, that is, by currents, and they push on moving charges. For steady currents, this chapter has followed the path that Chapter 3 followed for charges at rest: a force law, a law for the field of a small piece of the source, and ways of doing the resulting integrals. The structure is similar, but not identical: the source is a vector, $I\,d\boldsymbol{\ell}$, and cross products appear in both the force and the field.

Forces and currents

$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B})$	the Lorentz force law; the magnetic part is perpendicular to $\mathbf{v}$ and does no work
$\mathbf{F} = I\mathbf{L} \times \mathbf{B}, \quad \dfrac{F}{L} = \dfrac{\mu_0 I_1 I_2}{2\pi d}$	the force on a straight wire in a uniform field; the force per unit length between long parallel wires, attractive for currents in the same direction
$R = \dfrac{mv}{qB}, \quad \omega_c = \dfrac{qB}{m}$	radius and angular frequency of circular motion in a uniform field
$I = \lambda v, \quad \mathbf{K} = \sigma\mathbf{v}, \quad \mathbf{J} = \rho\mathbf{v}$	current, surface current density, volume current density
$I = \displaystyle\int K_\perp \, d\ell, \quad I = \int \mathbf{J} \cdot d\mathbf{a}$	the current crossing a line or a surface: integrate across the flow
$\nabla \cdot \mathbf{J} = -\dfrac{\partial \rho}{\partial t}, \quad \nabla \cdot \mathbf{J} = 0$	continuity; steady currents
$\mathbf{F} = \displaystyle\int I \, d\boldsymbol{\ell} \times \mathbf{B}$	the force on a current, with $\mathbf{K}\,da$ or $\mathbf{J}\,d\tau$ in place of $I\,d\boldsymbol{\ell}$ for surface and volume currents; zero on a closed loop in a uniform field

Fields

$\mathbf{B}(\mathbf{r}) = \dfrac{\mu_0}{4\pi} \displaystyle\int \frac{I \, d\boldsymbol{\ell}' \times \boldsymbol{\mathscr{r}}}{\mathscr{r}^3}$	the Biot–Savart law, for steady currents; $\mathbf{K}\,da'$ or $\mathbf{J}\,d\tau'$ in place of $I\,d\boldsymbol{\ell}'$ for surface and volume currents
$\mu_0 = 4\pi \times 10^{-7}\,\mathrm{N/A^2}$	the permeability of free space (very nearly)

The script-r method. Write down $\mathbf{r}$, $\mathbf{r}'$, $\boldsymbol{\mathscr{r}}$, $\mathscr{r}$, and $d\boldsymbol{\ell}'$ (or da', $d\tau'$) separately, form the cross product, and integrate each Cartesian component. Find the direction with the right-hand rule first when you can. *Colton's First Commandment* applies: thou shalt not move a variable unit vector outside the integral. Either convert to Cartesian components first, or take the cross product first and then convert. When the integrals cannot be done by hand, chop the current into short pieces and add their fields numerically, with pieces short compared with their distance from the field point. To find the motion of a charge in a magnetic field numerically, use a method, such as the Boris method, that keeps the speed constant.

Results worth remembering. A long straight wire: $\mu_0 I/2\pi s$, circling the wire. A circular loop of radius R, at its center: $\mu_0 I/2R$; far away on its axis, the field falls off as $1/z^3$.

Field lines of **B** never begin or end. Near a wire they circle it; through a loop or a

solenoid they run along the axis and return outside. They are not lines of force: the force on a moving charge is perpendicular to them.

Exercises

Exercises for 10.1 Forces on moving charges: the Lorentz force law

Warm-up questions Read Section 10.1, then answer before class

W10.1 At the point $(0,2,0)$, write $\hat{\mathbf{s}}$ and $\hat{\boldsymbol{\phi}}$ in Cartesian components, and check that $\hat{\mathbf{s}} \times \hat{\boldsymbol{\phi}} = \hat{\mathbf{z}}$, as the rules of Section 10.1.1 say. Now let $\hat{\boldsymbol{\phi}}'$ be the unit vector $\hat{\boldsymbol{\phi}}$ at the point $(0,-2,0)$. Find $\hat{\boldsymbol{\phi}}' \times \hat{\mathbf{s}}$, with $\hat{\mathbf{s}}$ still taken at $(0,2,0)$, and compare it with the rule $\hat{\boldsymbol{\phi}} \times \hat{\mathbf{s}} = -\hat{\mathbf{z}}$. Which is right, and what went wrong?

W10.2 A proton and an electron are each released from rest at the same point in the crossed fields of Figure 10.6c, with $\mathbf{E}$ up the page and $\mathbf{B}$ into it. Which way does each one start to move? Which way does each one drift on average, and why does the direction of the drift not depend on the sign of the charge? Which of the two makes the larger arches?

W10.3 The simple update of Section 10.1.5 uses the force at the old velocity $\mathbf{v}_n$, and every step makes the speed grow. Suppose instead that the magnetic force were evaluated at the new velocity, $\mathbf{v}_{n+1} = \mathbf{v}_n + (q/m)\,\mathbf{v}_{n+1} \times \mathbf{B}\,\Delta t$, with $\mathbf{E} = 0$. Sketch $\mathbf{v}_n$, $\mathbf{v}_{n+1}$, and their difference for each rule. With the new rule, does the speed grow, shrink, or stay the same? Why?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P10.1 In a magnetic-sector mass spectrometer, singly charged positive ions start essentially at rest, are accelerated through a potential difference ΔV, and then enter a region of uniform magnetic field $\mathbf{B}$ through a slit, moving perpendicular to $\mathbf{B}$ (Figure 10.20). Each ion moves on a half circle and strikes a detector plate in the plane of the slit.

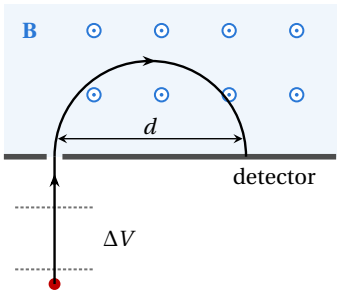


Figure 10.20 The mass spectrometer of Exercise P10.1, with $\mathbf{B}$ out of the page.

(a) Show that an ion of mass m and charge q lands a distance

$$d = \frac{2}{B}\sqrt{\frac{2m\Delta V}{q}}$$

from the slit.

(b) For $\Delta V = 20.0\,\text{kV}$ and $B = 0.500\,\text{T}$, find d for the ions $^{235}\text{U}^+$ and $^{238}\text{U}^+$, whose masses are 235.044 u and 238.051 u ($1\,\text{u} = 1.6605 \times 10^{-27}\,\text{kg}$), and the distance between their landing points.

(c) If instead all of the ions entered the field with the same speed, by what factor would d change for an ion whose charge-to-mass ratio is twice as large?

(d) With the arrangement of part (a), by what factor does d change for a doubly charged ion of the same mass?

P10.2 In the auroral zone, at the heights where aurorae glow, Earth's magnetic field has a magnitude of roughly $50\,\mu$T. Take $B = 50.0\,\mu$T, and treat the field as uniform. A proton with a kinetic energy of 1.00 keV moves with its velocity at 30.0° to **B**.

(a) Find its cyclotron frequency and period.

(b) Find the radius of its helical path.

(c) Find the pitch of the helix, the distance it advances along **B** in one revolution.

(d) Repeat parts (a)–(c) for an electron with the same kinetic energy, moving at the same angle to **B**. (Nonrelativistic mechanics is adequate.)

(e) Show that, for equal kinetic energies, the ratio of the radii is $r_p/r_e = \sqrt{m_p/m_e}$, and the ratio of the periods is $T_p/T_e = m_p/m_e$.

P10.3 A particle of charge $q > 0$ and mass m moves in the uniform fields $\mathbf{E} = E\hat{\mathbf{y}}$ (up the page) and $\mathbf{B} = -B\hat{\mathbf{z}}$ (into the page), the arrangement of Figure 10.6c. Let $\omega = qB/m$.

(a) Show that a particle moving with the velocity $-(E/B)\hat{\mathbf{x}}$ is not deflected, and explain how this arrangement can select particles of one speed.

(b) The particle is released from rest at the origin. Show that

$$x(t) = -\frac{E}{\omega B}(\omega t - \sin\omega t), \qquad y(t) = \frac{E}{\omega B}(1 - \cos\omega t)$$

satisfies the equations of motion and the initial conditions. Sketch the path, and find the greatest height the particle reaches and its average velocity. Evaluate both for a proton with $E = 1.00 \times 10^5$ V/m and $B = 0.0500$ T.

(c) Now the particle starts at the origin with the velocity $-v_0\hat{\mathbf{x}}$, where v_0 may be positive or negative. For which values of v_0 does the path have loops? Cusps? For which is it a straight line? (Hint: write the velocity as $-(E/B)\hat{\mathbf{x}}$ plus a remainder **w**, and show that **w** obeys the equation of motion for the magnetic field alone.)

(d) Show that in parts (b) and (c) the kinetic energy at any time equals its initial value plus qEy, so that the magnetic force does no work.

P10.4 Write a program that integrates the equation of motion $m\,d\mathbf{v}/dt = q(\mathbf{E} + \mathbf{v}\times\mathbf{B})$ in time steps, as in Section 10.1.5, and plots the trajectory. Use units in which $q = m = 1$, take $|\mathbf{E}| = |\mathbf{B}| = 1$ where they are present, start at the origin, and run to $t = 20$.

(a) Plot the trajectories for (i) $\mathbf{E} = \hat{\mathbf{x}}$, $\mathbf{B} = 0$, starting from rest; (ii) $\mathbf{E} = \hat{\mathbf{x}}$, $\mathbf{B} = 0$, $\mathbf{v}_0 = \hat{\mathbf{y}}$; (iii) $\mathbf{E} = 0$, $\mathbf{B} = \hat{\mathbf{z}}$, $\mathbf{v}_0 = \hat{\mathbf{x}}$; (iv) $\mathbf{E} = 0$, $\mathbf{B} = \hat{\mathbf{z}}$, $\mathbf{v}_0 = \hat{\mathbf{x}} + \hat{\mathbf{z}}$; and (v) $\mathbf{E} = \hat{\mathbf{x}}$, $\mathbf{B} = \hat{\mathbf{z}}$, starting from rest. Describe each one, and check (iii) against (10.5).

(b) With the simple update of Section 10.1.5, each step of case (iii) multiplies the speed by $\sqrt{1 + (\omega_c \Delta t)^2}$. Find the largest Δt that keeps the kinetic energy within one part in 10^3 of its initial value after ten cyclotron periods, and confirm it with your program. How many steps per period is that?

(c) Replace the velocity update by the Boris method of Section 10.1.5. Show that, with $\mathbf{E} = 0$, the step from $\mathbf{v}^-$ to $\mathbf{v}^+$ does not change the magnitude of the velocity, and confirm with your program that the speed in case (iii) stays constant, even with $\Delta t = 0.1$.

(d) The exact solution of case (v) is the cycloid $x = 1 - \cos t$, $y = -(t - \sin t)$. Compare your position at $t = 20$ with it, for both methods and several time steps. How does the error of each method depend on Δt?

P10.5 Protons with energies of millions of electron volts are trapped in Earth's magnetic field, in the inner Van Allen belt (Section 10.1.4). In this exercise you follow one (Figure 10.21). Model Earth's field as the field of a magnetic dipole at its center (Section 12.6.3),

$$\mathbf{B} = \frac{\mu_0}{4\pi}\,\frac{3(\mathbf{m}\cdot\hat{\mathbf{r}})\,\hat{\mathbf{r}} - \mathbf{m}}{r^3},$$

with $\mathbf{m} = -(8.0\times 10^{22}\,\mathrm{A\cdot m^2})\,\hat{\mathbf{z}}$, where $+\hat{\mathbf{z}}$ points toward geographic north, and take Earth's radius to be $R_E = 6.371\times 10^6\,\mathrm{m}$. A proton with a kinetic energy of 1.00 MeV ($v/c = 0.046$, so nonrelativistic mechanics is adequate) starts at $\mathbf{r}_0 = 2R_E\,\hat{\mathbf{x}}$, above the equator, with the velocity $\mathbf{v} = v\sin\alpha\,\hat{\mathbf{x}} + v\cos\alpha\,\hat{\mathbf{z}}$, where $\alpha = 30.0°$ is the angle between $\mathbf{v}$ and $\mathbf{B}$.

(a) Find $\mathbf{B}(\mathbf{r}_0)$, and the radius and period of the proton's circular motion there.

(b) Integrate the motion with the Boris method of Section 10.1.5, using 50 steps per period of the circular motion, for two complete bounces between the northern and southern hemispheres, and plot the trajectory. Find the greatest magnetic latitude the proton reaches, and the bounce period, the time between successive northward crossings of the equatorial plane.

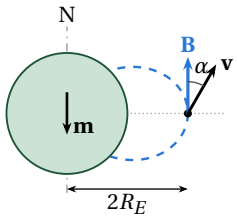


Figure 10.21 Exercise P10.5: Earth's dipole, the field line through $2R_E$ on the equator (dashed), and the proton's starting point and velocity.

(c) A theory of slow motion along field lines says that $v_\perp^2/B$ stays constant as the particle moves, where $v_\perp$ is the component of its velocity perpendicular to **B**. Along the field line $r = 2R_E\cos^2\lambda$, where λ is the magnetic latitude, the field strength is $B = B_{\text{eq}}\sqrt{1+3\sin^2\lambda}/\cos^6\lambda$ (Exercise P10.7). Use these facts to show that the proton turns back, or mirrors, at the latitude λ_m where $\cos^6\lambda_m/\sqrt{1+3\sin^2\lambda_m} = \sin^2\alpha$, and find λ_m. Then evaluate the bounce period, $T_b = 4\int_0^{\lambda_m} d\ell/v_\parallel$, with $d\ell = 2R_E\cos\lambda\sqrt{1+3\sin^2\lambda}\,d\lambda$ and $v_\parallel = v\sqrt{1-B/B_m}$, where B_m is the field at λ_m, and compare it with part (b).

(d) The proton also drifts slowly around Earth. Which way does it drift, east or west? Run the simulation for at least 20 bounces, find the average change in the proton's longitude per bounce, and estimate how long it takes to go once around Earth.

(e) Repeat part (b) with 25 and with 100 steps per period. How much do the results change?

(f) The field line through $r = 2R_E$ at the equator meets Earth's surface at a latitude of 45°. If α is too small, the proton reaches the dense lower atmosphere before it mirrors, and is lost. Find the smallest α for which it mirrors above Earth's surface.

Exercises for 10.2 Why the force and the field need cross products

Warm-up questions Read Section 10.2, then answer before class

W10.4 This section ends by saying that the field "splits a complicated problem into two simpler steps." For the force that a current loop exerts on an electron moving near it, what are the two steps, and which cross product does each use? Why does the second step not depend on whether the field comes from the loop, a magnet, or Earth?

W10.5 Suppose every physics book had always used a left-hand rule for cross products, so that what we write as $\mathbf{A}\times\mathbf{C}$ were replaced everywhere by its negative: in the field of a short piece of current, $d\boldsymbol{\ell}'\times\boldsymbol{\hat{r}}$, and in the force $q\mathbf{v}\times\mathbf{B}$ alike. Would any measured force, on a moving charge or between two currents, change? Would the direction of **B** change? What does this say about whether an experiment can see which way **B** points?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P10.6 Section 10.2 argued that any force that is linear in the velocity and always perpendicular to it has the form $q\mathbf{v}\times\mathbf{B}$. This exercise fills in the details.

Suppose that the force on a charge q moving with velocity $\mathbf{v}$ is $F_i = q\sum_j M_{ij}v_j$, where the matrix M may depend on position but not on $\mathbf{v}$, and that $\mathbf{F}$ is perpendicular to $\mathbf{v}$ for every $\mathbf{v}$.

(a) By choosing $\mathbf{v}$ along $\hat{\mathbf{x}}$, then along $\hat{\mathbf{x}}+\hat{\mathbf{y}}$, and so on, show that M is antisymmetric: $M_{ij} = -M_{ji}$.

(b) Show that if $\mathbf{B} = (M_{23}, M_{31}, M_{12})$, then $\sum_j M_{ij}v_j$ is the i component of $\mathbf{v}\times\mathbf{B}$, for every $\mathbf{v}$.

(c) A charge moving along $+\hat{\mathbf{x}}$ with speed v feels the force $\mathbf{F} = qv\,(2.0\,\hat{\mathbf{y}} + 1.0\,\hat{\mathbf{z}})$ T. Moving along $+\hat{\mathbf{y}}$ with the same speed, it feels $\mathbf{F} = qv\,(-2.0\,\hat{\mathbf{x}} - 3.0\,\hat{\mathbf{z}})$ T. Find $\mathbf{B}$, check that the two measurements are consistent with each other, and predict the force when the charge moves along $+\hat{\mathbf{z}}$.

(d) Explain why a measurement of the force for a single velocity can never determine $\mathbf{B}$ completely.

Exercises for 10.3 The magnetic field and field lines

Warm-up questions Read Section 10.3, then answer before class

W10.6 Far from the circular loop of Figure 10.7b, the field lines have the same shape as those of an electric dipole. Close in, they differ. Does $\mathbf{B}$ at the center of the loop point the same way as $\mathbf{B}$ far out on the axis, or the opposite way? Answer the same question for $\mathbf{E}$ at the midpoint between the two charges of an electric dipole, compared with $\mathbf{E}$ far out on its axis. Explain the difference with what each kind of field line can and cannot do.

W10.7 Every field line that runs up the inside of a solenoid (Figure 10.7c) has to come back down outside it. For a long solenoid of radius R and length L, suppose that on their way back the lines cross the plane through the middle of the solenoid spread out over a disk of radius about L. Use the number of lines to estimate the field just outside the middle of the solenoid, as a fraction of the field inside. What happens to that fraction if the solenoid is made twice as long?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P10.7 Far from Earth's surface, Earth's magnetic field is approximately the field of a magnetic dipole (Section 12.6.3). In spherical coordinates with the z axis along the dipole (Figure 10.22), the field is, apart from an overall sign,

$$\mathbf{B} = \frac{\mu_0 m}{4\pi r^3}\left(2\cos\theta\,\hat{\mathbf{r}} + \sin\theta\,\hat{\boldsymbol{\theta}}\right),$$

with $m = 8.0\times10^{22}\,\mathrm{A\cdot m^2}$. Earth's radius is $R_E = 6.371\times10^6\,\mathrm{m}$.

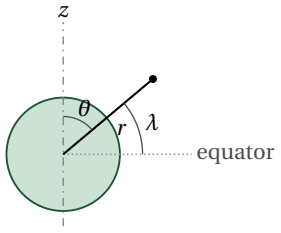


Figure 10.22 The angles of Exercise P10.7: the polar angle θ, measured from the dipole axis, and the magnetic latitude $\lambda = 90^\circ - \theta$, measured from the equator.

(a) A field line is everywhere parallel to **B**, so along it $dr/(r\,d\theta) = B_r/B_\theta$. Show that the field lines are the curves $r = r_{\text{eq}}\sin^2\theta$, where r_{eq} is the distance at which the line crosses the equator.

(b) Show that the field line that crosses the equator at $r_{\text{eq}} = LR_E$ reaches Earth's surface at the magnetic latitude $\lambda = \arccos(1/\sqrt{L})$, where $\lambda = 90° - \theta$. Evaluate λ for $L = 4$, in the outer Van Allen belt, and for $L = 6.6$, the distance of geosynchronous orbit, whose field line reaches the ground near the auroral zone.

(c) Show that along a field line $B/B_{\text{eq}} = \sqrt{1+3\sin^2\lambda}/\cos^6\lambda$, where B_{eq} is the field where the line crosses the equator. Evaluate the ratio of the field at Earth's surface to B_{eq} for the line with $L = 4$.

(d) Find the magnitude of **B** at Earth's surface on the magnetic equator, and on the equator at geosynchronous orbit.

Exercises for 10.4 Currents and current densities

Warm-up questions Read Section 10.4, then answer before class

W10.8 A wire of radius R carries a total current I, all of it in a thin layer of thickness $t \ll R$ at its surface, with a uniform current density in the layer. What is J in the layer? As t is made smaller, with I fixed, what happens to J, and what describes the current instead? Find it, and check that it gives the total current I.

W10.9 Which of these are steady currents, in the sense of Section 10.4.1? Give a reason for each. (a) The current in a wire that charges a capacitor at a constant rate. (b) A current that has circulated for an hour in a superconducting ring. (c) The current of an electron moving on a circle around a nucleus, as in Bohr's early model of hydrogen. (d) The same current averaged over one orbit.

W10.10 A rectangular loop of width w carries a current I clockwise and hangs in the plane of the page, with its bottom side in a region of uniform magnetic field that points into the page and its top side outside the field. Section 10.4.2 shows that a closed loop in a uniform field feels no net force. Does this loop feel one? Which parts of the loop contribute, and which way does the net force point?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P10.8 Find the total current in each case.

(a) An infinitely long thick cylindrical shell, $a < s < b$, carries the current density $\mathbf{J} = (k/s^3)\,\hat{\mathbf{z}}$. What are the units of k?

(b) A flat strip, $-a < x < a$, infinitely long in the y direction, carries the surface current density $\mathbf{K} = kx^2\,\hat{\mathbf{y}}$. What are the units of k?

(c) A ball of radius R with uniform charge density ρ spins about the z axis with angular velocity ω. Find its current density, and the current that crosses the half-plane $\phi = 0$.

(d) A spherical shell of radius R with uniform surface charge density σ spins in the same way. Find $\mathbf{K}$ as a function of θ, and the current that crosses the half-plane $\phi = 0$.

(e) Explain why the answers to (c) and (d) are both $Q\omega/2\pi$, where Q is the total charge.

P10.9 For each current density, decide whether it describes a steady current. If it does not, find $\partial\rho/\partial t$. Here k and J_0 are constants.

(a) $\mathbf{J} = J_0\,(x/a)^2\,\hat{\mathbf{x}}$

(b) $\mathbf{J} = ks\hat{\boldsymbol{\phi}}$

(c) $\mathbf{J} = k\,(y\hat{\mathbf{x}} + x\hat{\mathbf{y}})$

(d) $\mathbf{J} = kr\,\hat{\mathbf{r}}$

(e) A uniformly charged ball of radius $R(t)$ expands, so that the charge at position $\mathbf{r}$ moves with the velocity $(\dot{R}/R)\,\mathbf{r}$. Show that the continuity equation requires ρR^3 to stay constant, and explain why that makes sense.

P10.10 A flat conducting strip of width w and thickness t lies in the xy plane, with its length along x and its width along y (Figure 10.23). It carries a current I in the $+x$ direction, in a uniform magnetic field $\mathbf{B} = B\hat{\mathbf{z}}$. The current is carried by particles of charge q, which may be positive or negative, with n of them per unit volume, drifting along the strip with speed v_d.

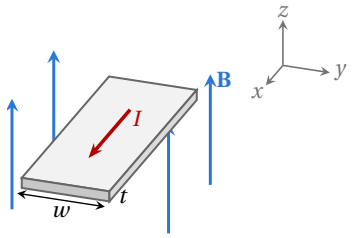


Figure 10.23 The strip of Exercise P10.10, with its current and the applied field.

(a) Show that $I = n|q|v_d\,wt$.

(b) The magnetic force pushes the moving charges toward one edge of the strip, until the electric field $\mathbf{E}_H$ of the charge that piles up at the edges balances it. Show that in the steady state $E_H = v_d B$, and that the voltage across the width of the strip, the Hall voltage, is $V_H = IB/n|q|t$. (Treat $\mathbf{E}_H$ as uniform across the strip.)

(c) Which edge, $+y$ or $-y$, becomes positively charged if the moving charges are positive? If they are electrons? Explain how the Hall voltage reveals the sign of the charges that carry the current.

(d) Copper has 8.5×10^{28} conduction electrons per cubic meter. For a copper strip with $w = 1.0\,\text{cm}$ and $t = 0.10\,\text{mm}$, carrying $I = 10\,\text{A}$ in a field of 1.0 T, find the drift speed and the Hall voltage.

(e) Repeat part (d) for a semiconductor strip of the same size, with $n = 1.0 \times 10^{22}\,\text{m}^{-3}$, carrying 1.0 mA in a field of 0.10 T. Why are Hall probes for measuring magnetic fields made of semiconductors rather than metals?

Exercises for 10.5 The Biot–Savart law

Warm-up questions Read Section 10.5, then answer before class

W10.11 A circular loop of radius a carries a current I. The current enters and leaves through a small gap in the loop, along two long straight leads that run radially outward from the gap, side by side. Do the leads change the field at the center of the loop? Do they produce much field at a point beside them, far from the loop?

W10.12 Biot and Savart measured the field of a long straight wire, and from it Laplace and Biot inferred a law for a single short piece of current. Could an experiment measure the field of one short piece of a steady current by itself? If not, what do experiments on steady currents actually test, and could some other law for a single piece agree with all of them? (The footnote in Section 10.2 is a hint.)

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

Exercises for 10.6 Solving for B when given J: script-r method

Warm-up questions Read Section 10.6, then answer before class

W10.13 The ring of Example 10.7 now carries its current clockwise as seen from above. Three classmates set up the integral. The first keeps $d\boldsymbol{\ell}' = a\,d\phi'\,\hat{\boldsymbol{\phi}}'$ and runs ϕ' from 2π down to 0; the second writes $d\boldsymbol{\ell}' = -a\,d\phi'\,\hat{\boldsymbol{\phi}}'$ and runs ϕ' from 0 to 2π; the third writes $-a\,d\phi'\,\hat{\boldsymbol{\phi}}'$ and runs ϕ' from 2π down to 0. Which of them get the right field, and what field does the other one get?

W10.14 Far from the ring of Example 10.7, its field falls off as $1/z^3$. A classmate concludes that the Biot–Savart law is not really an inverse-square law. What is wrong with the argument? Use the directions of the contributions in the example's sketch to explain why they add up to a field that falls off faster than $1/z^2$.

W10.15 Example 10.8 shows that the field at the end of a long solenoid is half the field at its center. Explain why without the formula. (Imagine a second, identical solenoid placed end to end with the first, so that together they make one solenoid twice as long.)

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P10.11 Two long straight wires (Figure 10.24), parallel to the z axis, pass through the points $(\pm d, 0, 0)$.

(a) Both carry a current I in the $+z$ direction. Find **B** at every point of the x axis and of the y axis. Where is the field zero? Where on the y axis is it strongest? What does the field look like far from both wires?

(b) Repeat part (a) with the current $+I$ in the wire at $x = +d$ and $-I$ in the wire at $x = -d$. How does the field far away depend on the distance?

(c) Sketch the field lines in both cases.

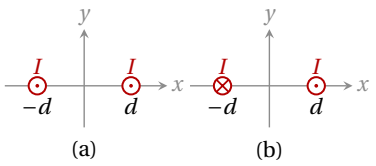


Figure 10.24 The two wires of Exercise P10.11, seen end on, with the currents of parts (a) and (b).

P10.12 A wire bent into an equilateral triangle with sides of length L carries a current I (Figure 10.25).

(a) Use (10.21) to find **B** at the center of the triangle.

(b) Compare your answer with the field at the center of a circular loop carrying the same current, (i) with the same perimeter as the triangle, and (ii) with the same area.

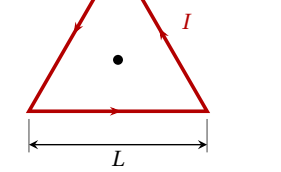


Figure 10.25 The triangular loop of Exercise P10.12 and its center.

P10.13 A long wire carries a current I in from $x = +\infty$ along the x axis to the origin, turns through 90° there, and leaves along the $+y$ axis (Figure 10.26). Find **B** at the points

(a) $(0, -s, 0)$, with $s > 0$;

(b) $(a, a, 0)$;

(c) $(-a, -a, 0)$;

(d) $(a, -a, 0)$.

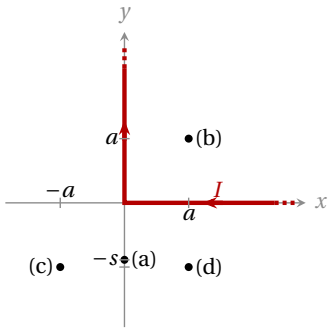


Figure 10.26 The bent wire of Exercise P10.13, with the points of parts (a)–(d).

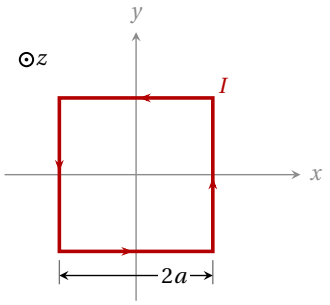


Figure 10.27 The square loop of Exercise P10.14.

P10.14 A square loop with sides of length $2a$ lies in the xy plane, centered on the origin, with its sides parallel to the axes (Figure 10.27). It carries a current I counterclockwise as seen from $+z$.

(a) Use (10.21) for each side to show that at points on the x axis, inside or outside the loop (but not at $x = \pm a$),

$$B_z(x) = \frac{\mu_0 I}{2\pi a}\left[\frac{\sqrt{a^2+(a-x)^2}}{a-x} + \frac{\sqrt{a^2+(a+x)^2}}{a+x}\right].$$

(b) Evaluate B_z at $x = 0$, $a/2$, $0.9a$, and $0.99a$, in units of $\mu_0 I/\pi a$. Compare the last with the field at the same distance from an infinitely long straight wire, and explain the comparison.

(c) Evaluate B_z at $x = 2a$ and $x = 10a$. Far from the loop, in its plane, the field should approach $-\mu_0 m/4\pi x^3$, where $m = (2a)^2 I$ is the loop's magnetic dipole moment (Section 12.6.2). Does it?

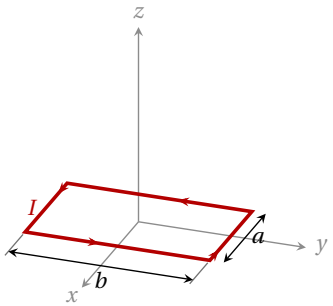


Figure 10.28 The rectangular loop of Exercise P10.15.

P10.15 A rectangular loop with sides a (parallel to x) and b (parallel to y) lies in the xy plane, centered on the origin, and carries a current I counterclockwise as seen from $+z$ (Figure 10.28).

(a) Use the Biot–Savart law and symmetry to find $\mathbf{B}$ at points on the z axis. (You may use (10.21) for each side, but take care with the direction of each side's contribution.)

(b) Show that for $z \gg a, b$ the field becomes $\mathbf{B} = (\mu_0 m/2\pi z^3)\hat{\mathbf{z}}$, with $m = Iab$, the same as for a circular loop with the same area and current (Example 10.7).

(c) Find $\mathbf{B}$ at the center. For a square loop of side w, compare it with the field at the center of a circular loop with the same area.

(d) For the square, beyond what value of z/w is the far-field formula of part (b) good to 1%?

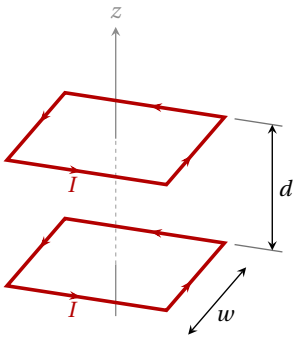


Figure 10.29 The two square coils of Exercise P10.16.

P10.16 Two identical square coils with sides of length w, each with N turns, are parallel to the xy plane and centered on the z axis, a distance d apart (Figure 10.29). They carry the same current I in the same sense. Measure z from the point midway between them.

(a) Use the result of Exercise P10.15 to write B_z on the z axis.

(b) Show that $dB_z/dz = 0$ at $z = 0$, whatever the value of d.

(c) Find, numerically, the value of d/w for which d^2B_z/dz^2 is also zero at $z = 0$. With this spacing the field near the center is as uniform as possible.

(d) For that spacing, find B_z at the center, and the distance from the center along the axis within which B_z stays within 1% of its central value.

(e) Evaluate d and the central field for $w = 20.0\,\text{cm}$, $N = 50$, and $I = 1.00\,\text{A}$.

P10.17 A thin nonconducting annulus, a flat ring with inner radius a and outer radius b, carries a uniform surface charge density σ and spins with angular velocity ω about its axis, the z axis (Figure 10.30).

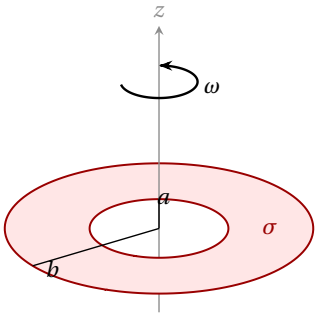


Figure 10.30 The spinning annulus of Exercise P10.17.

(a) Find the surface current density $\mathbf{K}(s)$ and the total current.

(b) Find $\mathbf{B}$ at the center.

(c) Treat the annulus as a set of rings, and use (10.22) to show that on the axis

$$B_z = \frac{\mu_0 \sigma \omega}{2}\left[\frac{b^2+2z^2}{\sqrt{b^2+z^2}} - \frac{a^2+2z^2}{\sqrt{a^2+z^2}}\right].$$

(d) Show that for $z \gg b$ this becomes $\mu_0 m/2\pi z^3$, with $m = \pi\sigma\omega(b^4 - a^4)/4$, and show that this m is $\int \pi s^2\, dI$, the sum of the dipole moments $I\pi s^2$ of the rings.

(e) For $\sigma = 1.0\,\mu\text{C/m}^2$, a rotation rate of 100 revolutions per second, $a = 5.0\,\text{cm}$, and $b = 10.0\,\text{cm}$, evaluate the field at the center and compare it with Earth's field.

P10.18 A long straight wire along the y axis carries a current I_1 in the $+y$ direction (Figure 10.31). A circular loop of radius R, carrying a current I_2, lies in the xy plane, centered on the point $(d, 0, 0)$, with $d > R$; its current runs counterclockwise as seen from $+z$.

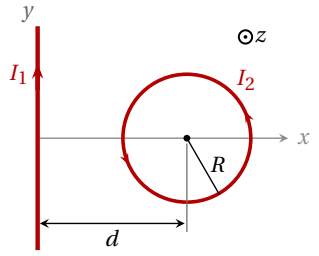


Figure 10.31 The wire and the loop of Exercise P10.18.

(a) Use (10.14) to show that the net force on the loop is

$$\mathbf{F} = \mu_0 I_1 I_2 \left(\frac{d}{\sqrt{d^2 - R^2}} - 1\right)\hat{\mathbf{x}}.$$

You may use $\int_0^{2\pi} \cos\phi\, d\phi/(d + R\cos\phi) = (2\pi/R)\left(1 - d/\sqrt{d^2 - R^2}\right)$.

(b) Explain the direction of the force, from the directions of the currents on the near and far sides of the loop.

(c) Show that for $d \gg R$, $F \approx \mu_0 I_1 I_2 R^2/2d^2$.

(d) Evaluate $F/\mu_0 I_1 I_2$ for $d = 10R$, $2R$, and $1.1R$, and compare each with the approximation of part (c).

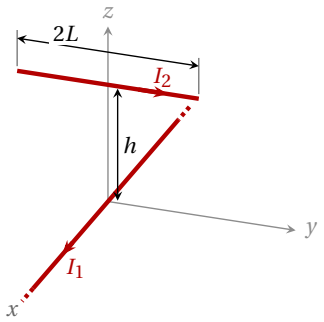


Figure 10.32 The long wire and the short piece of Exercise P10.19.

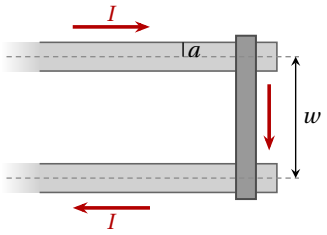


Figure 10.33 The railgun of Exercise P10.20, seen from above. The rails continue to the left, to the distant power supply.

P10.19 An infinitely long straight wire along the x axis carries a current I_1 in the $+x$ direction (Figure 10.32). A straight piece of wire of length $2L$, parallel to the y axis and centered on the point $(0,0,h)$, carries a current I_2 in the $+y$ direction. (The rest of its circuit is far away, and the forces on it can be ignored.)

(a) Find the magnetic force per unit length on the piece, as a function of the position y along it.

(b) Show that the net force on the piece is zero.

(c) Find the torque on the piece about the vertical line through its center, and describe which way it tends to turn the piece.

(d) What happens to the force and the torque as $L \to \infty$?

P10.20 In a railgun, two long parallel rails of radius a, with their axes a distance w apart, carry a current I to and from a conducting armature that bridges the gap between them and can slide along them (Figure 10.33). Treat each rail as a straight line current that runs from the armature off to infinity, and ignore the fields of the armature itself and of the distant power supply.

(a) Show that at a point on the armature a distance u from the axis of one rail, the field is $B = (\mu_0 I/4\pi)\,[1/u + 1/(w-u)]$, perpendicular to the plane of the rails.

(b) Show that the force on the armature is $F = (\mu_0 I^2/2\pi)\ln[(w-a)/a]$, and find its direction.

(c) For $I = 1.0\,\text{MA}$, $w = 10\,\text{cm}$, and $a = 1.0\,\text{cm}$, evaluate F, the acceleration of a 1.0 kg armature, and its speed after it has moved 5.0 m from rest. (Ignore friction.)

Exercises for 10.7 Numerical method: direct integration

Warm-up questions Read Section 10.7, then answer before class

W10.16 This section suggests testing a program for (10.24) on the axis of a circular loop, where (10.22) gives the answer, before trusting it anywhere else. Name two mistakes in a program that such a test would catch. Why test against a known answer at all, rather than just checking that the output looks reasonable?

W10.17 A real coil often has many layers of windings: N turns of wire, each carrying a current I, packed uniformly into a rectangular cross section of width a and height b, wound around the coil's axis. To find its field with (10.24), you could add up N separate loops, or treat the winding as a volume current. What is **J**, in size and direction? Where in space would the volume current be a poor model?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P10.21 A short solenoid is wound as a helix of radius $R = 1.00\,\text{cm}$, with 20 turns over a length of 2.00 cm, and carries a current $I = 1.00\,\text{A}$ (Figure 10.34). With the axis of the coil along z and its center at the origin, the wire follows $x' = R\cos u$, $y' = R\sin u$, $z' = (p/2\pi)u$, for $-20\pi \le u \le 20\pi$, where $p = 1.00\,\text{mm}$ is the distance between turns.

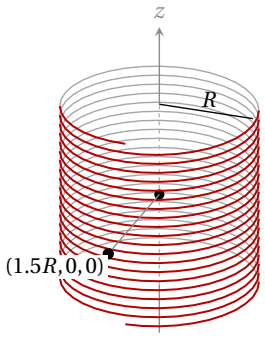


Figure 10.34 The helical coil of Exercise P10.21. The front of the wire is red, the back gray.

(a) Write a program that finds **B** at any point by adding the contributions (10.24) of N pieces of the wire.

(b) Find **B** at the center of the coil and at the point $(1.5R, 0, 0)$, for $N =$ 200, 1000, and 10^4, first with each piece evaluated at its starting end and then at its midpoint. Which rule converges faster? Why is $N = 200$ enough at the center but not near the windings?

(c) Compare B_z at the center, and at the end of the coil on its axis, with (10.23) for a solenoid with the same number of turns per unit length, and with the field $\mu_0 nI$ of an infinitely long one.

(d) The field at the center has a small y component, which a stack of flat circular loops would not produce. Show that it gets smaller as the coil is made longer with the same spacing of turns, and suggest where it comes from.

(e) Make a three-dimensional plot of the field vectors in and around the coil.

P10.22 A **Helmholtz pair** is two coaxial circular coils of radius R, a distance R apart, carrying the same current in the same sense (Figure 10.35). The field near its center is nearly uniform. On the axis the field is easy to find from (10.22); off the axis it is not.

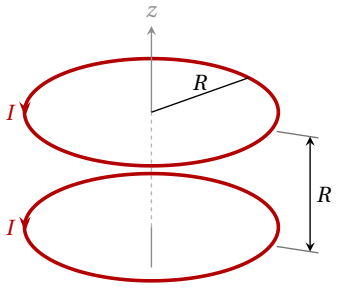


Figure 10.35 The Helmholtz pair of Exercise P10.22.

(a) Write a program that finds **B** at any point by adding the contributions (10.24) of N pieces of each loop. Test it on a single loop, against (10.22) on the axis and, off the axis, against the exact result

$$B_z = \frac{\mu_0 I}{2\pi\sqrt{(R+s)^2+z^2}}\left[K(m) + \frac{R^2 - s^2 - z^2}{(R-s)^2+z^2}E(m)\right],$$

and

$$B_s = \frac{\mu_0 I}{2\pi\sqrt{(R+s)^2+z^2}}\,\frac{z}{s}\left[-K(m)+\frac{R^2+s^2+z^2}{(R-s)^2+z^2}\,E(m)\right],$$

where $m = 4Rs/\left[(R+s)^2+z^2\right]$, s and z are measured from the center of the loop, and $K(m)$ and $E(m)$ are complete elliptic integrals (`EllipticK[m]` and `EllipticE[m]` in Mathematica, `scipy.special.ellipk(m)` and `ellipe(m)` in Python). How large must N be to give seven significant figures in the plane of the loop at $s = 0.5R$, $0.9R$, and $0.99R$?

(b) For the pair, find $|\mathbf{B}|/B(0)$ in the midplane at $s = 0.1R$, $0.2R$, $0.3R$, $0.4R$, and $0.5R$, where $B(0)$ is the field at the center.

(c) Find how far from the center, in the midplane and along the axis, the field strength differs from $B(0)$ by 1%, and by 0.1%. Also find $|\mathbf{B}|/B(0)$ at the point $s = 0.3R$, $z = 0.3R$. Is the region of nearly uniform field a sphere?

(d) For coils of radius 15 cm, each with 100 turns and a current of 1.00 A, find $B(0)$ and the size of the region in which the field is uniform to 1%.

Exercises for 10.8 Summary

P10.23 True or false? Explain your reasoning.

(a) If $\nabla\cdot\mathbf{J} > 0$ throughout a region, the charge density in that region is decreasing.

(b) To find the current carried by a surface current, integrate $\mathbf{K}$ along the direction in which the charges move.

(c) A long, flat strip carries a current along its length, spread uniformly across its width. Close to the strip, its magnetic field is parallel to the strip's surface at every point.

(d) A uniform magnetic field exerts no net force on a rectangular loop of current, whatever the field's direction.

(e) If the field of part (d) is parallel to two sides of the loop, the forces on the other two sides are equal and opposite, so the field exerts no torque on the loop.

(f) A charged particle spirals along a line of a static magnetic field into a region where the field is stronger. Its motion along the line slows, so its kinetic energy decreases.

(g) A charged particle in a uniform magnetic field always moves on a circle.

(h) If a particle's charge-to-mass ratio is doubled while its speed stays the same, the radius of its circular orbit in a given uniform field is halved.

P10.24 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as the motion of a charged particle in crossed electric and magnetic fields, the trapping of charged particles in Earth's magnetic field, the field lines of a circular loop of current, the region of nearly uniform field between a pair of Helmholtz coils, the field of a short solenoid wound as a helix, or the force between two loops of current.

Chapter 11

Ampère's Law

Chapter 10 found the magnetic field of a current distribution by adding up the fields of all of its pieces with the Biot–Savart law. That method always works, but the integrals can be long, and for the solenoid of Example 10.8 they gave the field only on its axis. This chapter takes the approach that Chapter 4 took for the electric field. Instead of asking what the field is at a particular point, it asks what the Biot–Savart law implies about the divergence and the curl of the field. For the fields predicted by the Biot–Savart law, the divergence is zero and the curl leads to Ampère's law. Another experimental fact goes with these results: sources other than currents can also give a magnetic field a curl (a changing electric field, as Chapter 16 explains), but no source that gives a magnetic field a divergence has ever been found. Like Gauss's law, Ampère's law has two main uses. Given a field, it gives the current density that produces it, with no integrals at all. And given a current distribution with enough symmetry, it gives the field in a few lines: inside a long solenoid, for instance, everywhere and not only on the axis. The boundary conditions that **B** obeys at a surface current close the chapter.



André-Marie Ampère (1775–1836, French) Ampère was largely self-taught, from the books in his father's library; his father was guillotined in 1793, during the Revolution. In September 1820 he heard of Hans Christian Ørsted's discovery that an electric current deflects a compass needle, and he soon showed that two parallel currents attract or repel each other. He went on to work out the forces between currents in general, and in an 1826 memoir he named the new science *electrodynamics*. Maxwell called him "the Newton of electricity." (Wikipedia)

11.1 Ampère's law

Sections 4.1 and 5.1 took the divergence and the curl of Coulomb's law, and the results, Gauss's law and $\nabla \times \mathbf{E} = 0$, made many problems far easier than the script-r integrals. This section does the same for the Biot–Savart law. The answers are the reverse of the electric ones: the magnetic field has no divergence, and its curl is proportional to the current density.

11.1.1 The divergence of B

Start from the volume form of the Biot–Savart law, (10.18); line and surface currents are special cases of volume currents, with delta functions in **J**. The divergence is taken with respect to the field point **r**, so it can be moved inside the

integral, where it acts on $\mathbf{J}(\mathbf{r}') \times \hat{\boldsymbol{\imath}}/\imath^2$:

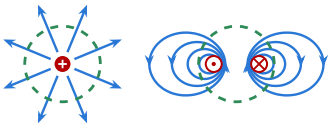
$$\nabla \cdot \mathbf{B} = \frac{\mu_0}{4\pi} \int \nabla \cdot \left(\mathbf{J}(\mathbf{r}') \times \frac{\hat{\boldsymbol{\imath}}}{\imath^2} \right) d\tau'.$$

By the product rule for the divergence of a cross product (Section 1.7), $\nabla \cdot (\mathbf{A} \times \mathbf{C}) = \mathbf{C} \cdot (\nabla \times \mathbf{A}) - \mathbf{A} \cdot (\nabla \times \mathbf{C})$,

$$\nabla \cdot \left(\mathbf{J}(\mathbf{r}') \times \frac{\hat{\boldsymbol{\imath}}}{\imath^2} \right) = \frac{\hat{\boldsymbol{\imath}}}{\imath^2} \cdot \left[\nabla \times \mathbf{J}(\mathbf{r}') \right] - \mathbf{J}(\mathbf{r}') \cdot \left[\nabla \times \frac{\hat{\boldsymbol{\imath}}}{\imath^2} \right].$$

Be careful about which variables each factor depends on. The current density depends only on the source point $\mathbf{r}'$, while ∇ takes derivatives with respect to the field point, so $\nabla \times \mathbf{J}(\mathbf{r}') = 0$. And $\hat{\boldsymbol{\imath}}/\imath^2$ is, apart from constants, the field of a point charge at $\mathbf{r}'$, whose curl Section 5.1 showed to be zero everywhere, with no delta function hiding at the charge. Both terms vanish, so

$$\nabla \cdot \mathbf{B} = 0. \qquad \text{(Gauss's law for magnetism)} \quad (11.1)$$



(a) (b)

Figure 11.1 (a) Electric field lines begin and end on charges; those of a positive charge cross a closed surface around it (dashed) only outward. (b) Magnetic field lines, here those of a current loop seen edge on, have no ends: every line that leaves a closed surface comes back in, and the flux out of the surface is zero.

With words, the magnetic field has no sources and no sinks: there is nothing for magnetic field lines to begin or end on (Figure 11.1). In particular, there is no magnetic counterpart of electric charge, a **magnetic monopole** from which **B** would point straight out, the way **E** points out from a point charge. We derived (11.1) for the fields of steady currents, but it is one of Maxwell's equations, and as far as anyone knows it holds for all magnetic fields, including those that change in time. People have searched for magnetic monopoles in cosmic rays, in particle accelerators, and in ancient rocks, and none has ever been found.

Like Gauss's law, (11.1) has an integral form. The **magnetic flux** through a surface is defined in the same way as the electric flux,

$$\Phi_B = \int_{\mathcal{S}} \mathbf{B} \cdot d\mathbf{a}, \quad (11.2)$$

and its SI unit is the **weber** ($1\,\mathrm{Wb} = 1\,\mathrm{T} \cdot \mathrm{m}^2$). Integrating (11.1) over any volume and using the divergence theorem gives

$$\oint_{\mathcal{S}} \mathbf{B} \cdot d\mathbf{a} = 0 : \quad (11.3)$$

the magnetic flux out of every closed surface is zero, as much field going in as coming out. One consequence will matter in later chapters. Any two surfaces bounded by the same closed loop together form a closed surface, so the flux through one of them equals the flux through the other. The magnetic flux through a loop is therefore the same for every surface that the loop bounds, and we can speak of the flux *through the loop*.

11.1.2 The curl of B

Start, as Section 5.1 did, with the simplest case, the field of a long straight wire on the z axis, $\mathbf{B} = (\mu_0 I/2\pi s)\,\hat{\boldsymbol{\phi}}$ from (10.20). Around a circle of radius s centered on the wire, $\mathbf{B}$ is parallel to the path and constant in magnitude, so

$$\oint \mathbf{B}\cdot d\boldsymbol{\ell} = \frac{\mu_0 I}{2\pi s}(2\pi s) = \mu_0 I,$$

whatever the radius: the field gets weaker farther out, but the circle gets longer in exactly the same proportion. In fact the same is true of any closed path that goes once around the wire (Figure 11.2a). Write a step along the path in cylindrical coordinates, $d\boldsymbol{\ell} = ds\,\hat{\mathbf{s}} + s\,d\phi\,\hat{\boldsymbol{\phi}} + dz\,\hat{\mathbf{z}}$, from (1.12). Then

$$\mathbf{B}\cdot d\boldsymbol{\ell} = \frac{\mu_0 I}{2\pi s}\,s\,d\phi = \frac{\mu_0 I}{2\pi}\,d\phi.$$

Only the part of each step that goes around the wire contributes, and it contributes in proportion to the angle $d\phi$ that it sweeps out, however far it is from the wire. Going once around the wire counterclockwise, as seen with the current coming toward you, ϕ increases by 2π, and

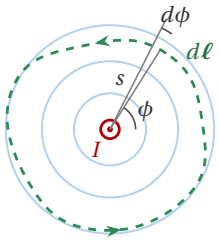
$$\oint \mathbf{B}\cdot d\boldsymbol{\ell} = \frac{\mu_0 I}{2\pi}\oint d\phi = \mu_0 I.$$

For a closed path that does not go around the wire (Figure 11.2b), ϕ comes back to its starting value, and the integral is zero. This is the same argument that Section 5.1 used for the field of a point charge, where only the radial part of each step contributed and the integral around every closed path was zero. Here only the azimuthal part contributes, and the integral around a path that encloses the current is not zero. By superposition, the integral of $\mathbf{B}$ around any closed path in the field of several long straight wires is μ_0 times the total current of the wires that the path encircles.

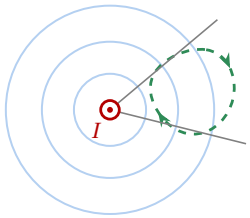
That suggests a general law, and it is one: for any distribution of steady current,

$$\nabla\times\mathbf{B} = \mu_0\mathbf{J}. \qquad \text{(Ampère's law, differential form)} \quad (11.4)$$

The derivation, from the curl of the Biot–Savart law, follows.



(a) a path around the wire



(b) a path that misses the wire

Figure 11.2 End view of a long wire carrying a current I out of the page. (a) Along any closed path around the wire, only the part $s\,d\phi$ of each step that goes around the wire contributes to $\mathbf{B}\cdot d\boldsymbol{\ell}$, and it contributes in proportion to the angle $d\phi$. Going once around adds up 2π. (b) Along a path that does not enclose the wire, ϕ increases and then decreases again, and the contributions cancel. Animation of a loop that moves and changes shape: its circulation is $\mu_0 I$ whenever it goes once around the wire, and zero whenever it does not enclose the wire.

The curl of the Biot–Savart law

Take the curl of (10.18). As with the divergence, the curl acts only on $\mathbf{J}(\mathbf{r}')\times\hat{\boldsymbol{\mathscr{r}}}/\mathscr{r}^2$:

$$\nabla\times\mathbf{B} = \frac{\mu_0}{4\pi}\int \nabla\times\left(\mathbf{J}(\mathbf{r}')\times\frac{\hat{\boldsymbol{\mathscr{r}}}}{\mathscr{r}^2}\right)d\tau'.$$

The product rule for the curl of a cross product (Section 1.7) has four terms, but two of them contain derivatives of $\mathbf{J}(\mathbf{r}')$ with respect to the field point, which are zero. What is left is

$$\nabla\times\left(\mathbf{J}(\mathbf{r}')\times\frac{\hat{\boldsymbol{\mathscr{r}}}}{\mathscr{r}^2}\right) = \mathbf{J}(\mathbf{r}')\left(\nabla\cdot\frac{\hat{\boldsymbol{\mathscr{r}}}}{\mathscr{r}^2}\right) - \left(\mathbf{J}(\mathbf{r}')\cdot\nabla\right)\frac{\hat{\boldsymbol{\mathscr{r}}}}{\mathscr{r}^2}.$$

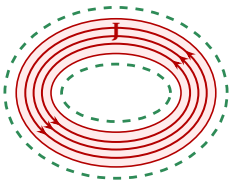


Figure 11.3 A steady current flowing around a closed circuit (shaded), and the surface (dashed) of a region slightly larger than the circuit. No current crosses that surface: $\mathbf{J}\cdot d\mathbf{a}' = 0$ everywhere on it.

The divergence in the first term is the delta function of (1.45), $4\pi\delta^3(\mathbf{r}-\mathbf{r}')$, so by the sifting property (1.42) the first term contributes

$$\frac{\mu_0}{4\pi}\int \mathbf{J}(\mathbf{r}')\,4\pi\delta^3(\mathbf{r}-\mathbf{r}')\,d\tau' = \mu_0\mathbf{J}(\mathbf{r}).$$

The second term must contribute nothing. The function $\hat{\boldsymbol{\imath}}/\imath^2$ depends on $\mathbf{r}$ and $\mathbf{r}'$ only through their difference $\mathbf{r}-\mathbf{r}'$, so a derivative with respect to x is the negative of the derivative with respect to x', and $(\mathbf{J}\cdot\nabla) = -(\mathbf{J}\cdot\nabla')$ when acting on it. Consider the x component, $f = (\hat{\boldsymbol{\imath}}/\imath^2)_x$. By the product rule (1.52), written with primed derivatives,

$$(\mathbf{J}\cdot\nabla')f = \nabla'\cdot(f\mathbf{J}) - f\,(\nabla'\cdot\mathbf{J}).$$

For steady currents, $\nabla'\cdot\mathbf{J} = 0$ by (10.13), and the last term vanishes. The integral of the first term over the volume of integration is, by the divergence theorem, an integral of $f\mathbf{J}$ over the surface of that volume. The volume has to contain all of the current, so it can be made slightly larger than the region where the current flows (Figure 11.3), and then no current crosses its surface, $\mathbf{J}\cdot d\mathbf{a}' = 0$ everywhere on it, and the surface integral is zero. The same argument works for the y and z components, which leaves $\nabla\times\mathbf{B} = \mu_0\mathbf{J}$.

Notice where the assumption of steady currents entered: through $\nabla'\cdot\mathbf{J} = 0$. For currents that are not steady, such as the current charging a capacitor, (11.4) is wrong, or at least incomplete. Maxwell found the missing term, and it is the subject of Section 16.1. Exercise P16.7 shows that the term dropped here supplies it exactly whenever the electric field is, at each instant, the field that the charges would produce by Coulomb's law.

With words, (11.4) says that the curl of the magnetic field at any point is μ_0 times the current density at that point. Magnetostatic fields have curl only where current flows. That can be surprising. The field lines of a long wire circle it everywhere, and it is natural to think that the field has curl everywhere. It does not: a small paddle wheel placed beside the wire does not turn. Around a small loop beside the wire, made of two arcs centered on the wire and two radial segments, the field is stronger along the inner arc, but the inner arc is shorter in exactly the same proportion, so the circulation is zero, as Section 1.4.3 showed for the field $\hat{\boldsymbol{\phi}}/s$ (Figure 1.14b). The only curl is on the wire itself, where it is a delta function: by (1.47), $\nabla\times(\mu_0 I\,\hat{\boldsymbol{\phi}}/2\pi s) = \mu_0 I\,\delta(x)\,\delta(y)\,\hat{\mathbf{z}}$, which is μ_0 times the current density of a thin wire on the z axis. Inside a thick wire with a uniform current density, the field turns out to grow in proportion to s (Example 11.1), and its curl is uniform throughout the wire, just as the field inside a uniformly charged ball has a uniform divergence.

11.1.3 The integral form

Integrate (11.4) over any surface $\mathcal{S}$, and use the curl theorem, (1.37), to turn the surface integral of the curl into a line integral around the closed loop that bounds the surface:

$$\int_{\mathcal{S}}(\nabla\times\mathbf{B})\cdot d\mathbf{a} = \oint \mathbf{B}\cdot d\boldsymbol{\ell} = \mu_0\int_{\mathcal{S}}\mathbf{J}\cdot d\mathbf{a}.$$

The last integral is the current I_{enc} that passes through the surface, the current *enclosed* by the loop, so

$$\oint \mathbf{B} \cdot d\boldsymbol{\ell} = \mu_0 I_{\text{enc}}. \qquad \text{(Ampère's law, integral form)} \quad (11.5)$$

A loop used this way is called an **Amperian loop**. The directions are linked by the right-hand rule of the curl theorem: if your fingers curl along the direction in which you go around the loop, your thumb points in the direction in which current counts as positive. Any surface bounded by the loop gives the same I_{enc}, for steady currents: two such surfaces together form a closed surface, and by (10.13) as much current flows into a closed surface as out of it. With words, (11.5) says that *the circulation of the magnetic field around any closed loop equals μ_0 times the current through the loop.* It holds for every loop, in the field of every steady current, just as Gauss's law holds for every closed surface.

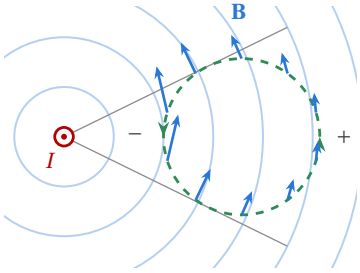


Figure 11.4 An Amperian loop (dashed) beside a long wire, seen end on, with the current out of the page. The loop encloses no current, but **B** (blue arrows, drawn in proportion to its magnitude) is nonzero everywhere on it. Going around the loop counterclockwise, $\mathbf{B} \cdot d\boldsymbol{\ell}$ is positive on the far arc (+) and negative on the near arc (−), which meet where the gray lines from the wire touch the loop. The net circulation is zero.

A caution, as with Gauss's law: a zero enclosed current means a zero circulation, *not* a zero field (Section 4.2.3). The loop of Figure 11.4 encloses no current, but **B** is nonzero everywhere on it. $\mathbf{B} \cdot d\boldsymbol{\ell}$ is positive along the far part of the loop and negative along the near part, and the two parts cancel, for the reason shown in Figure 11.2b. A loop in a uniform field is a simpler example: it encloses no current and has no circulation, but the field on it is not zero.

The law is named for Ampère, whose experiments and theory of 1820–1826 established the forces between currents, but this form of it was first written down by Maxwell, in 1855–56. It plays the same role in magnetostatics that Gauss's law plays in electrostatics. Its differential form tells you the current density if you know the field, with no integrals at all, and its integral form can be solved for the field when the current distribution has enough symmetry, which is the subject of the next section.

11.1.4 Maxwell's equations for static fields

Together with the results of Sections 4.1 and 5.1, this gives the divergence and the curl of both fields, for static charges and steady currents:

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0}, \qquad \nabla \times \mathbf{E} = 0, \qquad \nabla \cdot \mathbf{B} = 0, \qquad \nabla \times \mathbf{B} = \mu_0 \mathbf{J}.$$

In statics, electric fields have no curl and magnetic fields have no divergence. The sources of the electric field are charges, from which it spreads out; the sources of the magnetic field are currents, around which it curls. By the Helmholtz theorem (Section 2.3), the divergence and curl determine each field completely, so these four equations contain everything in Coulomb's law and the Biot–Savart law. Two of them, the two divergence equations, are true in general. The two curl equations hold only when nothing changes in time; their full forms, Faraday's law (2.3) and the Ampère–Maxwell law (2.4), are the subjects of Chapters 15 and 16. Chapter 14 sets all the electric and magnetic equations side by side.

11.2 Solving for B when given J: Ampère's law method (high symmetry)

Ampère's law in integral form, (11.5), holds for every closed loop in the field of every steady current, but like Gauss's law it can be solved for the field only in a few cases of high symmetry. To get **B** out of the integral, the Amperian loop has to be divisible into pieces on each of which one of two things is true:

1. **B** is parallel to the loop (parallel to $d\boldsymbol{\ell}$), and its magnitude is the same everywhere on that piece. Then $\int \mathbf{B}\cdot d\boldsymbol{\ell}$ over that piece is simply $\pm B\ell$, where ℓ is the piece's length.

2. **B** is perpendicular to the loop, or zero, everywhere on that piece. Then the piece contributes nothing.

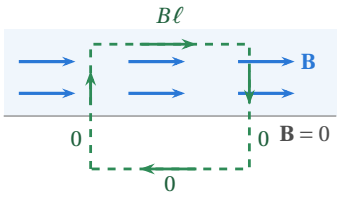


Figure 11.5 An Amperian loop (dashed) made of the two kinds of pieces. On the top side **B** is parallel to the loop and constant, so that side contributes $B\ell$; on the other sides **B** is perpendicular to the loop or zero, and they contribute nothing.

These are the conditions of Section 4.3 with one dimension fewer (Figure 11.5): a closed loop in place of a closed surface, and circulation in place of flux. Table 11.1 sets the two methods side by side.

	Gauss's law, $\oint \mathbf{E}\cdot d\mathbf{a} = q_{\text{enc}}/\varepsilon_0$	**Ampère's law**, $\oint \mathbf{B}\cdot d\boldsymbol{\ell} = \mu_0 I_{\text{enc}}$
Choose	a closed surface, divided into pieces	a closed loop, divided into pieces
The left side	the sum of the fluxes through the pieces	the sum of the circulations along the pieces
Simplify it	on each piece, either the flux is obviously zero, or **E** is parallel to $d\mathbf{a}$ and constant, so that $\int \mathbf{E}\cdot d\mathbf{a} = \pm EA$	on each piece, either the circulation is obviously zero, or **B** is parallel to $d\boldsymbol{\ell}$ and constant, so that $\int \mathbf{B}\cdot d\boldsymbol{\ell} = \pm B\ell$
The right side	find q_{enc}, often by integrating $\rho\, d\tau'$	find I_{enc}, often by integrating $\mathbf{J}\cdot d\mathbf{a}'$
Solve	for E	for B

Table 11.1 Finding a field with Gauss's law and with Ampère's law: the same steps, with a closed loop in place of a closed surface.

As with Gauss's law, you have to know the direction of the field, and what its magnitude can depend on, before you start: Ampère's law gives the magnitude, but the direction comes from symmetry. The procedure is:

1. Use symmetry to answer two separate questions, each with its own reason, as in Section 4.3.2: which components of **B** are zero, and which coordinates **B** can depend on.

2. Choose an Amperian loop that passes through the point where you want the field and that matches the symmetry, so that each piece of it satisfies condition 1 or 2 above. Draw it, list its pieces, and choose the direction in which you will go around it.

3. Calculate the enclosed current I_{enc}, if necessary by integrating $\mathbf{J}\cdot d\mathbf{a}'$ over a surface bounded by the loop (or $K_\perp\, d\ell'$ along the line where that surface

crosses a sheet of current). Current that passes through in the direction given by the right-hand rule counts as positive.

4. Write down Ampère's law, simplify the left side to B times a length, solve for B, and give **B** its direction.

For a long straight wire, the whole calculation takes two lines. The field circles the wire and depends only on the distance s from it (Example 11.1 gives the argument), so on a circle of radius s around the wire, **B** is parallel to $d\boldsymbol{\ell}$ and constant, and the left side of Ampère's law is $B(2\pi s)$. The enclosed current is I. So $B(2\pi s) = \mu_0 I$, and $\mathbf{B} = (\mu_0 I/2\pi s)\,\hat{\boldsymbol{\phi}}$, the result that took an integral in Example 10.5.

11.2.1 Colton's Third Commandment

Step 2 deserves the same emphasis that it did for Gauss's law. Ampère's law is about a specific closed loop, and an argument that never says what that loop is cannot check whether the field really is parallel and constant, or perpendicular, on each piece of it, or whether the loop catches the right current. This is **Colton's Third Commandment**:

> *When using Ampère's law to determine magnetic fields, thou shalt always draw an Amperian loop.*

Draw it explicitly, mark the direction in which you go around it, and account for every piece of it.

! **Colton's Third Commandment:** When using Ampère's law to determine magnetic fields, thou shalt always draw an Amperian loop.

11.2.2 The four geometries

There are only four kinds of current distribution for which this works, and therefore only four kinds of Amperian loop (Figure 11.6).

- **Current along the axis of a cylinder**: an infinitely long wire, or a cylinder or tube whose current density points along the axis and depends only on s. The field has only a ϕ component, so it circles the axis, and that component depends only on s: $\mathbf{B} = B(s)\,\hat{\boldsymbol{\phi}}$. The Amperian loop is a coaxial circle.
- **Current around the axis of a cylinder**: an infinitely long solenoid, or a cylinder whose current density points along $\hat{\boldsymbol{\phi}}$ and depends only on s. The field has only a z component, so it points along the axis, and that component depends only on s: $\mathbf{B} = B(s)\,\hat{\mathbf{z}}$. The Amperian loop is a rectangle with two sides parallel to the axis.
- **Planar symmetry**: an infinite plane of current, or a slab, centered on the xy plane, whose current density points in one fixed direction parallel to the plane and depends only on $|z|$. The field has only one component, parallel to the plane and perpendicular to the current, and that component depends only on z: it has the same magnitude at z and $-z$, in opposite directions on the two sides. The Amperian loop is a rectangle that straddles the plane symmetrically.

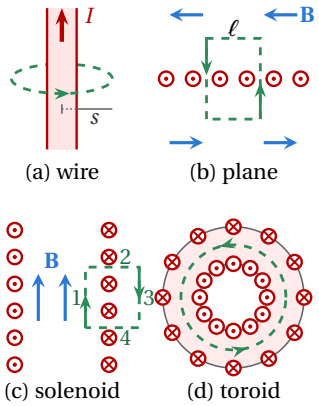


Figure 11.6 The four geometries in which Ampère's law can be solved for **B**, with their Amperian loops (dashed). (a) A long straight wire or cylinder with current along its axis: a coaxial circle. (b) An infinite plane of current, seen edge on, with the current coming out of the page: a rectangle straddling the plane. (c) A long solenoid, cut along its axis: a rectangle with one side inside and one outside, with its sides numbered as in Example 11.3. (d) A toroid seen from above: a circle inside the windings.

- **Toroidal symmetry**: a toroid, a solenoid bent around into a ring. The field has only a ϕ component, so inside the windings it circles the toroid's axis, and symmetry allows that component to depend on both s and z. The Amperian loop is a circle around that axis.

The first two are related. In a wire the current runs along the axis and the field circles it; in a solenoid the current circles the axis and the field runs along it. The roles are reversed.

Each item in this list answers the two questions of step 1 separately, as in Section 4.3.2, and the worked examples below give each answer its reason. For **B**, the usual reason that no piece of the current contributes to a component is the Biot–Savart law: each piece contributes a field perpendicular to itself. Keep the two questions apart here too. The toroid's field has only one component, and yet symmetry allows it to depend on two coordinates; that is enough for Ampère's law, because the Amperian circle needs the field to be the same only along the circle itself.

Note that, as with Gauss's law, the symmetry must be in the current distribution, not only in the shape of the object. A single circular loop of wire looks very symmetric, but its field lines (Figure 10.7) are neither circles around the wire nor straight lines. On any circle coaxial with the loop, **B** is perpendicular to the path (Figure 11.7), and Ampère's law says only that $0 = 0$; and no Amperian loop through a general point has **B** parallel and constant on some pieces and perpendicular on the rest. Ampère's law is still true for the loop, but it cannot be solved for the field: the field on its axis had to come from the Biot–Savart law, in Example 10.7. The same goes for a finite solenoid. For these, use the script-r method or its numerical version. (A finite piece of wire is different again. On its own it is not a steady current, and its field, (10.21), does not obey Ampère's law at all; only the field of a complete circuit does.)

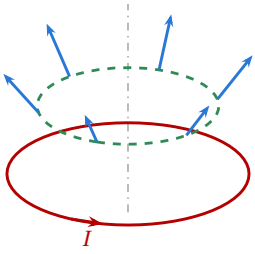


Figure 11.7 A current loop and a circle coaxial with it (dashed). At every point of the circle, **B** lies in the plane that contains the axis and has no component along the circle, so Ampère's law for the circle reads $0 = 0$.

11.2.3 Worked examples

Example 11.1

A long straight wire of radius R carries a current I, spread uniformly over its cross section. Find the magnetic field inside and outside the wire.

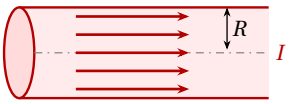
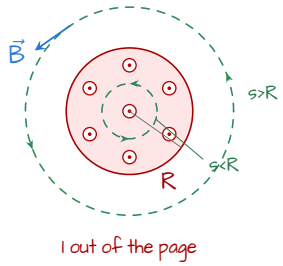


Figure 11.8 The wire of Example 11.1, with its current spread uniformly over the cross section.

Solution:



We begin by drawing what we are given: the wire end-on, with its current spread over the cross section and coming out of the page. As argued below, the field lines are circles around the axis, so the Amperian loops are circles of radius s, one inside the wire and one outside, traversed counterclockwise as seen here, with the current coming toward us, so that the current counts as positive.

First, the symmetry argument, in its two separate parts. The field can depend only on s: every value of z looks the same, because the wire is infinite, and every value of ϕ looks the same, because the wire is straight and round. As for its components, the field has no z component, because by the Biot-Savart law each piece of the current contributes a field perpendicular to itself, and every piece points along $\hat{\mathbf{z}}$. It has no s component either. If it did, a closed cylinder of radius s and length L, coaxial with the wire, would have a net flux $2\pi sL\,B_s$ through its curved side, on which B_s is the same everywhere, and none through its ends, which (11.3) forbids. So $\mathbf{B} = B(s)\,\hat{\boldsymbol{\phi}}$: the field lines are circles around the axis.

On each Amperian loop, then, $\mathbf{B}$ is parallel to $d\boldsymbol{\ell}$ and constant, and

$$\oint \mathbf{B}\cdot d\boldsymbol{\ell} = B\,(2\pi s).$$

The current density is $J = I/\pi R^2$. Inside the wire ($s < R$), the loop encloses the current that flows through a disk of radius s,

$$I_{\text{enc}} = J\pi s^2 = I\,\frac{s^2}{R^2},$$

so $B\,(2\pi s) = \mu_0 I s^2/R^2$, and

$$\mathbf{B} = \frac{\mu_0 I s}{2\pi R^2}\,\hat{\boldsymbol{\phi}}, \qquad s < R.$$

Outside ($s > R$), the loop encloses all of the current, and

$$\mathbf{B} = \frac{\mu_0 I}{2\pi s}\,\hat{\boldsymbol{\phi}}, \qquad s > R,$$

the same as for a thin wire. As with a uniformly charged ball and Gauss's law, the field outside is as if all of the current ran down the axis, and inside, the field grows in proportion to the distance from the axis (Figure 11.9).

Reality check: The field is continuous at $s = R$, as it should be, since there is no surface current there (Section 11.3). The curl in cylindrical coordinates (Section 1.4.6) gives,

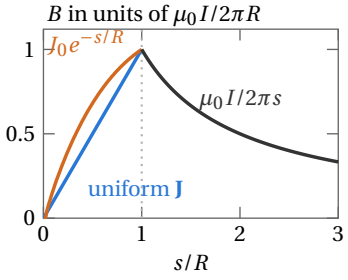


Figure 11.9 The field of a long wire of radius R carrying a current I. For a uniform current density (Example 11.1) it grows in proportion to s inside and falls as $1/s$ outside. For the current density $J_0 e^{-s/R}$ of Example 10.4, with the same total current, it rises faster near the axis. Outside, both are the field of a thin wire.

inside the wire,

$$\begin{aligned}\nabla\times\mathbf{B} &= \left(\frac{1}{s}\frac{\partial B_z}{\partial\phi}-\frac{\partial B_\phi}{\partial z}\right)\hat{\mathbf{s}}+\left(\frac{\partial B_s}{\partial z}-\frac{\partial B_z}{\partial s}\right)\hat{\boldsymbol{\phi}}\\ &\quad+\frac{1}{s}\left(\frac{\partial(sB_\phi)}{\partial s}-\frac{\partial B_s}{\partial\phi}\right)\hat{\mathbf{z}}\\ &= \frac{1}{s}\frac{\partial(sB_\phi)}{\partial s}\hat{\mathbf{z}}\\ &= \frac{1}{s}\frac{\mu_0 I}{2\pi R^2}(2s)\,\hat{\mathbf{z}}\\ &= \frac{\mu_0 I}{\pi R^2}\hat{\mathbf{z}}\\ &= \mu_0\mathbf{J},\end{aligned}$$

as Ampère's law requires. This is the field of Figure 1.14a, whose curl is the same everywhere.

Any current density that points along $\hat{\mathbf{z}}$ and depends only on s can be handled the same way; only I_{enc} changes. For the current density $J_0 e^{-s/R}$ of Example 10.4, the enclosed current is

$$I_{\text{enc}} = 2\pi J_0\int_0^s s'\,e^{-s'/R}\,ds' = 2\pi J_0 R^2\left[1-e^{-s/R}\left(1+\frac{s}{R}\right)\right],$$

and the field inside is $\mu_0 I_{\text{enc}}/2\pi s$. Notice the prime: inside the integral the radius is the integration variable s', which runs from 0 to the radius s of the Amperian loop. Figure 11.9 compares the two fields for the same total current.

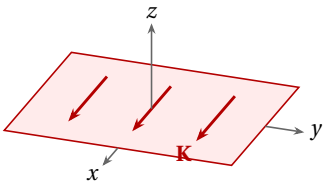
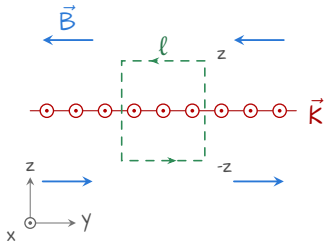


Figure 11.10 Part of the plane of current of Example 11.2.

Example 11.2

An infinite plane, the xy plane, carries a uniform surface current density $\mathbf{K} = K\hat{\mathbf{x}}$. Find the magnetic field everywhere.

Solution:



First, a sketch: we look at the sheet edge-on from the $+x$ side, so that the current comes out of the page. The field points along $-\hat{\mathbf{y}}$ above the sheet and along $+\hat{\mathbf{y}}$ below it, as argued below, so the Amperian loop is a rectangle that straddles the sheet, perpendicular to the current, with sides of length ℓ at heights z and $-z$.

Again the symmetry argument has two separate parts. First, which components are zero. The field has no x component, because every piece of the current points along $\hat{\mathbf{x}}$ and contributes a field perpendicular to itself. To find its other components, think of

the sheet as a great many parallel wires running in the x direction, side by side. At a point above the sheet, each wire's field circles that wire, so it points partly in the $-y$ direction and partly up or down, depending on which side of the point the wire is on. Pair each wire with the one the same distance away on the other side: their z components cancel, and their y components add. So above the sheet the field points along $-\hat{\mathbf{y}}$, and, by the same argument, below it the field points along $+\hat{\mathbf{y}}$. Second, what the field can depend on. Sliding the field point any distance in the x or y direction leaves the infinite sheet and its current looking exactly the same, so the field can depend only on z. A rotation by $180°$ about the x axis, the direction of the current, also leaves the sheet and its current unchanged, and it carries the point at height z to the point at $-z$, so the field has the same magnitude at z and $-z$.

Go around the loop counterclockwise as seen from the $+x$ side, as in the sketch, so that current in the $+x$ direction counts as positive. On the top side, $\mathbf{B}$ points along $-\hat{\mathbf{y}}$, and so does $d\boldsymbol{\ell}$; on the bottom side both point along $+\hat{\mathbf{y}}$. Each of these sides contributes $B\ell$. On the two vertical sides, $\mathbf{B}$ is perpendicular to $d\boldsymbol{\ell}$, and each contributes zero. (They do not cancel each other; each is zero by itself.) The enclosed current is $K\ell$, so $2B\ell = \mu_0 K\ell$, and $B = \mu_0 K/2$:

$$\mathbf{B} = \begin{cases} -\dfrac{\mu_0 K}{2}\hat{\mathbf{y}}, & z > 0, \\ +\dfrac{\mu_0 K}{2}\hat{\mathbf{y}}, & z < 0. \end{cases} \tag{11.6}$$

In vector form, $\mathbf{B} = (\mu_0/2)\,\mathbf{K} \times \hat{\mathbf{n}}$, where $\hat{\mathbf{n}}$ is the unit vector perpendicular to the sheet that points toward the field point. Like the electric field of an infinite plane of charge, (4.5), the field does not depend on the distance from the plane at all, and it reverses across the plane. The difference is in the direction: the electric field points away from a plane of positive charge, but the magnetic field is parallel to the plane of current (Figure 11.11).

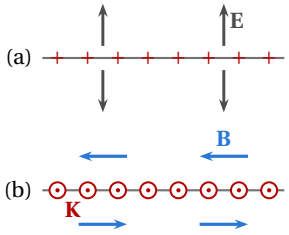


Figure 11.11 Seen edge on: (a) the electric field of a plane of positive charge points away from the plane; (b) the magnetic field of a plane of current, here coming out of the page, is parallel to the plane. Both reverse across the plane.

Reality check: The parallel wires give the same magnitude. Put the field point at $(0, 0, z)$, with $z > 0$. The strip of the sheet between y' and $y' + dy'$ is a wire that carries the current $K\,dy'$, a distance $\sqrt{y'^2 + z^2}$ away, so by (10.20) its field there has the magnitude $\mu_0 K\,dy'/2\pi\sqrt{y'^2 + z^2}$, and the fraction $z/\sqrt{y'^2 + z^2}$ of it is parallel to the sheet. Adding up the strips,

$$B = \frac{\mu_0 K}{2\pi}\int_{-\infty}^{\infty} \frac{z\,dy'}{y'^2 + z^2} = \frac{\mu_0 K}{2}.$$

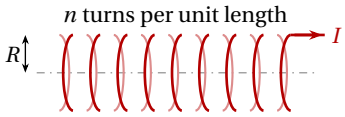
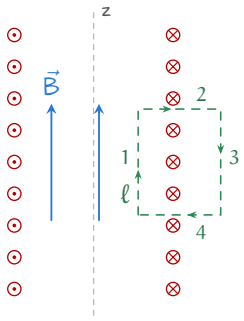


Figure 11.12 A few turns of the long solenoid of Example 11.3.

Example 11.3

A long solenoid of radius R has n turns of wire per unit length and carries a current I. Find the magnetic field inside and outside it, far from its ends.

Solution:



To get started, we draw the solenoid cut along its axis: the turns come out of the page on the left (dots) and go into it on the right (crosses), and inside, the field points along the axis. The Amperian loop in the sketch has side 1 inside the solenoid and side 3 outside it, and we go up side 1. Treat the solenoid as infinitely long, and treat the closely wound turns as a surface current $\mathbf{K} = nI\hat{\boldsymbol{\phi}}$ on a cylinder of radius R, with the z axis along the axis of the solenoid.[1] The symmetry argument has the same two parts.

The field can depend only on s, because the surface current looks the same at every z and at every ϕ. As for its components, it has no ϕ component: a circle of radius s around the axis encloses no current at all, because the current flows around the axis and never along it, so $2\pi s B_\phi = 0$. It has no s component either: a closed cylinder coaxial with the solenoid has equal and opposite fluxes through its two ends, because B_z does not depend on z, so by (11.3) the flux $2\pi sL B_s$ through its curved side must be zero. So $\mathbf{B} = B(s)\hat{\mathbf{z}}$.

Next, the field outside. A rectangular Amperian loop with two sides of length ℓ parallel to the axis, both outside the solenoid, at distances s_1 and s_2 from the axis, encloses no current. Its other two sides are perpendicular to $\mathbf{B}$ and contribute nothing, so $B(s_2)\,\ell - B(s_1)\,\ell = 0$: the field is the same at every distance outside the solenoid. It must in fact be zero, because the field outside a solenoid gets weaker the longer the solenoid is. The field lines that run up the inside have to come back down outside, and for a short coil they return close by (Figure 10.7). For a long coil they spread out over a much wider region before they get back around, and the field outside is correspondingly weak. Just outside the middle of a solenoid ten times as long as its radius, it is about 2% of the field inside, pointing the other way; for one 40 times as long, about 0.1%. In the limit of an infinite solenoid, the returning field lines spread over all of space, and the field outside is zero.

Now take the loop of the sketch, with side 1 inside the solenoid and side 3 outside, going up side 1. That is clockwise as drawn, so current into the page through the loop counts as positive. Side 1 contributes $B\ell$. Sides 2 and 4 are perpendicular to $\mathbf{B}$ inside the solenoid and lie in zero field outside it, and side 3 lies in zero field, so they contribute nothing. The

[1] A real winding is a helix, so besides going around, the current also progresses along the solenoid, from one end to the other; the solenoid carries a net current I along its length. Outside the solenoid this adds the weak field $\mu_0 I/2\pi s$ of a single straight wire, which we ignore.

loop encloses $n\ell$ turns, each carrying I through it, so $I_{enc} = n\ell I$. (Why is the enclosed current $n\ell I$, and not I or $2I$? Answer in footnote.[2]) So $B\ell = \mu_0 n\ell I$, and

$$\mathbf{B} = \begin{cases} \mu_0 n I \hat{\mathbf{z}}, & \text{inside,} \\ 0, & \text{outside.} \end{cases} \tag{11.7}$$

Side 1 can be at any distance from the axis, so the field inside is uniform, the same everywhere across the solenoid, which is what makes solenoids useful. A long solenoid is the magnetic counterpart of a parallel-plate capacitor (Example 4.6): a uniform field inside and, ideally, no field outside. In a real solenoid of finite length, the field lines that come out of one end return to the other outside the coil, and make a weak field there that is strongest near the ends.

Reality check: The field inside agrees with the on-axis field of a long finite solenoid, (10.23), far from its ends.

Example 11.4

A toroid is a coil of N turns wound around a ring-shaped core, like a solenoid bent around into a circle (Figure 11.6d). It carries a current I. Find the magnetic field inside the windings and outside them.

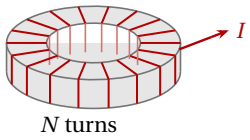
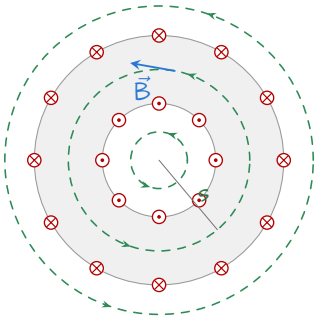


Figure 11.13 The toroid of Example 11.4, here with a rectangular cross section; only a few of its N turns are drawn.

Solution:



We sketch the toroid from above. The current goes up the inner side of each turn (dots) and down the outer side (crosses). The field circles the axis, as argued below, so the Amperian loops are circles of radius s around the axis: one inside the core, one in the hole, and one outside the toroid, all traversed in the $+\phi$ direction.

Let the z axis be the toroid's axis of symmetry, and treat the turns as closely spaced, so that the current is a smooth surface current over the surface of the core.[3] Each turn lies in a plane that contains the axis, so the current has no ϕ component, and the current distribution looks the same at every ϕ. So the field cannot depend on ϕ, though symmetry allows it to depend on s and z. As for its components, only B_ϕ is left: pair each

[2] The current passes through the loop once for every turn the loop encloses, into the page each time, so each of the $n\ell$ turns adds I. The other half of each turn comes back out of the page through the opposite wall, outside the loop.

[3] As with the solenoid, a real winding also progresses once around the ring, which adds the weak field of a single circular loop carrying I. We ignore it.

turn with its mirror image across the plane that contains the axis and the field point; their contributions to the field add up to a vector perpendicular to that plane, which is the $\hat{\boldsymbol{\phi}}$ direction at the field point (Exercise P11.9 works it out). The field therefore circles the axis, $\mathbf{B} = B(s,z)\,\hat{\boldsymbol{\phi}}$.

Each Amperian loop is a circle of radius s around the axis, at height z, traversed in the $+\phi$ direction. On it $\mathbf{B}$ is parallel to $d\boldsymbol{\ell}$ and constant, so $\oint \mathbf{B}\cdot d\boldsymbol{\ell} = B(2\pi s)$. Suppose the current goes up the inner side of each turn, along $+\hat{\mathbf{z}}$, and down the outer side. If the circle runs inside the core, the disk it bounds is crossed by the inner side of every turn, each carrying I, so $I_{\text{enc}} = NI$. If the circle runs through the hole in the middle of the toroid, no current crosses its disk. If it runs around the outside of the toroid, the current crosses its disk twice for every turn, once going up and once going down, for a total of zero; and if it runs above or below the toroid, no current crosses its disk at all. So

$$\mathbf{B} = \begin{cases} \dfrac{\mu_0 NI}{2\pi s}\,\hat{\boldsymbol{\phi}}, & \text{inside the core,} \\ 0, & \text{outside the core.} \end{cases} \tag{11.8}$$

The field is confined to the inside of the windings, whatever the shape of the core's cross section, round or rectangular. Unlike the field inside a long solenoid, it is not uniform: it falls off as $1/s$ across the core (Figure 11.14).

Reality check: If the toroid is large compared with its cross section, $N/2\pi s$ is nearly the number of turns per unit length around the ring, n, and the field is nearly $\mu_0 nI$, the field inside a long solenoid.

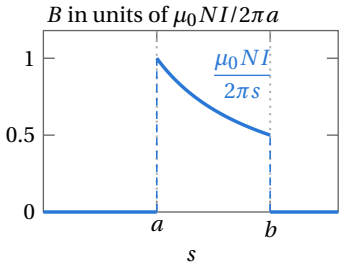


Figure 11.14 The field of the toroid, (11.8), against the distance s from its axis, for a core that runs from $s = a$ to $s = b = 2a$: zero in the hole and outside, and falling as $1/s$ across the core.

11.2.4 Superposition

As with Gauss's law, Ampère's law can sometimes be used when the current distribution as a whole has none of these symmetries, as long as each of its parts does: find the field of each part with its own Amperian loop, and add. Two parallel infinite sheets carrying equal and opposite surface currents, for example, give a field of magnitude $\mu_0 K$ between them and no field outside, just as two oppositely charged planes do. The caution at the end of Example 4.6 applies here too: a rectangle straddling just one of the two sheets does not have equal fields on its two long sides, because that symmetry belongs to a single sheet. Tutorial 11.1 practices the method on two kinds of long cylinder: a coaxial cable, whose two conductors carry currents along the axis in opposite directions, and a cylinder whose current circulates around its axis.

Tutorial 11.1 Ampère's law

Printable version

Background

The magnetostatics version of Ampère's law, (11.5), says that the circulation of **B** around a closed loop is $\mu_0 I_{\text{enc}}$, where I_{enc} is the amount of current passing through the surface defined by the loop. Just as Gauss's law can be used to determine **E** in certain limited situations of high symmetry, Ampère's law can be used to determine **B** in certain limited situations of high symmetry. The conditions required are quite similar to those for Gauss's law; the table in Section 11.2 sets them side by side. As with Gauss's law, the key is to properly identify the symmetry when the law can be used, and then to choose the correct loop for that symmetry.

Part 1 — Infinite cylinder, current in the z direction

Figure 11.15 shows the cross section of a long coaxial cable. The center conductor ($s < a$) carries a current I_0 out of the page. The outer conductor ($b < s < c$) carries the return current I_0 into the page. The current density **J** in each of the conductors is constant, although the two are not equal to each other.

(a) First consider each conductor separately. Sketch the center conductor with its current coming out of the page, and its magnetic field. On a second sketch, draw the outer conductor with its current going into the page, and its magnetic field. The total magnetic field will be a superposition of the fields you have drawn.

(b) Now consider the combined situation; sketch both currents on one figure. Why can you use Ampère's law to find **B** in this scenario? What changes to this scenario would make it so that Ampère's law could not be used to find it?

(c) What information are you given about the currents? Use this information to find the current density in the two regions. How does **J** being a vector make applying Ampère's law different from applying Gauss's law?

(d) Sketch an Amperian loop for the region $s < a$, and use it along with Ampère's law to find $\mathbf{B}(s)$ in this region.

(e) Now sketch another Amperian loop for the region $a < s < b$, and use it to find $\mathbf{B}(s)$ in this region.

(f) Sketch a third Amperian loop for the region $b < s < c$, and use it to find $\mathbf{B}(s)$ in this region.

(g) Without doing any calculations, what will $\mathbf{B}(s)$ be in the region $s > c$? Explain by applying Ampère's law.

(h) Sketch the magnitude of **B** as a function of s for all four regions. Specifically, consider whether it is (or should be) continuous. Can you use this as a partial check on your work?

Part 2 — Infinite cylinder, current in the ϕ direction

For this part, an infinite cylinder of radius R has a constant current density in the ϕ direction, $\mathbf{J} = J_0 \hat{\boldsymbol{\phi}}$ (Figure 11.16).

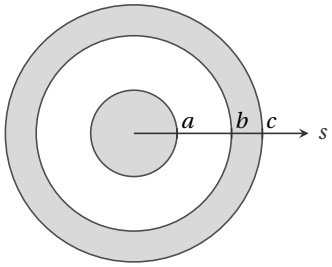


Figure 11.15 Cross section of the coaxial cable of Tutorial 11.1, Part 1. The center conductor ($s < a$) carries a current I_0 out of the page, and the outer conductor ($b < s < c$) carries the return current I_0 into the page.

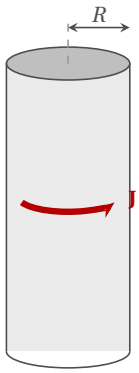


Figure 11.16 The long cylinder of Tutorial 11.1, Part 2, of radius R, with a uniform current density $\mathbf{J} = J_0 \hat{\boldsymbol{\phi}}$ circulating around its axis.

(a) The current here flows in the same direction as in a solenoid. In which ways is this similar to an infinite solenoid, and in which ways is it different?

(b) Which components of the magnetic field are zero, and why? Separately, which coordinates can it depend on, and why?

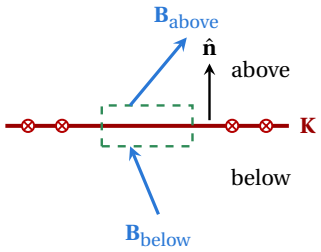
(c) Make an argument that the field outside the cylinder will be zero.

(d) Select an appropriate Amperian loop to find the field inside the cylinder. Explicitly divide the loop into sections, and make sure that each section has a trivial circulation integral, either because it is obviously zero or else because **B** is parallel to $d\ell$ and constant on that section, which makes the integral $B\ell$. Calculate the current passing through the loop, and solve for **B**.

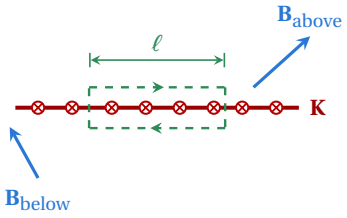
(e) What changes to the scenario would make this problem not doable with Ampère's law?

Part 3 — Wrap-up

(a) For your answer to 2(d), check to make sure it is reasonable, using at least one of the standard physics methods, such as limiting cases, units, or comparison with other problems.



(a) pillbox



(b) Amperian loop

Figure 11.17 A surface current **K** (into the page), seen edge on. (a) A short pillbox straddling it: the perpendicular components of the fields above and below must be equal. (b) A short, squat Amperian loop straddling it, in a plane perpendicular to **K**: the components of the fields along its long sides differ by $\mu_0 K$. The fields are drawn with the same perpendicular component and different parallel ones.

11.3 Boundary conditions on B

The fields of Examples 11.2 and 11.3 jump at a sheet of current, while the field of Example 11.1 does not jump anywhere. The divergence and curl of **B** say exactly how the field changes across any surface current, whatever its shape, just as Sections 4.4 and 5.2 did for **E** at a surface charge. The arguments are the same, but the roles of the two equations are reversed. Consider a small patch of a surface carrying the surface current density **K**, and let "above" be the side that a unit normal $\hat{\mathbf{n}}$ points toward (Figure 11.17).

First, draw a short pillbox that straddles the surface, with faces of area A just above and just below it. As the pillbox is made shorter, the flux through its wrapper goes to zero, and the flux out of its faces is $B^{\perp}_{\text{above}}A - B^{\perp}_{\text{below}}A$, where $B^{\perp}$ is the component of **B** along $\hat{\mathbf{n}}$. By (11.3), this must be zero. There is no magnetic counterpart of the surface charge that makes the perpendicular component of **E** jump, and so

$$B^{\perp}_{\text{above}} = B^{\perp}_{\text{below}}: \tag{11.9}$$

the perpendicular component of **B** is continuous across any surface, whatever current it carries.

Next, draw a short, squat Amperian loop that straddles the surface, in a plane perpendicular to **K**, with long sides of length ℓ just above and just below the surface (Figure 11.17b). As the loop is made shorter, the contributions of its short sides go to zero. Go around it in the direction for which **K** counts as positive: that means going along the upper side in the direction of $\mathbf{K}\times\hat{\mathbf{n}}$, and back along the lower side. The current through the loop is $K\ell$, and the two long sides, traversed in

opposite directions, contribute $B^{\parallel}_{\text{above}}\ell - B^{\parallel}_{\text{below}}\ell$, where $B^{\parallel}$ here is the component of $\mathbf{B}$ along $\mathbf{K}\times\hat{\mathbf{n}}$. So $B^{\parallel}_{\text{above}} - B^{\parallel}_{\text{below}} = \mu_0 K$. If the loop is instead turned to lie in a plane parallel to $\mathbf{K}$, no current passes through it, and the component of $\mathbf{B}$ along $\mathbf{K}$ is continuous. Altogether, the component of $\mathbf{B}$ that is parallel to the surface and perpendicular to the current jumps by $\mu_0 K$, and the component parallel to the current does not jump at all:

$$\mathbf{B}^{\parallel}_{\text{above}} - \mathbf{B}^{\parallel}_{\text{below}} = \mu_0\, \mathbf{K}\times\hat{\mathbf{n}}. \tag{11.10}$$

Combined with (11.9), the two boundary conditions can be written as one vector equation,

$$\mathbf{B}_{\text{above}} - \mathbf{B}_{\text{below}} = \mu_0\, \mathbf{K}\times\hat{\mathbf{n}}. \tag{11.11}$$

Compare (5.5). For $\mathbf{E}$, the perpendicular component jumps, by σ/ε_0, and the parallel components are continuous; for $\mathbf{B}$, the perpendicular component is continuous, and one of the parallel components jumps, by $\mu_0 K$. With words, a surface current makes the magnetic field jump in the direction that is parallel to the surface and perpendicular to the current. A jump in a component along a surface is a curl concentrated on the surface (Exercise P1.21), and here that curl is μ_0 times the surface current.

The examples of the previous section check these results. At the sheet of Example 11.2, with $\hat{\mathbf{n}} = \hat{\mathbf{z}}$, $\mathbf{K}\times\hat{\mathbf{n}} = K\,\hat{\mathbf{x}}\times\hat{\mathbf{z}} = -K\,\hat{\mathbf{y}}$, and the field jumps from $+(\mu_0 K/2)\,\hat{\mathbf{y}}$ below to $-(\mu_0 K/2)\,\hat{\mathbf{y}}$ above, a change of $-\mu_0 K\,\hat{\mathbf{y}}$. At the wall of the solenoid of Example 11.3, with $\hat{\mathbf{n}} = \hat{\mathbf{s}}$ pointing out, $\mathbf{K}\times\hat{\mathbf{n}} = nI\,\hat{\boldsymbol{\phi}}\times\hat{\mathbf{s}} = -nI\,\hat{\mathbf{z}}$, and the field jumps from $\mu_0 nI\,\hat{\mathbf{z}}$ inside to zero outside. At the surface of the thick wire of Example 11.1, which carries no surface current, the field is continuous. Like the boundary conditions on $\mathbf{E}$, these hold at any surface current, in any field, with or without symmetry. Section 13.9 extends them to the surfaces of magnetic materials.

11.4 Summary

The divergence and the curl of the Biot–Savart law are the reverse of those of Coulomb's law: the magnetic field has no divergence, and its curl is μ_0 times the current density, where the electrostatic field has a divergence and no curl. That curl equation is Ampère's law, which relates the field to the current in two equivalent ways.

$\nabla \cdot \mathbf{B} = 0, \quad \oint \mathbf{B} \cdot d\mathbf{a} = 0$	Gauss's law for magnetism: no magnetic monopoles; field lines never begin or end
$\Phi_B = \int \mathbf{B} \cdot d\mathbf{a}$	the magnetic flux through a surface, in webers
$\nabla \times \mathbf{B} = \mu_0 \mathbf{J}, \quad \oint \mathbf{B} \cdot d\boldsymbol{\ell} = \mu_0 I_{\text{enc}}$	Ampère's law, for steady currents: the field curls around currents, and has curl only where current flows
$B^{\perp}_{\text{above}} = B^{\perp}_{\text{below}}$, $\mathbf{B}^{\parallel}_{\text{above}} - \mathbf{B}^{\parallel}_{\text{below}} = \mu_0 \mathbf{K} \times \hat{\mathbf{n}}$	boundary conditions at a surface current

The Ampère's law method. Use symmetry to answer two separate questions: which components of the field are zero, and which coordinates it can depend on; choose an Amperian loop through the field point on which every piece has the field either parallel and constant or perpendicular (or zero); calculate I_{enc}; solve. There are only four geometries: current along the axis of a cylinder (a coaxial circle), current around the axis of a cylinder (a rectangle), a plane or slab (a rectangle straddling it), and a toroid (a circle around its axis). *Colton's Third Commandment:* when using Ampère's law to determine magnetic fields, thou shalt always draw an Amperian loop.

Results worth remembering. An infinite sheet: $\mu_0 K/2$, parallel to the sheet and independent of distance. A long solenoid: $\mu_0 nI$ inside, uniform, and zero outside. A toroid: $\mu_0 NI/2\pi s$ inside the windings, and zero outside.

Exercises

Exercises for 11.1 Ampère's law

Warm-up questions Read Section 11.1, then answer before class

W11.1 A classmate says that the field of a long solenoid must have a curl everywhere, because its field lines run up the inside and come back down outside, circling around the windings. Where is the curl of the solenoid's field actually not zero? Explain with (11.4), and with the circulation around a small loop inside the solenoid.

W11.2 An Amperian path winds twice around a long straight wire that carries a current I, like a stretched spring of two turns, and its two ends are joined by a straight piece parallel to the wire. What is $\oint \mathbf{B}\cdot d\boldsymbol{\ell}$ around it? Use the argument of Section 11.1.2 that only the part $s\,d\phi$ of each step contributes. What enclosed current does Ampère's law need here, and why?

W11.3 The text shows that, for steady currents, every surface bounded by an Amperian loop gives the same I_{enc}. Take a loop around the wire that carries current to one plate of a capacitor that is being charged, and two surfaces bounded by the loop: a flat disk that the wire passes through, and a bag that passes between the plates without touching the wire or either plate. What is I_{enc} for each? Which step of the argument for steady currents fails here?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P11.1 Which of the following could be a static magnetic field? For each one that could, find the current density that produces it. (k is a constant with suitable units.)

(a) $k(-y\hat{\mathbf{x}} + x\hat{\mathbf{y}})$

(b) $k(x\hat{\mathbf{x}} + y\hat{\mathbf{y}})$

(c) $ky^2\hat{\mathbf{x}}$

(d) $k(x^2 - y^2)\hat{\mathbf{z}}$

(e) $kz\hat{\mathbf{z}}$

(f) $k(yz\hat{\mathbf{x}} + xz\hat{\mathbf{y}} + xy\hat{\mathbf{z}})$

P11.2 A long straight wire on the z axis carries a current I in the $+z$ direction (Figure 11.18). Its field is (10.20).

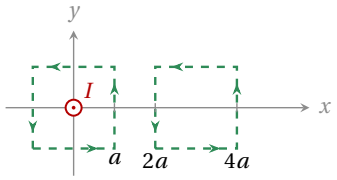


Figure 11.18 The wire and the two square paths of Exercise P11.2.

(a) Calculate the line integral of **B** directly, side by side, around the square with corners at $(\pm a, \pm a, 0)$, going counterclockwise as seen from $+z$.

(b) Repeat for the square with corners at $(2a, \pm a, 0)$ and $(4a, \pm a, 0)$, giving the contribution of each side as a multiple of $\mu_0 I$.

(c) Interpret both results with Ampère's law.

(d) Sketch the second square and the field lines of the wire near it, and mark the parts of the square where $\mathbf{B} \cdot d\boldsymbol{\ell}$ is positive and where it is negative. Use the sketch to explain how the contributions in (b) can add up to zero although **B** is nonzero all the way around.

Exercises for 11.2 Solving for B when given J: Ampère's law method (high symmetry)

Warm-up questions Read Section 11.2, then answer before class

W11.4 Could the method of this section find **B** for each of these, and why or why not: (a) a long straight wire whose current density grows from its axis toward its surface; (b) a single circular loop of wire; (c) a long straight wire whose current density is larger on one side of its axis than on the other?

W11.5 Example 11.1 shows that **B** has no s component by applying (11.3) to a closed cylinder coaxial with the wire. A long cylinder with a uniform charge density has the same symmetry. Why can the same argument not show that its electric field has no s component? Which law, applied to which closed path, shows instead that its electric field has no ϕ component?

W11.6 Suppose the wire of Example 11.1 runs along the axis of a long solenoid whose own field inside is $B_0\hat{\mathbf{z}}$, so that the field lines are helices. On the example's circular Amperian loop, is **B** still parallel to $d\boldsymbol{\ell}$? What does Ampère's law for that loop tell you, and what does it not tell you? How would you find the rest of the field?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

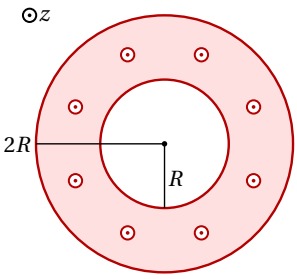


Figure 11.19 The hollow conductor of Exercise P11.3, in cross section.

P11.3 A long straight hollow conductor has inner radius R and outer radius $2R$ (Figure 11.19). It carries a total current I in the $+z$ direction, spread uniformly over the conducting material, with no surface current.

(a) Find **B** everywhere. Draw your Amperian loops.

(b) Sketch B as a function of s. Where is it largest?

(c) Show that B is continuous at $s = R$ and at $s = 2R$, and explain why it must be.

P11.4 A long wire of radius R carries the current density $\mathbf{J} = J_0(1 - s^2/R^2)\,\hat{\mathbf{z}}$, which is largest on the axis and falls to zero at the surface (Figure 11.20).

(a) Find the total current.

(b) Find **B** inside and outside the wire.

(c) Show that the field is largest inside the wire, at $s = \sqrt{2/3}\,R$, and find the ratio of its largest value to its value at the surface.

(d) Check that your field satisfies $\nabla \times \mathbf{B} = \mu_0 \mathbf{J}$ inside the wire.

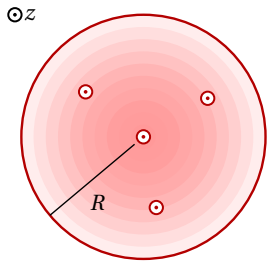


Figure 11.20 The wire of Exercise P11.4, in cross section. The shading is darker where the current density is larger.

P11.5 An infinitely long thick cylindrical shell, $a < s < b$, carries the current density $\mathbf{J} = J_0(s^2/b^2)\,\hat{\mathbf{z}}$, and its outer surface, $s = b$, also carries the surface current density $\mathbf{K} = K_0\,\hat{\mathbf{z}}$ (Figure 11.21). There is no current inside $s = a$.

(a) Find **B** in all three regions.

(b) Check that the jump in **B** at $s = b$ agrees with the boundary condition (11.11) (Section 11.3).

(c) What value of K_0 makes the field outside the shell zero? Where is the field nonzero then?

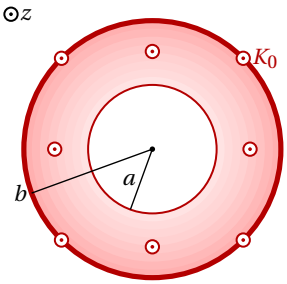


Figure 11.21 The shell of Exercise P11.5, in cross section. The shading is darker where the current density is larger, and the heavy circle carries the surface current.

P11.6 A long cylinder of radius a, with its axis along z, carries the current density $\mathbf{J} = J_0\,(s/a)\,\hat{\boldsymbol{\phi}}$, circulating around its axis (Figure 11.22). (Part 2 of Tutorial 11.1 had a uniform current density instead.)

(a) Argue that $\mathbf{B} = B(s)\,\hat{\mathbf{z}}$, and that the field outside the cylinder is zero.

(b) Use a rectangular Amperian loop to find **B** inside the cylinder.

(c) Compare the field on the axis with the field inside an ideal solenoid that carries the same current per unit length.

(d) Check that your field satisfies $\nabla \times \mathbf{B} = \mu_0 \mathbf{J}$.

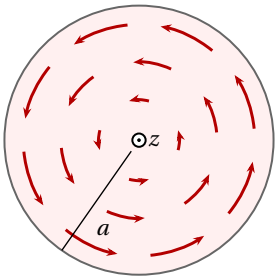


Figure 11.22 The cylinder of Exercise P11.6, in cross section, with its circulating current.

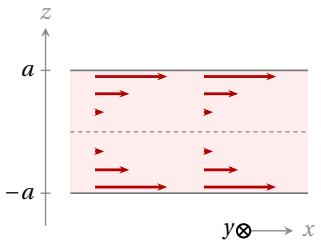


Figure 11.23 The slab of Exercise P11.7, edge on. The arrows show the current density.

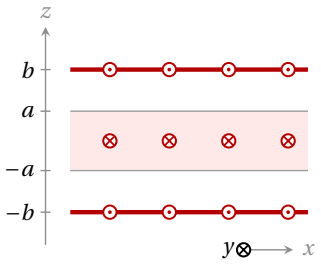


Figure 11.24 The strip line of Exercise P11.8, edge on.

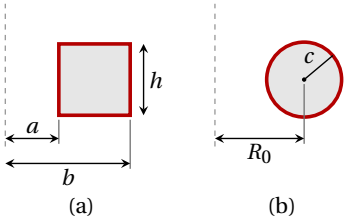


Figure 11.25 The toroid's core in parts (a) and (b) of Exercise P11.10, cut by a plane through its axis (dashed). The winding is red.

P11.7 A slab, $-a < z < a$, infinite in x and y, carries the current density $\mathbf{J} = J_0(z/a)^2\,\hat{\mathbf{x}}$ (Figure 11.23).

(a) Use symmetry to find the direction of **B** above, inside, and below the slab.

(b) Use Ampère's law to find **B** everywhere.

(c) Show that outside the slab the field is the same as that of an infinitely thin sheet, (11.6), carrying the same current per unit width.

P11.8 An idealized strip line consists of a conducting slab, $-a < z < a$, infinite in x and y, carrying the uniform current density $\mathbf{J} = J_0\,\hat{\mathbf{y}}$, between two thin conducting sheets at $z = \pm b$ ($b > a$) that carry the return current, each with the surface current density $\mathbf{K} = -J_0 a\,\hat{\mathbf{y}}$ (Figure 11.24).

(a) Find **B** in all five regions, and sketch B_x as a function of z.

(b) Check the boundary conditions of Section 11.3 at $z = a$ and at $z = b$.

(c) If the slab is ℓ wide (in the x direction) and carries the total current I, express the field between the slab and the sheets in terms of I and ℓ, and evaluate it for $I = 10\,\text{A}$ and $\ell = 5.0\,\text{mm}$. (Treat the strip line as infinitely wide.)

P11.9 This exercise fills in the symmetry argument of Example 11.4. In a toroid, the current flows in planes that contain the z axis, so the current density has no ϕ component, and it looks the same at every ϕ: $\mathbf{J} = J_s(s,z)\,\hat{\mathbf{s}} + J_z(s,z)\,\hat{\mathbf{z}}$. Let the field point P be the point $(x,0,z)$, in the xz plane.

(a) Consider two small pieces of the current, of equal volume $d\tau'$, at the points with cylindrical coordinates (s',ϕ',z') and $(s',-\phi',z')$, which are mirror images of each other across the xz plane. Sketch the two pieces, their current densities, and P, as seen from above. Then write the current density at each piece, and the separation vector $\boldsymbol{\imath}$ from each piece to P, in Cartesian components.

(b) Show that the sum of their contributions to **B** at P, from the Biot–Savart law, has only a y component.

(c) Explain why this means that $\mathbf{B} = B(s,z)\,\hat{\boldsymbol{\phi}}$ everywhere, inside and outside the toroid.

P11.10 A toroid has N turns and carries a current I, so that the field inside its core is (11.8).

(a) The core has a rectangular cross section, with inner radius a, outer radius b, and height h (Figure 11.25a). Find the magnetic flux through one turn. Evaluate the field at the inner and outer walls, and the flux, for $N = 1000$, $I = 0.50\,\text{A}$, $a = 5.0\,\text{cm}$, $b = 8.0\,\text{cm}$, and $h = 2.0\,\text{cm}$.

(b) Now the core has a circular cross section of radius c, centered a distance $R_0 > c$ from the axis (Figure 11.25b). Show that the flux through one turn is $\Phi = \mu_0 N I\big(R_0 - \sqrt{R_0^2 - c^2}\big)$. You may use

$$\int_{R_0-c}^{R_0+c} \frac{\sqrt{c^2-(s-R_0)^2}}{s}\,ds = \pi\left(R_0 - \sqrt{R_0^2-c^2}\right).$$

(c) Show that for $c \ll R_0$ the flux is approximately $B(R_0)\,\pi c^2$, the field at the center of the cross section times its area.

(d) For $N = 500$, $I = 2.00\,\text{A}$, $R_0 = 10.0\,\text{cm}$, and $c = 2.0\,\text{cm}$, find B at $s = 8.0$, 10.0, and 12.0 cm, and the error of the approximation in part (c).

P11.11 A straight segment of wire of length L carries a current I (Figure 11.26). Section 11.2.2 says that its field, (10.21), does not obey Ampère's law. Check this.

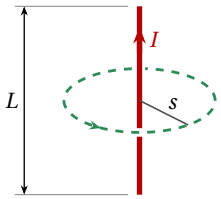


Figure 11.26 The segment of wire and the circle of Exercise P11.11.

(a) Find the circulation of the field around a circle of radius s, centered on the segment's midpoint and perpendicular to it. Is it $\mu_0 I$?

(b) What does the circulation approach for $s \ll L$, and for $s \gg L$?

(c) What is it about a lone segment of current that makes Ampère's law fail?

Exercises for 11.3 Boundary conditions on B

Warm-up questions Read Section 11.3, then answer before class

W11.7 Before it finds the size of the jump in **B** at a surface current, this section shows that the component of **B** perpendicular to the surface and the component along **K** are both continuous. Using only those two facts, explain why any jump in **B** must point along $\mathbf{K} \times \hat{\mathbf{n}}$ (or its negative). Compare the argument of Section 10.2 for the cross product in the field of a short piece of current.

W11.8 Answer each from the boundary conditions of this section, with a reason or an example from this chapter. (a) Can **B** be zero on both sides of a surface that carries a current? (b) Can it be zero on one side only? (c) Can **B** be parallel to **K** on both sides?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P11.12 The xy plane carries a surface current density **K**. Other currents are present too, so the fields near the sheet are not just the sheet's own.

(a) $\mathbf{K} = K_0\,\hat{\mathbf{x}}$, and just above the sheet $\mathbf{B} = B_0\,\hat{\mathbf{x}} + B_1\,\hat{\mathbf{z}}$. Find $\mathbf{B}$ just below it.

(b) Just above the sheet, $\mathbf{B} = (2.0\,\hat{\mathbf{x}} + 3.0\,\hat{\mathbf{y}} + 1.0\,\hat{\mathbf{z}})$ mT, and just below it, $\mathbf{B} = (-1.0\,\hat{\mathbf{x}} + 3.0\,\hat{\mathbf{y}} + 1.0\,\hat{\mathbf{z}})$ mT. Find $\mathbf{K}$.

(c) Could $(1.0\,\hat{\mathbf{x}} + 2.0\,\hat{\mathbf{z}})$ mT just above and $(1.0\,\hat{\mathbf{x}} + 3.0\,\hat{\mathbf{z}})$ mT just below be the fields on the two sides of a sheet of current? Explain.

Exercises for 11.4 Summary

P11.13 True or false? Explain your reasoning.

(a) Ampère's law, $\oint \mathbf{B} \cdot d\boldsymbol{\ell} = \mu_0 I_{\text{enc}}$, holds only for current distributions with a great deal of symmetry.

(b) Outside an ideal, infinitely long solenoid, B falls off as $1/s$.

(c) If a closed loop encircles no current, and $\int \mathbf{B} \cdot d\boldsymbol{\ell}$ is positive along part of the loop, then it is negative along another part.

(d) The net magnetic flux out of a sphere centered on one end of a long solenoid is zero.

(e) Across a surface current, every component of $\mathbf{B}$ parallel to the surface is continuous.

(f) Across a surface current, the component of $\mathbf{B}$ perpendicular to the surface is continuous.

(g) Ampère's law can be used to find the field at the center of a circular loop of current.

(h) Doubling the radius of a circular loop, at the same current, halves the field at its center but multiplies the field far away on its axis by four.

P11.14 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as the current density found by taking the curl of a map of the field that contains measurement noise, the circulation of the field of a wire tied in a trefoil knot around closed paths that thread the knot in different ways, the field inside a pipe that carries a current along its length and has a lengthwise slit cut in it, the twisted field lines that a tokamak's ring of coils and the current in its plasma make together, or the inward pull that the current in a plasma column exerts on the column itself (the pinch effect) and the pressure that balances it.

Chapter 12

Vector Potential

Chapter 5 began with one fact about the electrostatic field, that its curl is zero, and drew a long list of consequences from it, the most useful being that the field is the negative of the gradient of a scalar potential. This chapter does the same for the magnetic field, starting from the other half of what Chapter 11 found: the divergence of **B** is zero. A field with no divergence can be written as the curl of another vector field, the **magnetic vector potential A**.

The vector potential is less simple than the scalar potential in two ways. It is a vector, with three components to find instead of one, and it is not pinned down by the magnetic field even after something like a reference point has been chosen. But its components behave in a very familiar way. Each Cartesian component of **A** is produced by the same component of the current density exactly as V is produced by the charge density, and as a result much of Chapters 5, 7, and 8 carries over: the script-r integral, Poisson's equation and the relaxation method, the boundary conditions at a surface, and the multipole expansion. The last of these gives the field of a **magnetic dipole**, which Chapter 10 needed several times and could not yet derive. The chapter ends with a brief look at inductance and at the energy stored in a magnetic field, both of which wait for Faraday's law for their full treatment, and with a triangle diagram for **J**, **B**, and **A** to match the one for ρ, **E**, and V.

In magnetostatics the vector potential is mostly a convenience: it often makes a field easier to calculate, and it organizes the equations. When fields change in time it becomes a necessity, because the electric field is then no longer $-\nabla V$ alone (Section 15.4), and in quantum mechanics it appears in the equations of motion themselves.

! **Two potentials.** A field with zero curl is a gradient field, and a field with zero divergence is a curl:

$$\nabla \times \mathbf{E} = 0 \;\Rightarrow\; \mathbf{E} = -\nabla V,$$
$$\nabla \cdot \mathbf{B} = 0 \;\Rightarrow\; \mathbf{B} = \nabla \times \mathbf{A}.$$

12.1 The magnetic vector potential

The two differential equations of magnetostatics, from Section 11.1, are

$$\nabla \cdot \mathbf{B} = 0, \qquad \nabla \times \mathbf{B} = \mu_0 \mathbf{J}.$$

In Chapter 5 the statement $\nabla \times \mathbf{E} = 0$ led to $\mathbf{E} = -\nabla V$ (Section 5.3.1), and the identity that the curl of a gradient field is zero, (1.56), shows that every field of that form is curl-free. The first equation above pairs in the same way with the other identity of Section 1.7: the divergence of a curl is zero, (1.57). If

$$\mathbf{B} = \nabla \times \mathbf{A} \tag{12.1}$$

for some vector field $\mathbf{A}$, then $\nabla \cdot \mathbf{B} = \nabla \cdot (\nabla \times \mathbf{A}) = 0$ automatically. The field $\mathbf{A}$ is called the **magnetic vector potential**, or just the **vector potential**; to tell the two apart, V is sometimes called the **scalar potential**.

The converse is also true: a field whose divergence is zero everywhere, with no exceptions, can always be written as the curl of some other field. For a field that goes to zero far away, the Helmholtz theorem of Section 2.3 even constructs one. With $\nabla \cdot \mathbf{B} = 0$ and $\nabla \times \mathbf{B} = \mu_0 \mathbf{J}$, it gives (2.5) with $U = 0$ and

$$\mathbf{W}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int \frac{\mathbf{J}(\mathbf{r}')}{\mathscr{r}} \, d\tau',$$

so $\mathbf{B} = \nabla \times \mathbf{W}$. This $\mathbf{W}$ is the vector potential that the rest of the chapter uses most, and Section 12.2.2 arrives at it again by a different route. Fields that do not go to zero far away can be written as curls too; the uniform field of Example 12.1 is one. The words "with no exceptions" matter. The field $\hat{\mathbf{r}}/r^2$ has zero divergence everywhere except at the origin, where its divergence is $4\pi\delta^3(\mathbf{r})$ by (1.44), and it cannot be written as the curl of any field that is finite away from the origin: its flux through a sphere centered on the origin is 4π, but by the curl theorem the flux of a curl through a closed surface, which has no boundary curve, is zero (Figure 12.1). It is the field that a magnetic monopole would have, and $\nabla \cdot \mathbf{B} = 0$ says there are none.

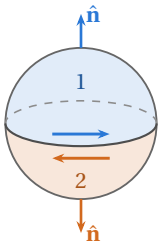


Figure 12.1 Why the flux of a curl through a closed surface is zero. Cut the surface into two caps, 1 and 2, each with its outward normal. By the curl theorem the flux of $\nabla \times \mathbf{A}$ through each cap is the circulation of $\mathbf{A}$ around the cut, taken in opposite directions for the two caps (arrows), so the two fluxes cancel.

Since $\mathbf{B}$ is measured in teslas, (12.1) gives $\mathbf{A}$ the units of tesla times meter, $\mathrm{T} \cdot \mathrm{m}$, which is the same as a weber per meter, or a newton per ampere.[1]

12.1.1 Gauge freedom and the Coulomb gauge

The potential V is defined only up to an added constant: V and $V + C$ give the same field, which is why a reference point has to be chosen (Section 5.3.2). The vector potential is less determined still. If ψ is any single-valued scalar function, then

$$\mathbf{A}' = \mathbf{A} + \nabla\psi \tag{12.2}$$

has the same curl as $\mathbf{A}$, because the curl of a gradient field is zero:

$$\nabla \times \mathbf{A}' = \nabla \times \mathbf{A} + \nabla \times (\nabla\psi) = \nabla \times \mathbf{A}.$$

[1] A newton per ampere is also a unit of momentum per unit charge, $(\mathrm{kg} \cdot \mathrm{m/s})/\mathrm{C}$, and that is no accident. If the currents that produce $\mathbf{A}$ at the location of a charge q at rest are switched off, the electric field that the changing magnetic field induces gives the charge a momentum $q\mathbf{A}$ (Example 15.6), with $\mathbf{A}$ in the Coulomb gauge of Section 12.1.1, going to zero far away. In 1865 Maxwell called $\mathbf{A}$ the "electromagnetic momentum."

So $\mathbf{A}$ and $\mathbf{A}'$ describe exactly the same magnetic field. Adding a constant vector $\mathbf{c}$ to $\mathbf{A}$ is a special case, with $\psi = \mathbf{c}\cdot\mathbf{r}$. A particular choice of ψ, and with it of $\mathbf{A}$, is called a choice of **gauge**, and the replacement (12.2) is a **gauge transformation**.[2]

Since every gauge gives the same field, we are free to pick whichever makes $\mathbf{A}$ simplest. The usual choice in magnetostatics is the **Coulomb gauge**,

$$\nabla\cdot\mathbf{A} = 0. \qquad \text{(Coulomb gauge)} \quad (12.3)$$

This choice can always be made. Suppose a vector potential $\mathbf{A}_0$ has the right curl but a divergence $\nabla\cdot\mathbf{A}_0 = f(\mathbf{r})$ that is not zero. Then $\mathbf{A} = \mathbf{A}_0 + \nabla\psi$ has the same curl, and its divergence is $f + \nabla^2\psi$, so we need a function ψ with

$$\nabla^2\psi = -f.$$

That is Poisson's equation (5.26), with f in place of ρ/ε_0, and we already know a solution: by (5.16), $\psi = (1/4\pi)\int f(\mathbf{r}')/\mathscr{r}\, d\tau'$, provided f goes to zero fast enough far away. In a sense, choosing the Coulomb gauge does for $\mathbf{A}$ what choosing a reference point does for V: it removes most of the arbitrariness. It does not remove all of it, though. A gradient field $\nabla\psi$ with $\nabla^2\psi = 0$ can still be added without spoiling (12.3), as the next example shows. For currents confined to a finite region, the further requirement that $\mathbf{A}$ go to zero far away removes that last freedom, as the corresponding requirement does for V.

The name refers to fields that change in time. In the Coulomb gauge the scalar potential still obeys $\nabla^2 V = -\rho/\varepsilon_0$ at every instant, so it is given at each moment by Coulomb's law (Section 15.4). Other gauges are more convenient for other problems; for radiation, the usual choice is the Lorenz gauge (Section 24.1.3), named for Ludvig Lorenz, whom we met in Section 9.4.1.

Example 12.1

A uniform magnetic field $\mathbf{B} = B_0\,\hat{\mathbf{z}}$ fills all of space. Show that

$$\mathbf{A}_1 = \frac{B_0}{2}\left(-y\,\hat{\mathbf{x}} + x\,\hat{\mathbf{y}}\right) \qquad \text{and} \qquad \mathbf{A}_2 = B_0\,x\,\hat{\mathbf{y}}$$

are both vector potentials for this field, and that both are in the Coulomb gauge. Find a function ψ that connects them.

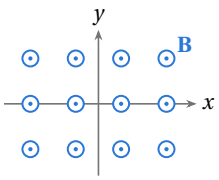


Figure 12.2 The uniform field of Example 12.1, seen in the xy plane: it points out of the page everywhere.

[2] The word comes from Hermann Weyl. In 1918 and 1919 he proposed a theory in which the scale of length, the "gauge" (*Eich-* in German), could change from place to place. The theory did not survive, but when a similar kind of freedom turned up in quantum mechanics, for the phase of the wave function, Weyl kept the name, calling the new freedom gauge invariance (*Eichinvarianz*) in 1929. See "Historical roots of gauge invariance," J. D. Jackson and L. B. Okun, Rev. Mod. Phys. 73, 663–680 (2001). https://doi.org/10.1103/RevModPhys.73.663.

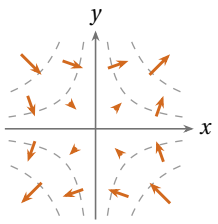
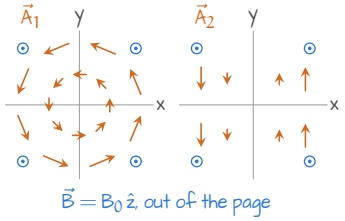


Figure 12.3 The difference between the two potentials of Example 12.1, $\mathbf{A}_2 - \mathbf{A}_1 = \nabla\psi$ with $\psi = B_0 xy/2$ (arrows), crosses the contours of ψ (dashed) at right angles. It has neither curl nor divergence, so adding it changes $\mathbf{A}$ but not $\mathbf{B}$, and it keeps the Coulomb gauge.

Solution:



We start by sketching both potentials in the xy plane, where $\mathbf{B}$ points out of the page everywhere. $\mathbf{A}_1$ circles the origin and grows with the distance from it; $\mathbf{A}_2$ points up the page to the right of the y axis and down it to the left, growing with the distance from that axis.

Take the curls, in Cartesian coordinates. For $\mathbf{A}_1$, only the z component survives:

$$\begin{aligned}\nabla\times\mathbf{A}_1 &= \left(\frac{\partial A_{1y}}{\partial x} - \frac{\partial A_{1x}}{\partial y}\right)\hat{\mathbf{z}} \\ &= \left(\frac{B_0}{2} + \frac{B_0}{2}\right)\hat{\mathbf{z}} \\ &= B_0\,\hat{\mathbf{z}}.\end{aligned}$$

For $\mathbf{A}_2$, likewise, $\nabla\times\mathbf{A}_2 = (\partial A_{2y}/\partial x)\,\hat{\mathbf{z}} = B_0\,\hat{\mathbf{z}}$. Both divergences are zero, since A_{1x} does not depend on x, A_{1y} does not depend on y, and A_{2y} does not depend on y. Their difference is

$$\mathbf{A}_2 - \mathbf{A}_1 = \frac{B_0}{2}\left(y\hat{\mathbf{x}} + x\hat{\mathbf{y}}\right) = \nabla\left(\frac{B_0\,xy}{2}\right),$$

so $\psi = B_0 xy/2$, plus any constant (Figure 12.3). As it must be, $\nabla^2\psi = 0$. (Why must ψ satisfy Laplace's equation? Answer in footnote.[3])

The first of these is the more symmetric. In cylindrical coordinates, $-y\hat{\mathbf{x}} + x\hat{\mathbf{y}} = s\hat{\boldsymbol{\phi}}$, so $\mathbf{A}_1 = \frac{1}{2}B_0 s\hat{\boldsymbol{\phi}}$: it circles the z axis and grows in proportion to the distance from it. In vector form, $\mathbf{A}_1 = \frac{1}{2}\mathbf{B}\times\mathbf{r}$. Neither potential goes to zero far away, any more than the potential of a uniform electric field does, because the field itself extends to infinity.

Reality check: Taking the curl of the vector form confirms the result, for a uniform field in any direction. Since $\mathbf{B}$ is constant, two of the four terms of the product rule for the curl of a cross product (Section 1.7) are zero, and with $\nabla\cdot\mathbf{r} = 3$ and $(\mathbf{B}\cdot\nabla)\mathbf{r} = \mathbf{B}$,

$$\begin{aligned}\nabla\times\left(\tfrac{1}{2}\mathbf{B}\times\mathbf{r}\right) &= \tfrac{1}{2}\left[\mathbf{B}\,(\nabla\cdot\mathbf{r}) - (\mathbf{B}\cdot\nabla)\mathbf{r}\right] \\ &= \tfrac{1}{2}\,(3\mathbf{B} - \mathbf{B}) \\ &= \mathbf{B}.\end{aligned}$$

[3] Both potentials are in the Coulomb gauge, so the divergence of their difference, $\nabla\cdot(\nabla\psi) = \nabla^2\psi$, must be zero. Conversely, the gradient of any function that satisfies Laplace's equation can be added to $\mathbf{A}$ without changing its divergence.

12.1.2 Circulation of A and magnetic flux

Equation (12.1) has an integral form, just as Gauss's law and Ampère's law do. Integrate it over any surface $\mathscr{S}$ bounded by a closed loop, and use the curl theorem, (1.37), on the integral of $\nabla \times \mathbf{A}$:

$$\oint \mathbf{A} \cdot d\boldsymbol{\ell} = \int_{\mathscr{S}} \mathbf{B} \cdot d\mathbf{a} = \Phi_B. \tag{12.4}$$

With words, the circulation of the vector potential around any closed loop equals the magnetic flux through the loop, (11.2). (Since the flux is the same for every surface the loop bounds, as Section 11.1.1 showed, it does not matter which surface is used.) The circulation of a gradient field around a closed loop is zero, by the gradient theorem, so a gauge transformation leaves the left side unchanged, as it must.

Compare (12.4) with Ampère's law in its integral form, (11.5):

$$\oint \mathbf{B} \cdot d\boldsymbol{\ell} = \mu_0 I_{\text{enc}} = \int \mu_0 \mathbf{J} \cdot d\mathbf{a}.$$

The vector potential is related to the magnetic field in the same way that the magnetic field is related to $\mu_0\mathbf{J}$ (Figure 12.4): the curl of $\mathbf{A}$ is $\mathbf{B}$, and the curl of $\mathbf{B}$ is $\mu_0\mathbf{J}$; in the Coulomb gauge the divergence of each is zero. That parallel will give us a method for finding $\mathbf{A}$ from $\mathbf{B}$ (Section 12.4), and it gives the boundary conditions next.

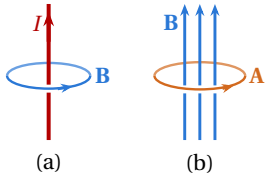


Figure 12.4 (a) The magnetic field circles a current, and its circulation is $\mu_0 I$. (b) The vector potential circles a bundle of magnetic flux in the same way, and its circulation is the flux, (12.4).

12.1.3 Continuity of the vector potential

Section 11.3 found how $\mathbf{B}$ behaves at a surface current: its perpendicular component is continuous, and its parallel components jump, as summarized in (11.11). The vector potential behaves more simply, as the potential V does at a surface charge.

For the parallel components, use a thin rectangular loop like the one in Figure 11.17b: two sides of length ℓ parallel to the surface, one just above it and one just below, joined by two short sides of length δ. As $\delta \to 0$, the short sides contribute nothing to $\oint \mathbf{A} \cdot d\boldsymbol{\ell}$, as long as $\mathbf{A}$ is finite, and the flux through the loop goes to zero too, since $\mathbf{B}$ is finite and the area $\ell\delta$ shrinks to zero. So (12.4) gives $(A^{\parallel}_{\text{above}} - A^{\parallel}_{\text{below}})\,\ell = 0$: the parallel components of $\mathbf{A}$ are continuous, in any gauge. For the perpendicular component, use a flat pillbox straddling the surface. In the Coulomb gauge, $\oint \mathbf{A} \cdot d\mathbf{a} = \int \nabla \cdot \mathbf{A}\, d\tau = 0$, and as the pillbox gets thinner only its top and bottom contribute, so $A^{\perp}_{\text{above}} = A^{\perp}_{\text{below}}$. In the Coulomb gauge, then, all of $\mathbf{A}$ is continuous:

$$\mathbf{A}_{\text{above}} = \mathbf{A}_{\text{below}}. \qquad \text{(Coulomb gauge)} \tag{12.5}$$

In other gauges the perpendicular component can jump (Exercise P12.4).

What jumps is the slope, as the slope of V does at a surface charge. Put the surface in the xy plane, with the unit normal $\hat{\mathbf{n}} = \hat{\mathbf{z}}$ pointing toward the side called

"above." Because **A** is continuous across the surface, its derivatives along the surface, with respect to x and y, are continuous too. So the only parts of $\nabla\times\mathbf{A}$ that can jump are the ones containing z derivatives,

$$-\frac{\partial A_y}{\partial z}\hat{\mathbf{x}}+\frac{\partial A_x}{\partial z}\hat{\mathbf{y}}=\hat{\mathbf{z}}\times\frac{\partial\mathbf{A}}{\partial z},$$

and by (11.11) the jump in this quantity must be $\mu_0\mathbf{K}\times\hat{\mathbf{z}}$, where **K** lies in the xy plane. In the Coulomb gauge, $\partial A_z/\partial z=-\partial A_x/\partial x-\partial A_y/\partial y$ is also continuous. Writing the jump in $\partial\mathbf{A}/\partial z$ as Δ, these say that $\hat{\mathbf{z}}\times\Delta=\mu_0\mathbf{K}\times\hat{\mathbf{z}}=\hat{\mathbf{z}}\times(-\mu_0\mathbf{K})$ and that Δ has no z component. Together, $\Delta=-\mu_0\mathbf{K}$:

$$\frac{\partial\mathbf{A}_{\text{above}}}{\partial n}-\frac{\partial\mathbf{A}_{\text{below}}}{\partial n}=-\mu_0\mathbf{K}, \qquad \text{(Coulomb gauge)} \quad (12.6)$$

where $\partial/\partial n$ is the derivative along $\hat{\mathbf{n}}$. Compare the jump in the slope of the potential at a surface charge, (7.2): the slope of V jumps by $-\sigma/\varepsilon_0$, and the slope of each Cartesian component of **A** jumps by $-\mu_0$ times the same component of **K**.

As a check, take the current sheet of Example 11.2, with $\mathbf{K}=K\hat{\mathbf{x}}$ in the xy plane and the field (11.6). The vector potential

$$\mathbf{A}=-\frac{\mu_0 K}{2}|z|\hat{\mathbf{x}}$$

has zero divergence, and its curl is $(\partial A_x/\partial z)\hat{\mathbf{y}}$, which is $-\frac{1}{2}\mu_0 K\hat{\mathbf{y}}$ above the sheet and $+\frac{1}{2}\mu_0 K\hat{\mathbf{y}}$ below it, as (11.6) requires. It is continuous at $z=0$, and its slope $\partial A_x/\partial z$ changes from $+\frac{1}{2}\mu_0 K$ below to $-\frac{1}{2}\mu_0 K$ above, a jump of $-\mu_0 K$, as (12.6) requires. It looks just like the potential of a charged plane, $-(\sigma/2\varepsilon_0)|z|$, with σ/ε_0 replaced by $\mu_0 K$ (Figure 12.5). Like that potential, it cannot be made to vanish at infinity, because the sheet is infinite; this one is chosen to be zero on the sheet itself.

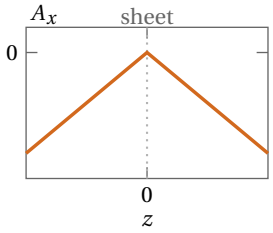


Figure 12.5 The vector potential of the current sheet, $A_x=-\frac{1}{2}\mu_0 K|z|$: continuous at the sheet, with slope $+\frac{1}{2}\mu_0 K$ below it and $-\frac{1}{2}\mu_0 K$ above, a jump of $-\mu_0 K$.

12.2 Poisson's equation for A

In Section 5.10, combining $\mathbf{E}=-\nabla V$ with Gauss's law gave **Poisson's equation**. Do the same with $\mathbf{B}=\nabla\times\mathbf{A}$ and Ampère's law. By the curl of a curl, (1.58),

$$\begin{aligned}\nabla\times\mathbf{B}&=\nabla\times(\nabla\times\mathbf{A})\\&=\nabla(\nabla\cdot\mathbf{A})-\nabla^2\mathbf{A}\\&=\mu_0\mathbf{J}.\end{aligned}$$

That looks worse than what we started with, until we remember the gauge. In the Coulomb gauge the first term is zero, and what is left is

$$\nabla^2\mathbf{A}=-\mu_0\mathbf{J}. \qquad \text{(Poisson's equation for A)} \quad (12.7)$$

This is the main reason the Coulomb gauge is the usual choice in magnetostatics: it turns Ampère's law into Poisson's equation.

What does the Laplacian of a vector mean? In Cartesian coordinates the unit vectors are constants, so $\nabla^2\mathbf{A} = (\nabla^2 A_x)\,\hat{\mathbf{x}} + (\nabla^2 A_y)\,\hat{\mathbf{y}} + (\nabla^2 A_z)\,\hat{\mathbf{z}}$, and (12.7) is three equations,

$$\nabla^2 A_x = -\mu_0 J_x, \qquad \nabla^2 A_y = -\mu_0 J_y, \qquad \nabla^2 A_z = -\mu_0 J_z. \tag{12.8}$$

In cylindrical or spherical coordinates, $\nabla^2\mathbf{A}$ is *not* the Laplacian of each component times its unit vector, because those unit vectors change from place to place. For example, $\nabla^2\hat{\boldsymbol{\phi}} = -\hat{\boldsymbol{\phi}}/s^2$, not zero (Exercise P12.5). It is the same trap that Colton's First Commandment warns about, in a new place. The safe definition in any coordinates is the curl-of-a-curl identity read backward, $\nabla^2\mathbf{A} = \nabla(\nabla\cdot\mathbf{A}) - \nabla\times(\nabla\times\mathbf{A})$.

12.2.1 The components of A behave like V

Each equation in (12.8) is Poisson's equation (5.26), with ρ/ε_0 replaced by $\mu_0 J_x$, $\mu_0 J_y$, or $\mu_0 J_z$. Picture the x component of the current density, $J_x(\mathbf{r})$, as a scalar function spread through space, positive in some places and negative in others (Figure 12.6). It produces A_x in exactly the way that a charge density produces a potential; J_y produces A_y, and J_z produces A_z, the same way. The mathematical problems are identical, so every result we have for V gives a result for each Cartesian component of $\mathbf{A}$, as Table 12.1 shows.

Electrostatics	Magnetostatics, each component
$\nabla^2 V = -\rho/\varepsilon_0$	$\nabla^2 A_x = -\mu_0 J_x$
$V = \dfrac{1}{4\pi\varepsilon_0}\displaystyle\int \frac{\rho}{\mathscr{r}}\,d\tau'$	$A_x = \dfrac{\mu_0}{4\pi}\displaystyle\int \frac{J_x}{\mathscr{r}}\,d\tau'$
continuous; slope jumps by $-\sigma/\varepsilon_0$	continuous; slope jumps by $-\mu_0 K_x$
line charge: $V = -\dfrac{\lambda}{2\pi\varepsilon_0}\ln\dfrac{s}{s_0}$	wire along z: $A_z = -\dfrac{\mu_0 I}{2\pi}\ln\dfrac{s}{s_0}$
relaxation: $V_{i,j} = \frac{1}{4}(\ldots) + \rho h^2/4\varepsilon_0$	$A_{x\,i,j} = \frac{1}{4}(\ldots) + \mu_0 J_x h^2/4$

Table 12.1 Results for the potential V and the corresponding results for each Cartesian component of $\mathbf{A}$.

In every row the substitution is the same: $\rho/\varepsilon_0 \to \mu_0 J_x$, and likewise $\sigma/\varepsilon_0 \to \mu_0 K_x$ and $\lambda/\varepsilon_0 \to \mu_0 I$. This correspondence is a harder idea than it looks, and it catches many people off guard. The vector potential is not much more complicated than the scalar potential: it just has components, and each of them behaves like a scalar potential.

Three cautions go with it. First, it holds for Cartesian components only, for the reason given above (Figure 12.7). Second, (12.7) holds in the Coulomb gauge. Third, the integral in the table assumes that the currents are confined to a finite region, so that $\mathbf{A}$ goes to zero far away, just as the corresponding integral for V assumes a finite distribution of charge.

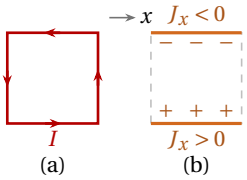


Figure 12.6 (a) A square loop of current. (b) The x component of its current alone: positive along the bottom side, negative along the top, and zero on the other two. It produces A_x as a pair of opposite line charges would produce V.

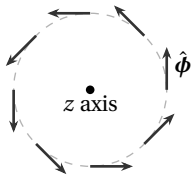


Figure 12.7 The unit vector $\hat{\boldsymbol{\phi}}$ at eight points around the z axis, seen from above: the same length everywhere, but a different direction at each point. That is why $\nabla^2(A_\phi\hat{\boldsymbol{\phi}})$ is not $(\nabla^2 A_\phi)\,\hat{\boldsymbol{\phi}}$.

12.2.2 The script-r integral for A

Read the other way, (12.7) is a differential equation for **A**, and the correspondence gives its solution. Section 5.10 showed that the script-r integral (5.16) is the solution of Poisson's equation that goes to zero far away; applying it to each component and putting the components back together,

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi}\int \frac{\mathbf{J}(\mathbf{r}')}{\mathscr{r}}\,d\tau'. \tag{12.9}$$

This is the vector **W** of the Helmholtz theorem, found in Section 12.1. There is a gap in the argument as it stands, though. We used the Coulomb gauge to get (12.7), so we should check that (12.9) really has zero divergence, and also that its curl really is the field of the Biot–Savart law.

Checking the script-r integral for A

Both checks rest on the same observation as in Section 11.1: ∇ acts on the field point **r**, so inside the integral it acts only on $1/\mathscr{r}$, and $\nabla(1/\mathscr{r}) = -\hat{\boldsymbol{\mathscr{r}}}/\mathscr{r}^2$.

The curl. By the product rule $\nabla\times(f\mathbf{C}) = f\nabla\times\mathbf{C} - \mathbf{C}\times\nabla f$ (Section 1.7), with $\mathbf{C} = \mathbf{J}(\mathbf{r}')$, which does not depend on **r**,

$$\nabla\times\frac{\mathbf{J}(\mathbf{r}')}{\mathscr{r}} = -\mathbf{J}(\mathbf{r}')\times\nabla\frac{1}{\mathscr{r}} = \frac{\mathbf{J}(\mathbf{r}')\times\hat{\boldsymbol{\mathscr{r}}}}{\mathscr{r}^2},$$

so the curl of (12.9) is $(\mu_0/4\pi)\int \mathbf{J}\times\hat{\boldsymbol{\mathscr{r}}}/\mathscr{r}^2\,d\tau'$, which is the Biot–Savart law (10.18).

The divergence. Since $1/\mathscr{r}$ depends on $\mathbf{r}-\mathbf{r}'$, its gradient with respect to the source point is the negative of its gradient with respect to the field point: $\nabla(1/\mathscr{r}) = -\nabla'(1/\mathscr{r})$. So

$$\begin{aligned}\nabla\cdot\frac{\mathbf{J}(\mathbf{r}')}{\mathscr{r}} &= \mathbf{J}(\mathbf{r}')\cdot\nabla\frac{1}{\mathscr{r}}\\ &= -\mathbf{J}(\mathbf{r}')\cdot\nabla'\frac{1}{\mathscr{r}}\\ &= -\nabla'\cdot\frac{\mathbf{J}(\mathbf{r}')}{\mathscr{r}} + \frac{\nabla'\cdot\mathbf{J}(\mathbf{r}')}{\mathscr{r}},\end{aligned}$$

using (1.52) in the last step. Integrate over a volume that contains all of the current. By the divergence theorem the first term becomes $-\oint \mathbf{J}/\mathscr{r}\cdot d\mathbf{a}'$ over the surface of that volume, where there is no current, so it is zero. The second term is zero because steady currents have $\nabla'\cdot\mathbf{J} = 0$, (10.13). So $\nabla\cdot\mathbf{A} = 0$.

Both checks need steady currents confined to a finite region. And as in Section 5.10, the Laplacian of (12.9) is $(\mu_0/4\pi)\int \mathbf{J}(\mathbf{r}')\nabla^2(1/\mathscr{r})\,d\tau' = -\mu_0\mathbf{J}$, by (1.46).

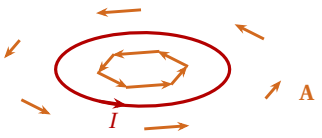


Figure 12.8 The vector potential (12.9) of a circular loop, at points in the plane of the loop inside and outside it: it circles in the same sense as the current. (The lengths are computed from (12.14) of Example 12.4.)

Equation (12.9) also shows which way **A** points (Figure 12.8). It adds up the current-density vectors $\mathbf{J}(\mathbf{r}')$, each weighted by the positive number $1/\mathscr{r}$. So the vector potential at a point is a weighted sum of the directions of the currents, with the nearby currents counting the most: near a straight wire **A** runs parallel to the

wire, and around a loop of current it circles in the same sense as the current. This is a statement about the particular potential (12.9), in the Coulomb gauge and going to zero far away. With another gauge, or another reference, it can fail; the potential of the current sheet above, which was chosen to vanish on the sheet, points opposite to **K**.

12.3 Solving for A when given J: script-r method

For a thin wire carrying a current I, or a surface current, $\mathbf{J}\,d\tau'$ in (12.9) is replaced by $I\,d\boldsymbol{\ell}'$ or $\mathbf{K}\,da'$, as in the Biot–Savart law:

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0 I}{4\pi}\int \frac{d\boldsymbol{\ell}'}{\mathscr{r}}, \tag{12.10}$$

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi}\int \frac{\mathbf{K}(\mathbf{r}')}{\mathscr{r}}\,da'. \tag{12.11}$$

Together with (12.9), these are the blueprints for the vector potential. Compare them with the Biot–Savart law, (10.16)–(10.18). In place of $\hat{\boldsymbol{\mathscr{r}}}/\mathscr{r}^2$ there is $1/\mathscr{r}$, and there is no cross product, so the integrals are usually easier. But they are still vector integrals. A piece of current whose direction changes from place to place, such as $d\boldsymbol{\ell}' = a\,d\phi'\,\hat{\boldsymbol{\phi}}'$ around a loop, must be written in Cartesian components before it is integrated: Colton's First Commandment applies here too. And compared with the integrals for V, they are the same integral done once for each component.

The steps follow those of Section 5.5:

1. Draw a picture, mark a piece of the current at an arbitrary location, and draw $\mathbf{r}$, $\mathbf{r}'$, and $\boldsymbol{\mathscr{r}}$ (Figure 12.9).
2. Write down $\mathbf{r}$, $\mathbf{r}'$, $\boldsymbol{\mathscr{r}} = \mathbf{r} - \mathbf{r}'$, and $\mathscr{r}$, each on a line of its own, and the element $d\boldsymbol{\ell}'$, da', or $d\tau'$. Only the distance $\mathscr{r}$ goes into the integral, but writing the vector $\boldsymbol{\mathscr{r}}$ first is the surest way to get it right. The element is the size of one piece of the current, and $I\,d\boldsymbol{\ell}'$, $\mathbf{K}\,da'$, or $\mathbf{J}\,d\tau'$ is the moving charge in that piece times its velocity, as in the Biot–Savart law.
3. Write the current element, $I\,d\boldsymbol{\ell}'$, **K**, or **J**, in Cartesian components, in terms of the primed variables. Then set the limits: integrate over the source, counting each piece of it once.
4. Do the integrals, one for each Cartesian component. Symmetry often shows that some of them are zero.
5. Check the result: its direction (roughly along the nearby currents), its units ($\mu_0 I$ times a pure number is a vector potential, in T·m), and its limits. If the field is wanted, take the curl.

The same limitation applies as for V. These integrals put **A** to zero far away, and for an infinite distribution of current, such as an infinitely long wire, they give an infinite result, so **A** has to be found some other way.

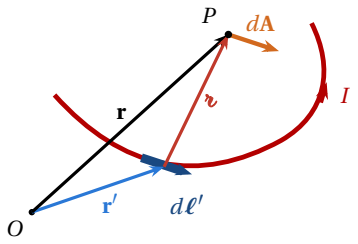


Figure 12.9 The script-r method for the vector potential. A short piece $d\boldsymbol{\ell}'$ of a wire carrying a current I, at the source point $\mathbf{r}'$, contributes $d\mathbf{A} = \mu_0 I\,d\boldsymbol{\ell}'/4\pi\mathscr{r}$ at the field point P. The contribution is parallel to the piece, and only the distance $\mathscr{r}$ enters.

12.3.1 Worked examples

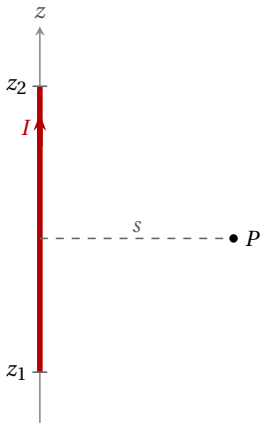


Figure 12.10 The straight segment of Example 12.2, from z_1 to z_2 on the z axis, and the field point P, a distance s from the axis.

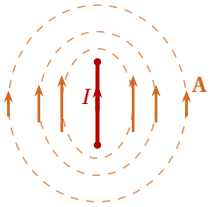
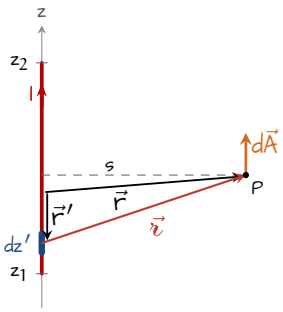


Figure 12.11 The vector potential (12.12) of the segment points along it everywhere. Its magnitude depends only on the sum of the distances from the two ends, so it is the same all around each ellipse (dashed) whose foci are the ends.

Example 12.2

A straight wire runs along the z axis from z_1 to z_2 and carries a current I in the $+z$ direction. Find the vector potential it produces at a point P a distance s from the axis, at height z. (A segment like this cannot carry a steady current by itself; think of it as one piece of a closed circuit.)

Solution:



We copy the segment of Figure 12.10 and, as the steps above say, mark a piece of the current at an arbitrary location, with the vectors $\mathbf{r}$ (to P), $\mathbf{r}'$ (to the piece) and $\boldsymbol{\ell}$ (from the piece to P).

In cylindrical coordinates centered on the wire,

$$\begin{aligned}\mathbf{r} &= s\hat{\mathbf{s}} + z\hat{\mathbf{z}}\\ \mathbf{r}' &= z'\hat{\mathbf{z}}\\ \boldsymbol{\ell} &= \mathbf{r} - \mathbf{r}' = s\hat{\mathbf{s}} + (z - z')\hat{\mathbf{z}}\\ \ell &= \sqrt{s^2 + (z - z')^2}.\end{aligned}$$

The piece of wire is $d\boldsymbol{\ell}' = dz'\,\hat{\mathbf{z}}$, whose direction does not change along the wire, so $\hat{\mathbf{z}}$ can come straight out of the integral. By (12.10),

$$\mathbf{A} = \frac{\mu_0 I}{4\pi}\hat{\mathbf{z}}\int_{z_1}^{z_2}\frac{dz'}{\sqrt{s^2 + (z' - z)^2}}.$$

With $u = z' - z$ and the standard integral $\int du/\sqrt{s^2 + u^2} = \ln\left(u + \sqrt{s^2 + u^2}\right)$,

$$\mathbf{A} = \frac{\mu_0 I}{4\pi}\ln\left[\frac{z_2 - z + \sqrt{s^2 + (z_2 - z)^2}}{z_1 - z + \sqrt{s^2 + (z_1 - z)^2}}\right]\hat{\mathbf{z}}. \tag{12.12}$$

The vector potential points along the wire (Figure 12.11).

Reality check: Take the curl. For $\mathbf{A} = A_z(s, z)\,\hat{\mathbf{z}}$, the curl in cylindrical coordinates (Section 1.4.6) is

$$\begin{aligned}\nabla \times \mathbf{A} &= \left(\frac{1}{s}\frac{\partial A_z}{\partial \phi} - \frac{\partial A_\phi}{\partial z}\right)\hat{\mathbf{s}} + \left(\frac{\partial A_s}{\partial z} - \frac{\partial A_z}{\partial s}\right)\hat{\boldsymbol{\phi}} + \frac{1}{s}\left(\frac{\partial (sA_\phi)}{\partial s} - \frac{\partial A_s}{\partial \phi}\right)\hat{\mathbf{z}}\\ &= -\frac{\partial A_z}{\partial s}\,\hat{\boldsymbol{\phi}}.\end{aligned}$$

Multiplying the top and bottom of the derivative by $\sqrt{s^2+u^2}-u$ gives

$$\frac{\partial}{\partial s}\ln\left(u+\sqrt{s^2+u^2}\right)=\frac{s}{\sqrt{s^2+u^2}\left(u+\sqrt{s^2+u^2}\right)}=\frac{1}{s}-\frac{u}{s\sqrt{s^2+u^2}},$$

so

$$\mathbf{B}=-\frac{\partial A_z}{\partial s}\hat{\boldsymbol{\phi}}=\frac{\mu_0 I}{4\pi s}\left[\frac{z_2-z}{\sqrt{s^2+(z_2-z)^2}}-\frac{z_1-z}{\sqrt{s^2+(z_1-z)^2}}\right]\hat{\boldsymbol{\phi}}.$$

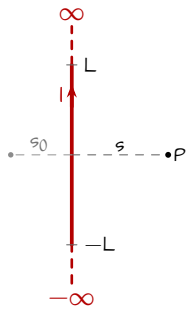
The two fractions are $\sin\theta_2$ and $\sin\theta_1$ of Example 10.6, so this is (10.21). Far away, at a distance r much greater than the length $L=z_2-z_1$, every piece of the wire is about the same distance away, and $\mathbf{A}\approx\mu_0 IL\hat{\mathbf{z}}/4\pi r$: the potential of a point charge, with $\mu_0 IL$ in place of q/ε_0.

Notice one more property. The divergence of (12.12) is $\partial A_z/\partial z=(\mu_0 I/4\pi)(1/\mathscr{r}_1-1/\mathscr{r}_2)$, where $\mathscr{r}_1$ and $\mathscr{r}_2$ are the distances from P to the two ends, so it is not zero. That is as it should be, since a lone segment is not a steady current: charge would pile up at its ends, and the check in Section 12.2.2 needed $\nabla\cdot\mathbf{J}=0$. For a closed circuit made of straight pieces, the end contributions cancel in pairs at every corner, and the sum of the pieces' potentials is in the Coulomb gauge (Exercise P12.7).

Example 12.3

Find the vector potential of an infinitely long straight wire on the z axis that carries a current I in the $+z$ direction.

Solution:



Our sketch shows the infinite wire and, on it, the long but finite segment from $-L$ to L that we will use in its place, with P opposite the middle of the segment, and a reference distance s_0. The obvious approach is to let $z_1\to-\infty$ and $z_2\to\infty$ in (12.12). It fails: the integral $\int dz'/\sqrt{s^2+z'^2}$ diverges at both ends, and $\mathbf{A}$ comes out infinite everywhere. The same thing happened to the potential of an infinite line of charge (Example 5.1), for the same reason: the source extends to infinity.

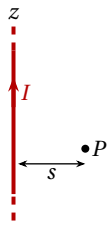


Figure 12.12 The wire of Example 12.3, and a point P at a distance s from it.

As there, the reference has to be changed. Take a long but finite segment, from $-L$ to L, and put P at $z=0$, opposite its middle. Then (12.12) gives, after the top and bottom of the fraction are multiplied by $L+\sqrt{s^2+L^2}$,

$$\begin{aligned}A_z&=\frac{\mu_0 I}{4\pi}\ln\left[\frac{L+\sqrt{s^2+L^2}}{-L+\sqrt{s^2+L^2}}\right]\\&=\frac{\mu_0 I}{4\pi}\ln\left[\frac{\left(L+\sqrt{s^2+L^2}\right)^2}{s^2}\right]\approx\frac{\mu_0 I}{2\pi}\ln\frac{2L}{s}\end{aligned}$$

for $L \gg s$. As $L \to \infty$ this grows without limit, but the part that grows, $(\mu_0 I/2\pi)\ln 2L$, is the same at every distance s. Differences stay finite:

$$A_z(s) - A_z(s_0) = \frac{\mu_0 I}{2\pi} \ln \frac{s_0}{s}.$$

Adding a constant vector to $\mathbf{A}$ changes nothing (it is a gauge transformation, (12.2), with ψ proportional to z), so we may subtract the value at some chosen distance s_0, and the limit $L \to \infty$ is then harmless:

$$\mathbf{A} = -\frac{\mu_0 I}{2\pi} \ln \frac{s}{s_0} \hat{\mathbf{z}}. \tag{12.13}$$

It is zero at the reference distance s_0. It is also the potential of an infinite line of charge, $-(\lambda/2\pi\varepsilon_0)\ln(s/s_0)$, with λ/ε_0 replaced by $\mu_0 I$, as Section 12.2.1 says it must be.

Reality check: $-\partial A_z/\partial s = \mu_0 I/2\pi s$, so the curl is (10.20), and $\partial A_z/\partial z = 0$, so this is the Coulomb gauge. Section 12.4 finds the same result directly from the field.

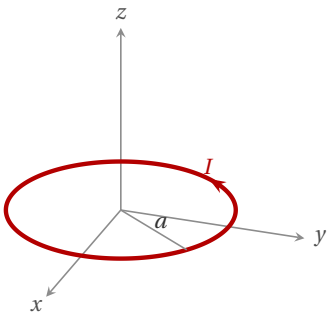
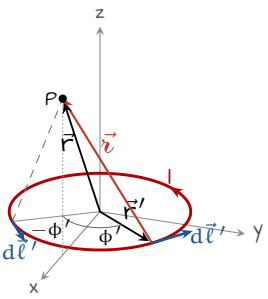


Figure 12.13 The loop of Example 12.4: radius a, in the xy plane, centered on the origin, carrying a current I counterclockwise as seen from $+z$.

Example 12.4

A circular loop of radius a lies in the xy plane, centered on the origin, and carries a current I counterclockwise as seen from $+z$. Find the vector potential at an arbitrary point.

Solution:



We copy the loop of Figure 12.13, put the field point in a convenient place, and mark two pieces of the loop placed symmetrically about it, with the vectors $\mathbf{r}$, $\mathbf{r}'$ and $\boldsymbol{\imath}$ for one of them. The loop is symmetric about the z axis, so nothing is lost by putting the field point in the xz plane.

For the piece of the loop at angle ϕ',

$$\begin{aligned}
\mathbf{r} &= s\hat{\mathbf{x}} + z\hat{\mathbf{z}} \\
\mathbf{r}' &= a\hat{\mathbf{s}}' = a\left(\cos\phi'\,\hat{\mathbf{x}} + \sin\phi'\,\hat{\mathbf{y}}\right) \\
\boldsymbol{\imath} &= \mathbf{r} - \mathbf{r}' = \left(s - a\cos\phi'\right)\hat{\mathbf{x}} - a\sin\phi'\,\hat{\mathbf{y}} + z\hat{\mathbf{z}} \\
\imath &= \sqrt{\left(s - a\cos\phi'\right)^2 + a^2\sin^2\phi' + z^2} = \sqrt{s^2 + a^2 + z^2 - 2as\cos\phi'}.
\end{aligned}$$

The piece of the loop is $d\boldsymbol{\ell}' = a\,d\phi'\,\hat{\boldsymbol{\phi}}'$. Since $\hat{\boldsymbol{\phi}}'$ changes direction around the loop, Colton's First Commandment requires it in Cartesian components:

$$d\boldsymbol{\ell}' = a\,d\phi'\left(-\sin\phi'\,\hat{\mathbf{x}} + \cos\phi'\,\hat{\mathbf{y}}\right).$$

By (12.10),

$$\mathbf{A} = \frac{\mu_0 I a}{4\pi}\int_0^{2\pi} \frac{-\sin\phi'\,\hat{\mathbf{x}} + \cos\phi'\,\hat{\mathbf{y}}}{\sqrt{s^2+a^2+z^2-2as\cos\phi'}}\,d\phi'.$$

The $\hat{\mathbf{x}}$ integral is zero. (Why is the $\hat{\mathbf{x}}$ integral zero? Answer in footnote.[4]) The $\hat{\mathbf{y}}$ term survives, and at P, which is in the xz plane, $\hat{\mathbf{y}} = \hat{\boldsymbol{\phi}}$. So $\mathbf{A} = A_\phi(s,z)\,\hat{\boldsymbol{\phi}}$, with

$$A_\phi(s,z) = \frac{\mu_0 I a}{4\pi}\int_0^{2\pi} \frac{\cos\phi'\,d\phi'}{\sqrt{s^2+a^2+z^2-2as\cos\phi'}}. \tag{12.14}$$

The vector potential circles the axis, in the same sense as the current. The integral cannot be done in terms of elementary functions. It can be written with complete elliptic integrals,[5] or evaluated numerically (Figure 12.14).

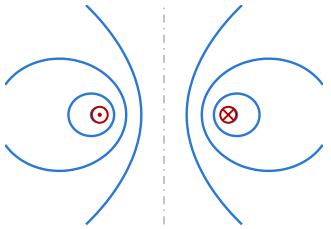


Figure 12.14 The vector potential of the loop of Example 12.4, computed from the elliptic integrals, in a plane through the axis: curves on which sA_ϕ is constant. They are also the field lines, because **B** is perpendicular to the gradient of sA_ϕ.

Reality check: Two limits test the result. Near the axis, $s \ll a$, expand the integrand to first order in s:

$$\frac{1}{\sqrt{a^2+z^2-2as\cos\phi'+s^2}} \approx \frac{1}{\sqrt{a^2+z^2}}\left(1+\frac{as\cos\phi'}{a^2+z^2}\right).$$

The 1 integrates to zero against $\cos\phi'$, and $\int_0^{2\pi}\cos^2\phi'\,d\phi' = \pi$, so

$$A_\phi \approx \frac{\mu_0 I a^2 s}{4\left(a^2+z^2\right)^{3/2}}.$$

For $\mathbf{A} = A_\phi\,\hat{\boldsymbol{\phi}}$ the curl in cylindrical coordinates is

$$\begin{aligned}\nabla\times\mathbf{A} &= \left(\frac{1}{\cancel{s}}\frac{\cancel{\partial A_z}}{\cancel{\partial\phi}} - \frac{\partial A_\phi}{\partial z}\right)\hat{\mathbf{s}} + \left(\frac{\cancel{\partial A_s}}{\cancel{\partial z}} - \frac{\cancel{\partial A_z}}{\cancel{\partial s}}\right)\hat{\boldsymbol{\phi}} + \frac{1}{s}\left(\frac{\partial(sA_\phi)}{\partial s} - \frac{\cancel{\partial A_s}}{\cancel{\partial\phi}}\right)\hat{\mathbf{z}} \\ &= -\frac{\partial A_\phi}{\partial z}\,\hat{\mathbf{s}} + \frac{1}{s}\frac{\partial(sA_\phi)}{\partial s}\,\hat{\mathbf{z}},\end{aligned}$$

and on the axis the second term gives

$$B_z = \frac{1}{s}\frac{\partial}{\partial s}\left(sA_\phi\right) = \frac{\mu_0 I a^2}{2\left(a^2+z^2\right)^{3/2}},$$

which is (10.22) (Figure 12.15).

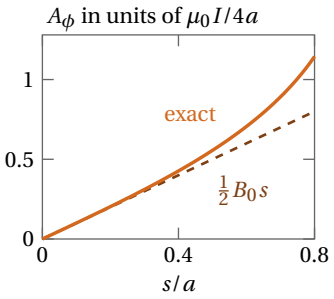


Figure 12.15 In the plane of the loop, A_ϕ from (12.14) (solid) and the near-axis approximation (dashed). The approximation is $\frac{1}{2}B_0 s$, where $B_0 = \mu_0 I/2a$ is the field at the center: near the axis the field is nearly uniform, and **A** is the symmetric potential $\mathbf{A}_1$ of Example 12.1.

Far away, $r = \sqrt{s^2+z^2} \gg a$, expand in a/r instead. Since $\mathscr{r}^2 = r^2 - 2as\cos\phi' + a^2$, to first order $1/\mathscr{r} \approx (1/r)(1 + as\cos\phi'/r^2)$, and with $s = r\sin\theta$,

$$A_\phi \approx \frac{\mu_0 I a}{4\pi}\frac{as}{r^3}\pi = \frac{\mu_0}{4\pi}\frac{(I\pi a^2)\sin\theta}{r^2}.$$

[4] Replacing ϕ' by $-\phi'$ changes the sign of $\sin\phi'$ but not the distance $\mathscr{r}$, so the pieces of the loop at ϕ' and $-\phi'$, placed symmetrically about the xz plane as in the sketch, cancel.

[5] $A_\phi = \frac{\mu_0 I}{\pi k}\sqrt{\frac{a}{s}}\left[\left(1-\frac{k^2}{2}\right)K(k^2) - E(k^2)\right]$, with $k^2 = 4as/\left[(a+s)^2+z^2\right]$, where K and E are the complete elliptic integrals of the first and second kinds, written as functions of the parameter $m = k^2$, as Mathematica's `EllipticK[m]` and `EllipticE[m]` and SciPy's `ellipk(m)` and `ellipe(m)` are (Section 5.6).

This is the vector potential of a magnetic dipole, with a dipole moment equal to the current times the area of the loop, $I\pi a^2$ (Section 12.6.2). Notice that in both expansions the leading term of $1/r$ integrated to zero. Far away, that means there is no term in $1/r$; Section 12.6.1 shows why there never is.

12.3.2 Tips for Mathematica and Python

The integral (12.14) is easy to evaluate numerically, with `NIntegrate` in Mathematica or `scipy.integrate.quad` in Python. When the integrals cannot be done by hand at all, the chunk-and-sum method of Section 10.7 works for **A** as it does for **B**, and more simply: divide the wire into short pieces $\Delta\boldsymbol{\ell}'_i$ and add up $\mathbf{A} \approx (\mu_0 I/4\pi)\sum_i \Delta\boldsymbol{\ell}'_i/r_i$, with no cross product to form. To take a curl in cylindrical coordinates, give Mathematica's `Curl` the coordinate system,

```
Curl[{0, Aphi[s, z], 0}, {s, phi, z}, "Cylindrical"],
```

with the components in the order s, ϕ, z. In SymPy, a coordinate system made with

```
CoordSys3D("C", transformation="cylindrical")
```

does the same; it names the coordinates r, θ, z.

12.4 Solving for A when given B: flux method (high symmetry)

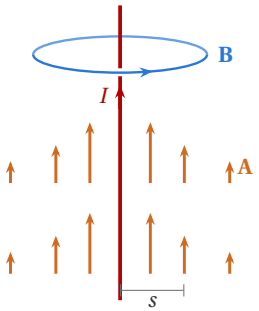


Figure 12.16 The vector potential beside a long straight wire, $\mathbf{A} = A_z(s)\,\hat{\mathbf{z}}$: parallel to the wire, the same at every z and on every side of the wire, and smaller farther out. (The arrows are drawn closer to the wire than the reference distance s_0, where $A_z = 0$.) Its curl is the field **B**, which circles the wire.

Often the field is already known, perhaps from Ampère's law, and the vector potential is wanted. The electrostatic counterpart is finding V from **E** with $V = -\int \mathbf{E}\cdot d\boldsymbol{\ell}$, (5.8). There is no equally simple formula for **A**, but when the symmetry is high enough, two routes work.

The first is to integrate $\mathbf{B} = \nabla\times\mathbf{A}$ directly. For the infinite wire the current is along $\hat{\mathbf{z}}$, so **A** is along $\hat{\mathbf{z}}$, and by symmetry it depends only on s: $\mathbf{A} = A_z(s)\,\hat{\mathbf{z}}$ (Figure 12.16). The curl of the vector potential in cylindrical coordinates, (1.27), is this, where we have crossed out all terms that involve components of **A** not in the z direction, and all terms that involve derivatives with respect to ϕ and z (which are zero):

$$\begin{aligned}\nabla\times\mathbf{A} &= \left(\frac{1}{s}\frac{\partial A_z}{\partial\phi} - \frac{\partial A_\phi}{\partial z}\right)\hat{\mathbf{s}} + \left(\frac{\partial A_s}{\partial z} - \frac{\partial A_z}{\partial s}\right)\hat{\boldsymbol{\phi}} \\ &\quad + \frac{1}{s}\left(\frac{\partial(sA_\phi)}{\partial s} - \frac{\partial A_s}{\partial\phi}\right)\hat{\mathbf{z}} \\ &= -\frac{dA_z}{ds}\,\hat{\boldsymbol{\phi}}.\end{aligned}$$

This must be the field of the wire, $(\mu_0 I/2\pi s)\,\hat{\boldsymbol{\phi}}$ from (10.20), so

$$\begin{aligned}
-\frac{dA_z}{ds} &= \frac{\mu_0 I}{2\pi s}, \\
dA_z &= -\frac{\mu_0 I}{2\pi s}\,ds, \\
A_z(s) &= \int_{s_0}^{s}\left(-\frac{\mu_0 I}{2\pi \bar{s}}\right)d\bar{s} \qquad \text{(where } s = s_0 \text{ is a reference point)} \\
&= -\frac{\mu_0 I}{2\pi}\ln\frac{s}{s_0},
\end{aligned}$$

which is (12.13) without the detour through a finite segment.

The second route uses the method of Section 11.2 for Ampère's law, with **A** in place of **B**. The two laws have the same form (Table 12.2).

Ampère's law	This law
$\nabla \times \mathbf{B} = \mu_0 \mathbf{J}$	$\nabla \times \mathbf{A} = \mathbf{B}$
$\oint \mathbf{B}\cdot d\boldsymbol{\ell} = \mu_0 I_{\text{enc}}$	$\oint \mathbf{A}\cdot d\boldsymbol{\ell} = ?$

Table 12.2 Ampère's law and the corresponding law for **A** (Figure 12.17).

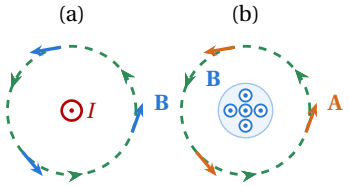


Figure 12.17 (a) An Amperian loop (dashed) around a current I that comes out of the page: **B** circles the current, and $\oint \mathbf{B}\cdot d\boldsymbol{\ell} = \mu_0 I$. (b) An "Amperian" loop around a bundle of magnetic flux that comes out of the page: **A** circles the flux in the same way, and $\oint \mathbf{A}\cdot d\boldsymbol{\ell} = \Phi_B$.

To figure out what goes on the right-hand side, by simple analogy we can treat **B** as a type of "current density," and so we must integrate it across the surface bounded by the loop of the line integral to get the "current enclosed." But the integral of **B** across a surface is just the magnetic flux, so

$$\oint \mathbf{A}\cdot d\boldsymbol{\ell} = \Phi_B,$$

which is (12.4), found in Section 12.1.2 with the curl theorem. We will call it the **flux method**. It needs what the Ampère's-law method needs: symmetry that tells us the direction of **A** and what it can depend on, and an "Amperian" loop along which **A** is either constant and parallel to the loop, or perpendicular to it. The steps:

1. From the symmetry and the directions of the currents, find the line along which **A** points and what it can depend on: around a solenoid it circles the axis, and beside straight currents it is parallel to them (Section 12.2.1). Its sign comes out of step 4, and beside straight currents it depends on the reference distance.

2. Draw a loop that follows **A** where **A** is constant, and crosses it at right angles elsewhere. Choose a direction around the loop. (The advice of Colton's Third Commandment applies here as much as to Ampère's law.)

3. Calculate the magnetic flux through the loop, with its sign given by the right-hand rule.

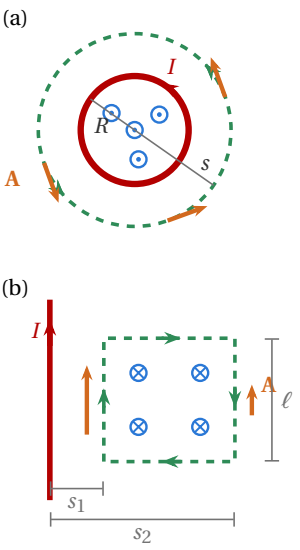


Figure 12.18 Loops for the flux method. (a) When **A** circles an axis, as around a solenoid (end view), use a circle of radius s around the axis. (b) When **A** is parallel to an axis, as beside a straight wire (side view), use a rectangle in a plane containing the axis. Going around it in the direction shown, up the side at s_1 and down the side at s_2, the right-hand rule makes the rectangle's normal point into the page, along **B**.

4. Set $\oint \mathbf{A}\cdot d\boldsymbol{\ell} = \Phi_B$, and solve for **A**.
5. Check that $\nabla\times\mathbf{A} = \mathbf{B}$, that $\nabla\cdot\mathbf{A} = 0$ if the Coulomb gauge is wanted, and that **A** is continuous.

Two loop shapes cover the usual cases, shown in Figure 12.18. When **A** circles an axis, $\mathbf{A} = A_\phi(s)\,\hat{\boldsymbol{\phi}}$, a circle of radius s centered on the axis gives $\oint \mathbf{A}\cdot d\boldsymbol{\ell} = 2\pi s A_\phi$. When **A** is parallel to an axis, $\mathbf{A} = A_z(s)\,\hat{\mathbf{z}}$, use a rectangle in a plane containing the axis, with two sides of length ℓ parallel to the axis at distances $s_1 < s_2$. Going up the inner side and down the outer one, $\oint \mathbf{A}\cdot d\boldsymbol{\ell} = [A_z(s_1) - A_z(s_2)]\,\ell$; the other two sides contribute nothing, because **A** is perpendicular to them. This loop gives the difference between **A** at two distances, so a reference distance enters on its own, as it does for V. For the wire, with the inner side at s_0 and the outer side at s, the flux through the rectangle is

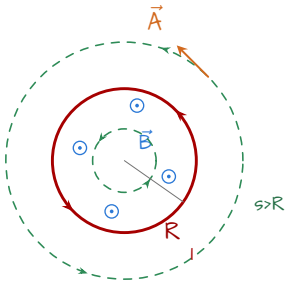
$$\Phi_B = \int_{s_0}^{s} \frac{\mu_0 I}{2\pi\bar{s}}\,\ell\,d\bar{s} = \frac{\mu_0 I\ell}{2\pi}\ln\frac{s}{s_0},$$

positive because **B** points along the rectangle's normal (Figure 12.18b). So $A_z(s_0) - A_z(s) = (\mu_0 I/2\pi)\ln(s/s_0)$, which is (12.13) once more.

Example 12.5

A long solenoid of radius R has n turns per unit length and carries a current I, so the field is $\mu_0 nI\hat{\mathbf{z}}$ inside and zero outside, (11.7). Find the vector potential inside and outside.

Solution:



As always, the first step is a sketch of what is given and what is asked: the solenoid end-on, with its current going around counterclockwise, and **B** out of the page inside it and zero outside. The currents circle the axis, so **A** does too, $\mathbf{A} = A_\phi(s)\,\hat{\boldsymbol{\phi}}$, and the loops are circles of radius s centered on the axis, one inside and one outside, traversed in the same sense as the current.

The right-hand rule makes the normal of each circle $\hat{\mathbf{z}}$, and $\oint \mathbf{A}\cdot d\boldsymbol{\ell} = 2\pi s A_\phi$.

Inside ($s < R$), the flux through the circle is $\mu_0 nI\pi s^2$, so $A_\phi = \frac{1}{2}\mu_0 nIs$. Outside ($s > R$), the circle encloses all of the flux, $\mu_0 nI\pi R^2$, however large it is, so $A_\phi = \mu_0 nIR^2/2s$. Together,

$$\mathbf{A} = \begin{cases} \dfrac{\mu_0 nI}{2}\,s\,\hat{\boldsymbol{\phi}}, & s < R, \\[2ex] \dfrac{\mu_0 nIR^2}{2s}\,\hat{\boldsymbol{\phi}}, & s > R. \end{cases} \tag{12.15}$$

Reality check: The curl of $A_\phi(s)\,\hat{\boldsymbol{\phi}}$ is $(1/s)\,d(sA_\phi)/ds\,\hat{\mathbf{z}}$, which is $\mu_0 nI$ inside and zero outside, as it should be. The divergence of $A_\phi(s)\,\hat{\boldsymbol{\phi}}$ is zero, so this is the Coulomb gauge. At $s = R$ both forms give $\frac{1}{2}\mu_0 nIR$, so $\mathbf{A}$ is continuous, as (12.5) requires. The slope dA_ϕ/ds is $\frac{1}{2}\mu_0 nI$ just inside and $-\frac{1}{2}\mu_0 nI$ just outside, a jump of $-\mu_0 nI$, the winding is a surface current $\mathbf{K} = nI\,\hat{\boldsymbol{\phi}}$, so that is (12.6). Inside, $\mathbf{A}$ is $\frac{1}{2}\mathbf{B}\times\mathbf{r}$, the symmetric potential $\mathbf{A}_1$ of Example 12.1, as it must be. (Why must the potential inside be $\mathbf{A}_1$, and not $\mathbf{A}_2$, which describes the same uniform field? Answer in footnote.[6])

The striking feature is outside: $\mathbf{A}$ is not zero there, although $\mathbf{B}$ is (Figure 12.19). It cannot be zero. Every loop around the solenoid has the flux $\mu_0 nI\pi R^2$ through it, so $\oint \mathbf{A}\cdot d\boldsymbol{\ell}$ around every such loop must be that large. A potential that was zero outside would have zero curl and zero divergence there, but it would violate (12.4), and it would jump at $s = R$. A potential depends on the field elsewhere, as the potential inside a charged spherical shell does (Example 5.2): it is not zero although the field there is. But that potential is a constant, and it becomes zero if the reference point is put inside the shell. No choice of gauge can make $\mathbf{A}$ zero outside the solenoid, because $\oint \mathbf{A}\cdot d\boldsymbol{\ell} = \Phi_B$ is the same in every gauge. (Outside, $\mathbf{A}$ is the gradient of $\Phi_B\phi/2\pi$, but that function is not single-valued: it increases by Φ_B on each trip around the axis, so subtracting its gradient is not a gauge transformation.)[7]

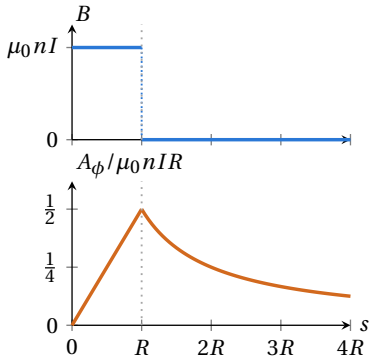


Figure 12.19 The field (top) and the vector potential (bottom) of a long solenoid of radius R, from Example 12.5. The field jumps at the winding, by $\mu_0 nI$. The vector potential is continuous there, but its slope changes abruptly, and it is not zero outside, where the field is zero.

The solenoid shows the parallel between $\mathbf{A}$ and $\mathbf{B}$ at work. Compare (12.15) with the field of a thick wire carrying a uniform current density J along its axis, which Ampère's law gives as $\frac{1}{2}\mu_0 Js$ inside and $\mu_0 JR^2/2s$ outside, circling the axis. With $\mu_0 J$ replaced by $B = \mu_0 nI$, the two are the same: the vector potential of a solenoid has the form of the magnetic field of a thick wire (Figure 12.20). In general, the Coulomb-gauge vector potential that goes to zero far away can be found from $\mathbf{B}$ by the Biot–Savart law, with $\mu_0\mathbf{J}$ replaced by $\mathbf{B}$, as long as $\mathbf{B}$ goes to zero fast enough far away (Exercise P12.12).

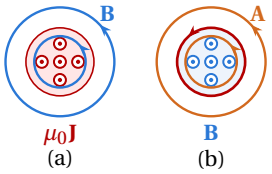


Figure 12.20 (a) A thick wire end on: $\mathbf{B}$ circles the current density $\mu_0\mathbf{J}$. (b) A long solenoid end on: $\mathbf{A}$ circles the field $\mathbf{B}$ inside it in the same way, growing in proportion to s inside and falling as $1/s$ outside.

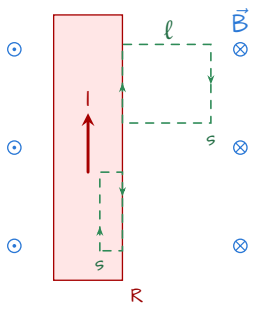
[6] The currents of the solenoid circle its axis, so its vector potential circles the axis too, $\mathbf{A} = A_\phi(s)\,\hat{\boldsymbol{\phi}}$, as step 1 of the method says. $\mathbf{A}_1 = \frac{1}{2}B_0 s\,\hat{\boldsymbol{\phi}}$ has that form; $\mathbf{A}_2 = B_0 x\,\hat{\mathbf{y}}$ does not, since it is parallel to the y axis everywhere.

[7] Is the vector potential outside a solenoid real, or only a device for calculating? In classical physics it has no effects, since forces depend only on $\mathbf{B}$. In quantum mechanics it does. Electrons passing on the two sides of a long solenoid, through regions where $\mathbf{B}$ is zero, acquire a relative phase $e\Phi_B/\hbar$, proportional to $\oint \mathbf{A}\cdot d\boldsymbol{\ell}$, which shifts their interference pattern. Werner Ehrenberg and Raymond Siday pointed this out in 1949: "The refractive index in electron optics and the principles of dynamics," W. Ehrenberg and R. E. Siday, Proc. Phys. Soc. B 62, 8–21 (1949). https://doi.org/10.1088/0370-1301/62/1/303. Yakir Aharonov and David Bohm found it independently in 1959: "Significance of electromagnetic potentials in the quantum theory," Y. Aharonov and D. Bohm, Phys. Rev. 115, 485–491 (1959). https://doi.org/10.1103/PhysRev.115.485. The most convincing test enclosed a tiny toroidal magnet in a superconductor, so that no field could leak out, and found the predicted shift: "Evidence for Aharonov–Bohm effect with magnetic field completely shielded from electron wave," A. Tonomura, N. Osakabe, T. Matsuda, T. Kawasaki, J. Endo, S. Yano, and H. Yamada, Phys. Rev. Lett. 56, 792–795 (1986). https://doi.org/10.1103/PhysRevLett.56.792.

Example 12.6

A long straight wire of radius R carries a current I in the $+z$ direction, spread uniformly over its cross section. Find the vector potential inside and outside the wire, taking it to be zero at the surface of the wire.

Solution:



We first draw the wire from the side, with its current going up the page. By Ampère's law (Example 11.1), the field circles the axis, with magnitude $\mu_0 I s/2\pi R^2$ inside and $\mu_0 I/2\pi s$ outside. The current is along $\hat{\mathbf{z}}$, so $\mathbf{A} = A_z(s)\,\hat{\mathbf{z}}$, and the loops are rectangles in a plane containing the axis, with sides of length ℓ: one for a point outside the wire and one for a point inside. Outside ($s > R$), put the inner side at the surface and the outer side at s.

The flux through the rectangle is the one found above for the thin wire, with $s_0 = R$. (Why is the flux the same as for a thin wire? Answer in footnote.[8]) So $A_z(R) - A_z(s) = (\mu_0 I/2\pi)\ln(s/R)$, and with $A_z(R) = 0$,

$$A_z = \frac{\mu_0 I}{2\pi}\ln\frac{R}{s}.$$

Inside ($s < R$), put the inner side at s and the outer side at the surface. The flux is

$$\Phi_B = \int_s^R \frac{\mu_0 I \bar{s}}{2\pi R^2}\,\ell\,d\bar{s} = \frac{\mu_0 I \ell}{4\pi R^2}\left(R^2 - s^2\right),$$

and $[A_z(s) - A_z(R)]\,\ell = \Phi_B$, so

$$A_z = \frac{\mu_0 I}{4\pi}\left(1 - \frac{s^2}{R^2}\right).$$

Reality check: Inside,

$$\begin{aligned}\nabla^2 A_z &= \frac{1}{s}\frac{\partial}{\partial s}\left(s\frac{\partial A_z}{\partial s}\right) + \cancel{\frac{1}{s^2}\frac{\partial^2 A_z}{\partial \phi^2}} + \cancel{\frac{\partial^2 A_z}{\partial z^2}} \\ &= -\frac{\mu_0 I}{\pi R^2} \\ &= -\mu_0 J,\end{aligned}$$

which is Poisson's equation (12.8); outside, $\nabla^2 A_z = 0$. At $s = R$ both forms are zero, and both slopes are $-\mu_0 I/2\pi R$: $\mathbf{A}$ and its slope are continuous, as they should be where there is no surface current. The result has the form of the potential of a uniformly charged cylinder (Exercise P5.2), with ρ/ε_0 replaced by $\mu_0 J$.

[8] Outside the wire the field has the magnitude $\mu_0 I/2\pi s$, the same as for a thin wire carrying the same current, because Ampère's law gives it from the enclosed current alone (Example 11.1).

12.5 Numerical method: relaxation for A in two dimensions

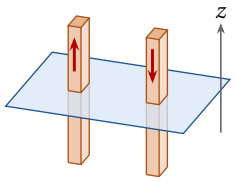


Figure 12.21 Two long conductors parallel to the z axis, carrying opposite currents, and one cross section (shaded). The current density $J_z(x,y)\,\hat{\mathbf{z}}$ is the same in every cross section, so the problem is two-dimensional.

Many practical conductors are long and straight: power cables, the bus bars that carry large currents in buildings and power stations, the traces on a circuit board, the windings of a long electromagnet. When all the conductors are parallel to the z axis, and long enough that their ends do not matter, the current density is $\mathbf{J} = J_z(x,y)\,\hat{\mathbf{z}}$ (Figure 12.21). Only J_z is nonzero, so by (12.8) only A_z is nonzero; it does not depend on z, so $\nabla\cdot\mathbf{A} = \partial A_z/\partial z = 0$ automatically; and

$$\frac{\partial^2 A_z}{\partial x^2} + \frac{\partial^2 A_z}{\partial y^2} = -\mu_0 J_z.$$

This is the two-dimensional Poisson equation, which the **relaxation method** of Section 7.3.1 handles with a source term, here with ρ/ε_0 replaced by $\mu_0 J_z$. On a square grid with spacing h, the relaxation rule (7.10), with the source term for Poisson's equation, becomes

$$A_{i,j} = \tfrac{1}{4}\left(A_{i+1,j} + A_{i-1,j} + A_{i,j+1} + A_{i,j-1}\right) + \frac{\mu_0 J_{i,j}\,h^2}{4}, \tag{12.16}$$

where $A_{i,j}$ and $J_{i,j}$ are A_z and J_z at the grid point (i,j). Once A_z is known on the grid, the field follows from $\mathbf{B} = \nabla\times(A_z\hat{\mathbf{z}})$,

$$\mathbf{B} = \frac{\partial A_z}{\partial y}\,\hat{\mathbf{x}} - \frac{\partial A_z}{\partial x}\,\hat{\mathbf{y}}, \tag{12.17}$$

with the derivatives taken by central differences, (5.18).

The contours of A_z have a useful property. By (12.17),

$$\mathbf{B}\cdot\nabla A_z = \frac{\partial A_z}{\partial y}\frac{\partial A_z}{\partial x} - \frac{\partial A_z}{\partial x}\frac{\partial A_z}{\partial y} = 0,$$

so $\mathbf{B}$ is everywhere perpendicular to the gradient of A_z: it runs along the curves on which A_z is constant (Figure 12.22). The contours of A_z are the field lines of $\mathbf{B}$. Moreover $|\mathbf{B}| = |\nabla A_z|$, so if the contours are drawn at equal steps of A_z, they crowd together where the field is strong, and the picture shows the strength of the field as well as its direction. The step between neighboring contours has a meaning too. A rectangle like the one in Figure 12.18b, with its long sides of length ℓ through two points 1 and 2 of the xy plane, has $\oint \mathbf{A}\cdot d\boldsymbol{\ell} = [A_z(1) - A_z(2)]\,\ell$, so the magnetic flux per unit length passing between two field lines is the difference of A_z between them.

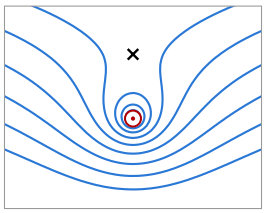


Figure 12.22 Contours of A_z at equal steps, for a long wire (current out of the page) in a uniform field pointing to the right. They are the field lines, and they crowd together where the field is strong, below the wire. The field is zero at the cross, where the wire's field cancels the uniform one.

The grid has to end somewhere, and A_z must be given on its edges. Holding $A_z = 0$ there means, by the last remark, that no flux crosses the edges, as if the conductors were enclosed in a box that magnetic flux cannot enter. That is right for some problems, such as a cable inside a superconducting shield, but not for conductors in open space. For those, put the edges of the grid well away

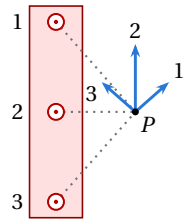


Figure 12.23 The fields at the point P of three thin filaments of current in a tall bar, seen in cross section with the current out of the page. Each field (blue) carries its filament's number and is perpendicular to the dotted line from that filament to P. The filament level with P gives a field straight up; the fields of the filaments near the top and bottom are weaker and tilted, and their sideways parts cancel.

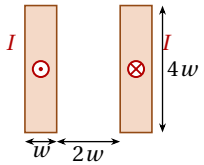


Figure 12.24 The bus bars of Example 12.7 in cross section.

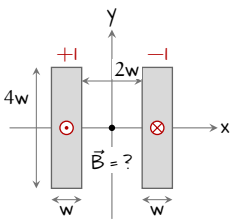
from the conductors and hold A_z there at the value an approximate solution gives, for example the thin-wire potential (12.13) of each conductor, carrying that conductor's total current. Far from conductors with a net current, A_z grows logarithmically with distance, so some reference has to be chosen anyway, and the s_0 of (12.13) does that. (Magnetic materials change the rule (12.16), just as dielectrics changed the relaxation rule for V in Section 9.9.2.)

Replacing a conductor by a thin wire at its center, as at the edges of the grid, works well far away, but close to a broad conductor it overestimates the field (Figure 12.23). Picture the conductor's current as many thin filaments, each with a field that circles it. At a point P to the right of a tall bar whose current comes out of the page, a filament level with P gives a field that points straight up. A filament near the top of the bar is farther from P, so its field there is weaker, and it is tilted away from the vertical (perpendicular to the line from the filament to P), so only part of it points up. A filament near the bottom gives the mirror image, and the sideways parts of the two cancel. Both effects make the field at P smaller than that of a thin wire at the bar's center carrying the same current.

Example 12.7

Two long, flat copper bus bars, each of width w and height $4w$, stand side by side with their broad faces $2w$ apart. They carry equal and opposite currents I, spread uniformly over their cross sections. Find the field lines, and the magnetic field midway between the bars.

Solution:



We draw the bars in cross section, with their currents and the coordinates we will use. Put the z axis along the bars, midway between them, with the bar carrying $+I$ (out of the page) centered at $x = -1.5w$ and the other at $x = +1.5w$. Inside the bars $J_z = \pm I/4w^2$, and elsewhere it is zero.

A grid with spacing $h = w/10$, covering $-15w < x, y < 15w$, puts 11 by 41 grid points in or on each bar. The square of side h around a point on a bar's edge lies only half inside the bar (a quarter, at a corner), so those points are given half the current density (a quarter at the corners), and each bar carries exactly I. On the edges of the grid, A_z is held at the potential of two thin wires at the centers of the bars. By (12.13) that is $(\mu_0 I/2\pi)\ln(s_R/s_L)$, where s_L and s_R are the distances to the centers of the left and right bars; the reference distance s_0 cancels. Seen from that far away, more than $13w$, the bars look very nearly like thin wires. On a grid this fine, relaxation needs a very large number of passes (Section 7.3.3 needed about 840 Gauss-Seidel passes on a grid of only 61 by 21 points), so the values in Figure 12.25 were found by solving the grid equations (12.16) directly with a sparse-matrix solver, which gives the values that

relaxation converges to. (In Python, `scipy.sparse.linalg.spsolve` does this; in Mathematica, `LinearSolve` with a `SparseArray`.)

Each bar is circled by field lines (Figure 12.25), and in the gap the lines crowd together and run nearly straight, upward: the field of the left bar circles it counterclockwise, the field of the right bar circles it clockwise, and in the gap both point up. Midway between the bars, (12.17) gives

$$\mathbf{B} = 0.149\,\frac{\mu_0 I}{w}\,\hat{\mathbf{y}}.$$

Reality check: The value is the same to three figures with grid spacings $w/5$, $w/10$, and $w/20$, and the same as a calculation in free space that adds up the fields of thin wires over the two cross sections.

Two simple estimates do badly. Thin wires at the centers of the bars would give $2 \times \mu_0 I/(2\pi \cdot 1.5w) = 0.212\,\mu_0 I/w$, too high by 43 percent. (Why do thin wires at the centers of the bars give too large a field? Answer in footnote.[9]) Infinite plates would give $\mu_0 K = \mu_0 I/4w = 0.250\,\mu_0 I/w$, too high by 68 percent, because plates $4w$ tall and $2w$ apart are far from infinite. For bars $w = 5.0\,\text{mm}$ wide carrying 400 A, the field midway between them is 15 mT. During a short circuit the current can be many times larger, and the forces between the bars become an engineering problem (Exercise P12.14).

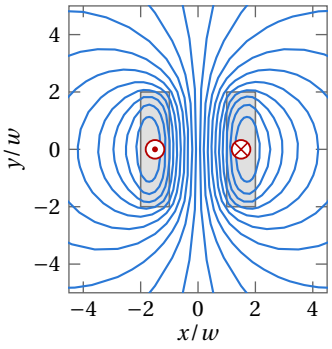


Figure 12.25 Field lines (blue) of the two bus bars (gray) of Example 12.7, carrying I out of the page (left) and into the page (right). The lines are contours of A_z, computed by solving the relaxation equations (12.16) on a grid with spacing $w/10$, at equal steps of $0.02\,\mu_0 I$; where they crowd together, as between the bars, the field is strong.

12.6 Multipole expansion of A; the magnetic dipole

Far from a localized distribution of current, such as a coil, the circulating electrons of an atom, or the molten iron in a planet's core, the details of the distribution should hardly matter, just as they hardly matter for a distribution of charge. Section 8.3 made that idea precise for V with the multipole expansion, and since each component of $\mathbf{A}$ comes from the same integral as V, the same expansion works. Put the expansion of $1/\mathscr{r}$, (8.8), into (12.9), with α again the angle between $\mathbf{r}$ and $\mathbf{r}'$ (Section 8.3.2):

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi}\sum_{n=0}^{\infty}\frac{1}{r^{n+1}}\int (r')^n P_n(\cos\alpha)\,\mathbf{J}(\mathbf{r}')\,d\tau'. \tag{12.18}$$

For a single loop of wire, (12.10) gives instead

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0 I}{4\pi}\sum_{n=0}^{\infty}\frac{1}{r^{n+1}}\oint (r')^n P_n(\cos\alpha)\,d\boldsymbol{\ell}'. \tag{12.19}$$

Compare (8.9). As there, the expansion is exact as long as the field point is farther from the origin than any of the current, and each term falls off one power of r faster than the one before. The one difference is that the integrals now add up vectors: each piece of current has a direction as well as a size.

[9] Most of each bar's current is farther from the midpoint than the bar's center is, and its field there is tilted away from the vertical, so only part of it adds to the field in the gap (Figure 12.23).

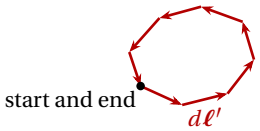


Figure 12.26 The steps $d\boldsymbol{\ell}'$ around a closed loop, laid head to tail, end where they began, so they add up to zero.

12.6.1 The monopole term is zero

For $n = 0$, $P_0 = 1$, and the loop form of the first term is $(\mu_0 I/4\pi r)\oint d\boldsymbol{\ell}'$. Adding up the steps $d\boldsymbol{\ell}'$ all the way around a closed loop brings you back to where you started, so $\oint d\boldsymbol{\ell}' = 0$ (Figure 12.26). The volume form, $(\mu_0/4\pi r)\int \mathbf{J}\, d\tau'$, vanishes too, for any steady current confined to a finite region. To see it, take the x component, and write $J_x = \mathbf{J}\cdot\nabla' x'$. By the product rule (1.52),

$$\int J_x\, d\tau' = \int \mathbf{J}\cdot\nabla' x'\, d\tau' = \int \nabla'\cdot\left(x'\mathbf{J}\right) d\tau' - \int x'\left(\nabla'\cdot\mathbf{J}\right) d\tau'.$$

By the divergence theorem, the first integral is the flux of $x'\mathbf{J}$ out of a surface that encloses all of the current, which is zero, since there is no current on that surface; the second is zero because $\nabla'\cdot\mathbf{J} = 0$ for steady currents. The same goes for the y and z components, so

$$\int \mathbf{J}\, d\tau' = 0. \tag{12.20}$$

The magnetic counterpart of the total charge is always zero: the expansion of $\mathbf{A}$ has no monopole term. That is the multipole expansion's version of $\nabla\cdot\mathbf{J} = 0$ (steady currents flow in closed paths). So $\mathbf{A}$ falls off at least as fast as $1/r^2$ far away, and $\mathbf{B}$ at least as fast as $1/r^3$, as Section 10.6.1 stated for a loop of any shape.

12.6.2 The dipole term

For $n = 1$, $P_1(\cos\alpha) = \cos\alpha$, and $r'\cos\alpha = \hat{\mathbf{r}}\cdot\mathbf{r}'$. The unit vector $\hat{\mathbf{r}}$ points toward the field point, so it is a constant as far as the integral is concerned. For a loop of wire, the dipole term is

$$\mathbf{A}_{\text{dip}}(\mathbf{r}) = \frac{\mu_0 I}{4\pi r^2}\oint \left(\hat{\mathbf{r}}\cdot\mathbf{r}'\right) d\boldsymbol{\ell}'.$$

An identity from vector calculus turns this into something simple.

An identity for the dipole term

For any constant vector $\mathbf{c}$ and any closed loop,

$$\oint \left(\mathbf{c}\cdot\mathbf{r}'\right) d\boldsymbol{\ell}' = \mathbf{a}\times\mathbf{c}, \qquad \text{where} \quad \mathbf{a} = \int d\mathbf{a}' \tag{12.21}$$

is the **vector area** of the loop, the integral of $d\mathbf{a}'$ over a surface that the loop bounds, oriented by the right-hand rule. To prove it, let $\mathbf{b}$ be another constant vector, and apply the curl theorem, (1.37), to the field $(\mathbf{c}\cdot\mathbf{r}')\,\mathbf{b}$. By the product rule for the curl, and since $\nabla'(\mathbf{c}\cdot\mathbf{r}') = \mathbf{c}$, $\nabla'\times\left[(\mathbf{c}\cdot\mathbf{r}')\,\mathbf{b}\right] = \mathbf{c}\times\mathbf{b}$, so

$$\begin{aligned}\mathbf{b}\cdot\oint \left(\mathbf{c}\cdot\mathbf{r}'\right) d\boldsymbol{\ell}' &= \int (\mathbf{c}\times\mathbf{b})\cdot d\mathbf{a}' \\ &= (\mathbf{c}\times\mathbf{b})\cdot\mathbf{a} \\ &= \mathbf{b}\cdot(\mathbf{a}\times\mathbf{c}),\end{aligned}$$

where the last step rearranges the triple product cyclically. This holds for every $\mathbf{b}$, so the two vectors dotted with $\mathbf{b}$ are equal.

With $\mathbf{c} = \hat{\mathbf{r}}$, the identity gives $\oint (\hat{\mathbf{r}} \cdot \mathbf{r}')\, d\boldsymbol{\ell}' = \mathbf{a} \times \hat{\mathbf{r}}$, and the dipole term becomes

$$\boxed{\mathbf{A}_{\text{dip}}(\mathbf{r}) = \frac{\mu_0}{4\pi} \frac{\mathbf{m} \times \hat{\mathbf{r}}}{r^2}}, \tag{12.22}$$

where

$$\mathbf{m} = I\mathbf{a} = I \int d\mathbf{a} \tag{12.23}$$

is the **magnetic dipole moment** of the loop. For a flat loop the vector area is just the area of the loop, times the unit vector perpendicular to it given by the right-hand rule: curl the fingers of your right hand along the current, and your thumb points along **a**. For a loop that is not flat, any surface bounded by the loop gives the same **a** (Figure 12.27), because the vector area of a closed surface is zero.[10] A loop made of flat pieces can be handled one piece at a time, as in Example 12.8, and for any loop $\mathbf{a} = \frac{1}{2} \oint \mathbf{r}' \times d\boldsymbol{\ell}'$ (Exercise P12.15). For a current spread through a volume, such as a spinning charged object, the dipole moment is

$$\mathbf{m} = \frac{1}{2} \int \mathbf{r}' \times \mathbf{J}(\mathbf{r}')\, d\tau', \tag{12.24}$$

which is what the $n = 1$ term of (12.18) gives, and which reduces to (12.23) for a loop (Exercise P12.15 shows both). Because $\int \mathbf{J}\, d\tau' = 0$, moving the origin by a vector **d** changes (12.24) by $-\frac{1}{2}\mathbf{d} \times \int \mathbf{J}\, d\tau' = 0$: the magnetic dipole moment does not depend on where the origin is. (An electric dipole moment has that property only when the total charge is zero.)

The unit of **m** is the ampere square meter, $\text{A} \cdot \text{m}^2$, which is the same as a joule per tesla. A loop carrying 1 A around $1\,\text{cm}^2$ has $m = 10^{-4}\,\text{A} \cdot \text{m}^2$; a small neodymium magnet a centimeter on a side has about $1\,\text{A} \cdot \text{m}^2$;[11] Earth has $8.0 \times 10^{22}\,\text{A} \cdot \text{m}^2$; and an electron, because of its spin, has $9.28 \times 10^{-24}\,\text{A} \cdot \text{m}^2$.

The terms beyond the dipole, the magnetic quadrupole and so on, fall off faster and are needed much less often. They exist, but in this book the dipole term is the only one we will use for magnetic fields. Since the monopole term is always zero, the first term that does not vanish is, for almost any current distribution, the dipole.

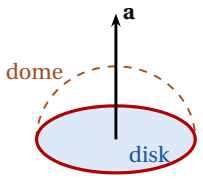


Figure 12.27 Two surfaces bounded by the same loop, a flat disk and a dome, have the same vector area **a**: turn the dome inside out, and the two make a closed surface, whose vector area is zero.

12.6.3 The field of a magnetic dipole

Point **m** along the z axis. Then $\mathbf{m} \times \hat{\mathbf{r}} = m \sin\theta\, \hat{\boldsymbol{\phi}}$, and (12.22) becomes

$$\mathbf{A}_{\text{dip}}(r, \theta) = \frac{\mu_0}{4\pi} \frac{m \sin\theta}{r^2} \hat{\boldsymbol{\phi}}. \tag{12.25}$$

[10] For any constant vector **c**, the divergence theorem gives $\mathbf{c} \cdot \oint d\mathbf{a} = \oint \mathbf{c} \cdot d\mathbf{a} = \int \nabla \cdot \mathbf{c}\, d\tau = 0$, so $\oint d\mathbf{a} = 0$. Two surfaces bounded by the same loop, one of them turned inside out, make a closed surface, so their vector areas are equal.

[11] A sintered neodymium–iron–boron magnet has a remanent magnetization $\mu_0 M \approx 1.2\,\text{T}$ ($4\pi M \approx 12\,\text{kG}$ in Gaussian units), so $M \approx 9.7 \times 10^5\,\text{A/m}$, and a cube 1 cm on a side has $m \approx 1.0\,\text{A} \cdot \text{m}^2$. The compound $Nd_2Fe_{14}B$, of which such magnets are mostly made, has a density of $7.60\,\text{g/cm}^3$, so the cube's mass is about 7.6 g (Example 15.7 uses 7.5 g). Both values are from "$R_2Fe_{14}B$ materials: Intrinsic properties and technological aspects," J. F. Herbst, Rev. Mod. Phys. 63, 819–898 (1991), Fig. 26 and Table III. https://doi.org/10.1103/RevModPhys.63.819

For a vector potential with only a ϕ component, the curl in spherical coordinates has only two terms,

$$B_r = \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\, A_\phi\right), \qquad B_\theta = -\frac{1}{r}\frac{\partial}{\partial r}\left(r A_\phi\right),$$

and they give

$$\mathbf{B}_{\text{dip}}(r,\theta) = \frac{\mu_0 m}{4\pi r^3}\left(2\cos\theta\,\hat{\mathbf{r}} + \sin\theta\,\hat{\boldsymbol{\theta}}\right). \tag{12.26}$$

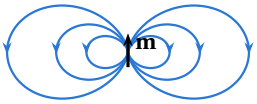


Figure 12.28 Field lines of a pure magnetic dipole **m**, in a plane through **m**: the curves $r = C\sin^2\theta$, the same pattern as for a pure electric dipole.

This has exactly the form of the electric dipole field (8.15), with p/ε_0 replaced by $\mu_0 m$ (Figure 12.28). That is remarkable: the curl of a field proportional to $\sin\theta\,\hat{\boldsymbol{\phi}}/r^2$ gives the same pattern as the gradient of a potential proportional to $\cos\theta/r^2$. As for the electric dipole, θ is measured from the direction of **m**, not from the z axis of whatever coordinate system is in use (Section 8.3.5). A form that holds for any direction of **m** is

$$\mathbf{B}_{\text{dip}}(\mathbf{r}) = \frac{\mu_0}{4\pi}\frac{3\,(\mathbf{m}\cdot\hat{\mathbf{r}})\,\hat{\mathbf{r}} - \mathbf{m}}{r^3}, \tag{12.27}$$

and the electric dipole field can be written the same way, with $\mu_0\mathbf{m}$ replaced by $\mathbf{p}/\varepsilon_0$ (Exercise P12.16).

Two special cases connect with Chapter 10. On the axis of the dipole, $\theta = 0$, the field is $\mu_0 m/2\pi r^3$ along **m**; for a ring, with $m = I\pi a^2$, that is $\mu_0 I a^2/2r^3$, the far-field limit of (10.22). In the plane through the dipole perpendicular to **m**, $\theta = 90°$ and $\hat{\boldsymbol{\theta}}$ points opposite to **m**, so the field there is $-\mu_0\mathbf{m}/4\pi r^3$, which Exercise P10.14 used.

Example 12.8

A wire follows six edges of a cube of side a, from $(a,0,0)$ to $(a,a,0)$, $(0,a,0)$, $(0,a,a)$, $(0,0,a)$, the origin, and back to $(a,0,0)$, and carries a current I in that order. Find its magnetic dipole moment, and its magnetic field far away on the x axis.

Solution:

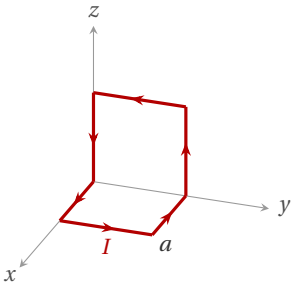
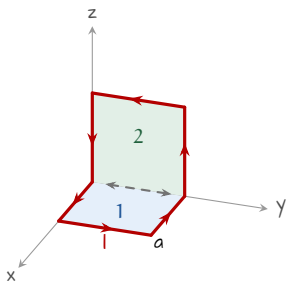


Figure 12.29 The bent loop of Example 12.8, which follows six edges of a cube of side a.



First we copy the loop of Figure 12.29. The loop is not flat, but it can be split into two flat ones. Add the edge from $(0,a,0)$ to the origin twice, once in each direction, which adds nothing (dashed). Then the current goes around two squares. The first, $(a,0,0) \to (a,a,0) \to (0,a,0) \to (0,0,0) \to (a,0,0)$, lies in the xy plane and runs counterclockwise as seen from $+z$, so its vector area is $a^2\hat{\mathbf{z}}$.

The second, $(0,a,0) \to (0,a,a) \to (0,0,a) \to (0,0,0) \to (0,a,0)$, lies in the yz plane, and its vector area is $a^2\hat{\mathbf{x}}$. (Why does the second square's vector area point along $+\hat{\mathbf{x}}$, and

not $-\hat{\mathbf{x}}$? Answer in footnote.[12]) The dipole moment is

$$\mathbf{m} = Ia^2(\hat{\mathbf{x}} + \hat{\mathbf{z}}),$$

of magnitude $\sqrt{2}\,Ia^2$, pointing halfway between the x and z axes. Far away on the x axis, at $\mathbf{r} = d\hat{\mathbf{x}}$ with $d \gg a$, $\hat{\mathbf{r}} = \hat{\mathbf{x}}$ and $\mathbf{m}\cdot\hat{\mathbf{r}} = Ia^2$, so by (12.27)

$$\mathbf{B} = \frac{\mu_0}{4\pi d^3}\left[3Ia^2\hat{\mathbf{x}} - Ia^2(\hat{\mathbf{x}} + \hat{\mathbf{z}})\right] = \frac{\mu_0 Ia^2}{4\pi d^3}(2\hat{\mathbf{x}} - \hat{\mathbf{z}}).$$

Reality check: The formula $\mathbf{a} = \frac{1}{2}\oint \mathbf{r}' \times d\boldsymbol{\ell}'$, applied to the six edges of the loop as it is, without splitting it, gives the same vector area, $a^2(\hat{\mathbf{x}} + \hat{\mathbf{z}})$.

Example 12.9

A thin spherical shell of radius R carries a uniform surface charge density σ and spins about the z axis with angular velocity ω. Find its magnetic dipole moment.

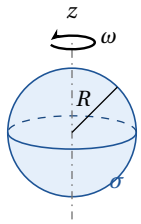
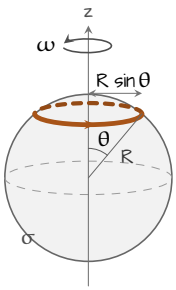


Figure 12.30 The spinning shell of Example 12.9.

Solution:



We sketch the spinning shell and one thin ring of it. Divide the shell into rings, as in Exercise P10.17.

The ring between the polar angles θ and $\theta + d\theta$ has radius $R\sin\theta$ and width $R\,d\theta$, so its charge is

$$dq = \sigma(2\pi R\sin\theta)(R\,d\theta).$$

(Why is the ring's width $R\,d\theta$, and not $R\sin\theta\,d\theta$? Answer in footnote.[13]) It goes around once in a time $2\pi/\omega$, so it is a current $dI = dq\,\omega/2\pi = \sigma\omega R^2\sin\theta\,d\theta$ around the area $\pi(R\sin\theta)^2$, and its dipole moment is

$$dm = \pi R^2\sin^2\theta\,dI = \pi\sigma\omega R^4\sin^3\theta\,d\theta,$$

along $\hat{\mathbf{z}}$. Every ring's moment points the same way, so they add (Figure 12.31):

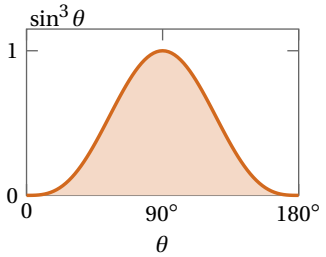


Figure 12.31 The contributions of the rings of the spinning shell to its dipole moment, $dm/d\theta \propto \sin^3\theta$: the rings near the equator, which are the largest and move the fastest, contribute the most.

[12] Seen from $+x$, with y to the right and z up, the current goes up the right side, across the top, and down the left side: counterclockwise. By the right-hand rule its normal then points toward $+x$.

[13] The width is measured along the shell, on a circle of radius R through the poles, so the angle $d\theta$ spans an arc of length $R\,d\theta$. The factor $\sin\theta$ belongs to the ring's radius, $R\sin\theta$, and so to its circumference.

$$\begin{aligned}\mathbf{m} &= \pi\sigma\omega R^4\,\hat{\mathbf{z}}\int_0^\pi \sin^3\theta\,d\theta \\ &= \frac{4\pi}{3}\,\sigma\omega R^4\,\hat{\mathbf{z}} \\ &= \frac{1}{3}\,Q\omega R^2\,\hat{\mathbf{z}},\end{aligned}$$

where $Q = 4\pi R^2\sigma$ is the total charge.

Reality check: For $\sigma > 0$ and $\omega > 0$ the charge turns counterclockwise as seen from $+z$, so the current does too, and by the right-hand rule $\mathbf{m}$ should point along $+\hat{\mathbf{z}}$, as it does. The units are right too: $\sigma\omega R^4$ has the units $(\mathrm{C/m^2})(\mathrm{s^{-1}})(\mathrm{m^4}) = \mathrm{C\cdot m^2/s} = \mathrm{A\cdot m^2}$.

This shell has a property in common with the uniformly polarized sphere of Example 9.3: its field outside is exactly the dipole field (12.26) at every distance, not only far away, and its field inside is uniform, $\frac{2}{3}\mu_0\sigma\omega R\,\hat{\mathbf{z}}$ (Exercise P12.17).

12.6.4 Near a dipole

Real dipoles have a size: a pair of charges $\pm q$ a distance d apart, with $p = qd$, or a loop of radius a, with $m = I\pi a^2$. Far away both give the same pattern of field lines. Near the source they are completely different, as Figure 12.32 shows. Between the two charges the electric field points from the positive charge to the negative one, opposite to $\mathbf{p}$. Inside the loop the magnetic field points along $\mathbf{m}$. The electric field lines begin and end on the charges; the magnetic field lines close on themselves, passing through the loop.

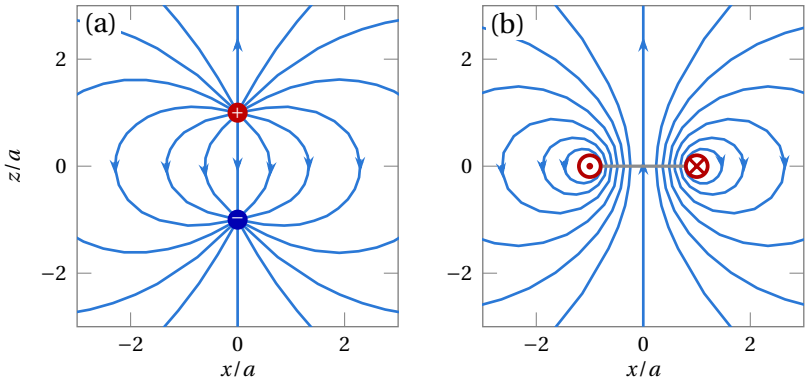


Figure 12.32 Field lines of (a) a physical electric dipole, charges $+q$ and $-q$ a distance $2a$ apart, and (b) a physical magnetic dipole, a loop of radius a seen edge on, with the current out of the page on the left and into it on the right. Both are drawn at levels chosen so that the lines coincide far away, where both fields approach the dipole pattern. Near the sources they differ completely: between the charges $\mathbf{E}$ points down, opposite to $\mathbf{p}$, while inside the loop $\mathbf{B}$ points up, along $\mathbf{m}$. The linked animation zooms out from this view and back in, showing that the two patterns become the same far from the sources.

Now shrink the sources to points, letting $d \to 0$ with qd fixed and $a \to 0$ with $I\pi a^2$ fixed. Equation (12.27) and its electric counterpart then hold everywhere except at the origin, where something must be added to account for the very different fields inside the two sources. The complete fields of point dipoles turn out to be

$$\mathbf{E}_{\text{dip}}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\frac{3(\mathbf{p}\cdot\hat{\mathbf{r}})\hat{\mathbf{r}}-\mathbf{p}}{r^3} - \frac{1}{3\varepsilon_0}\mathbf{p}\,\delta^3(\mathbf{r}), \tag{12.28}$$

$$\mathbf{B}_{\text{dip}}(\mathbf{r}) = \frac{\mu_0}{4\pi}\frac{3(\mathbf{m}\cdot\hat{\mathbf{r}})\hat{\mathbf{r}}-\mathbf{m}}{r^3} + \frac{2\mu_0}{3}\mathbf{m}\,\delta^3(\mathbf{r}). \tag{12.29}$$

The delta-function terms come from averages. The average of $\mathbf{E}$ over the volume of a sphere of radius R, centered on the origin, that contains a distribution of charge is $-\mathbf{p}/4\pi\varepsilon_0 R^3$, and the average of $\mathbf{B}$ over such a sphere that contains a distribution of current is $+(\mu_0/4\pi)(2\mathbf{m}/R^3)$, whatever the size of the source, as long as it is inside the sphere. The $1/r^3$ terms contribute nothing to these averages (their average over the directions of $\hat{\mathbf{r}}$ is zero), so the delta-function terms must supply all of it. Exercise P12.19 derives both averages. The two delta functions have opposite signs, matching the opposite directions of the fields inside the two sources in Figure 12.32.

Does the difference matter? For hydrogen in its ground state it decides the answer. The electron's probability cloud is spherically symmetric and sits right on top of the proton (Figure 12.33), so the $1/r^3$ part of the proton's magnetic field averages to zero over it, and the whole of the magnetic interaction between the proton and the electron comes from the delta-function term. That interaction splits the ground state into two levels 5.9×10^{-6} eV apart, and transitions between them emit radio waves with a wavelength of 21 cm, which astronomers use to map hydrogen in our galaxy. The measured splitting agrees with the current-loop form, $(2\mu_0/3)\,\mathbf{m}\,\delta^3(\mathbf{r})$. If the proton's magnetic moment came from a pair of magnetic charges, the delta-function term would instead be $-(\mu_0/3)\,\mathbf{m}\,\delta^3(\mathbf{r})$, by analogy with (12.28), giving a splitting of the opposite sign and half the size.

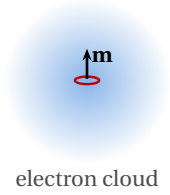


Figure 12.33 Hydrogen in its ground state, not to scale: the electron's spherically symmetric probability cloud (shaded) is centered on the proton, a magnetic dipole $\mathbf{m}$.

12.7 Inductance and the energy stored in magnetic fields

Two topics that belong with the vector potential get only a short look here, because their full treatment needs Faraday's law of induction, the subject of Chapter 15. They are introduced now so that they will be familiar when they appear among the parallel equations of Chapter 14. Chapter 19 takes them up properly (Section 19.4).

12.7.1 Inductance

Consider two loops of wire, as in Figure 12.34. A current I_1 in loop 1 produces a field proportional to I_1, since the Biot–Savart law is linear in the current, so the magnetic flux of that field through loop 2 is proportional to I_1 too:

$$\Phi_2 = M_{21} I_1. \tag{12.30}$$

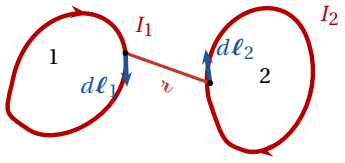


Figure 12.34 Two loops of wire, with an element $d\boldsymbol{\ell}_1$ of one, an element $d\boldsymbol{\ell}_2$ of the other, and the distance r between them. The Neumann formula (12.31) adds up $d\boldsymbol{\ell}_1 \cdot d\boldsymbol{\ell}_2/r$ over every pair of elements.

The constant M_{21} depends only on the sizes, shapes, and positions of the two loops. It is called the **mutual inductance**. The vector potential gives a formula for it at once. By (12.4), $\Phi_2 = \oint_2 \mathbf{A}_1 \cdot d\boldsymbol{\ell}_2$, where $\mathbf{A}_1$ is the vector potential of loop 1, and by (12.10), $\mathbf{A}_1 = (\mu_0 I_1/4\pi) \oint_1 d\boldsymbol{\ell}_1/r$. So

$$M_{21} = \frac{\mu_0}{4\pi} \oint_1 \oint_2 \frac{d\boldsymbol{\ell}_1 \cdot d\boldsymbol{\ell}_2}{r}, \qquad \text{(Neumann formula)} \quad (12.31)$$



Franz Ernst Neumann (1798–1895, German) Neumann fought as a volunteer in the Prussian army against Napoleon and was badly wounded at the battle of Ligny in 1815. He then studied theology in Berlin, turned to science at Jena, and returned to Berlin to study mineralogy and crystallography under Christian Samuel Weiss. From 1826 he taught at Königsberg, where he held the chair of mineralogy and physics from 1829. In the 1830s he and the mathematician Carl Jacobi founded a seminar there in mathematical physics that trained students in research and in the careful treatment of experimental errors; Gustav Kirchhoff was one of its students. In papers of 1845 and 1847 Neumann gave the first mathematical theory of electromagnetic induction. Later writers credited him with inventing the vector potential, although it appears in his formulas only implicitly. His formula for the mutual inductance of two circuits carries his name. (The Neumann boundary condition is named for his son Carl, a mathematician.) Neumann lived to the age of 96. (Wikipedia)

where r is the distance between the elements $d\boldsymbol{\ell}_1$ and $d\boldsymbol{\ell}_2$. Nothing in the formula changes if the loops trade places, so $M_{21} = M_{12}$, and we can drop the subscripts and write M. With words: the flux through loop 2 per unit current in loop 1 is the same as the flux through loop 1 per unit current in loop 2. That is far from obvious when one loop is large and the other small (Exercise P12.20).

A loop's own field also passes through the loop itself, and that flux is proportional to its own current,

$$\Phi = LI. \quad (12.32)$$

The constant L is the **self-inductance**, or just the **inductance**, of the loop. Both kinds of inductance are measured in **henrys**: $1\,\mathrm{H} = 1\,\mathrm{Wb/A} = 1\,\mathrm{T\cdot m^2/A}$. For a coil of many turns, Φ means the total flux through all of the turns. For a long solenoid of length ℓ, the flux through each turn is $\mu_0 nI\pi R^2$ and there are $n\ell$ turns, so

$$L = \mu_0 n^2 \pi R^2 \ell. \quad (12.33)$$

For a wire of zero thickness, L would be infinite, because the field near a thin wire falls off only as $1/s$ and the flux close to the wire diverges logarithmically; real wires have a thickness, and their inductance is finite. The name comes from what happens when the currents change. A changing current in loop 1 changes the flux through loop 2, and by Faraday's law that induces an EMF in loop 2, in proportion to M; a changing current in a single loop induces an EMF in that loop itself, in proportion to L (Section 15.2.2).

12.7.2 Energy stored in the magnetic field

In Section 5.9 the energy of a distribution of charge was the work needed to assemble it, and it could be written in terms of the charges or in terms of the field, (5.22) and (5.23). The corresponding results for steady currents are

$$U = \frac{1}{2}\int \mathbf{A}\cdot\mathbf{J}\, d\tau, \quad (12.34)$$

$$U = \frac{1}{2\mu_0}\int B^2\, d\tau, \quad (12.35)$$

with the second integral over all space. The energy per unit volume of the magnetic field is

$$u_B = \frac{B^2}{2\mu_0}, \quad (12.36)$$

to be compared with $\varepsilon_0 E^2/2$ for the electric field. For a single loop of wire, $\mathbf{J}\,d\tau$ becomes $I\,d\boldsymbol{\ell}$, and by (12.4) and (12.32), (12.34) becomes

$$\begin{aligned} U &= \frac{1}{2} I \oint \mathbf{A}\cdot d\boldsymbol{\ell} \\ &= \frac{1}{2} I\Phi \\ &= \frac{1}{2} LI^2, \end{aligned} \tag{12.37}$$

to be compared with $Q^2/2C$ for a capacitor.

The two forms (12.34) and (12.35) are equal, and showing it needs only Ampère's law, just as the step from $\frac{1}{2}\int \rho V\,d\tau$ to $\frac{1}{2}\varepsilon_0 \int E^2\,d\tau$ in Section 5.9.2 needed only Gauss's law. Put $\mu_0\mathbf{J} = \nabla\times\mathbf{B}$ into (12.34), and move the curl from $\mathbf{B}$ to $\mathbf{A}$ with the product rule

$$\nabla\cdot(\mathbf{A}\times\mathbf{B}) = \mathbf{B}\cdot(\nabla\times\mathbf{A}) - \mathbf{A}\cdot(\nabla\times\mathbf{B})$$

(Section 1.7). Then

$$\frac{1}{2}\int \mathbf{A}\cdot\mathbf{J}\,d\tau = \frac{1}{2\mu_0}\int \mathbf{A}\cdot(\nabla\times\mathbf{B})\,d\tau = \frac{1}{2\mu_0}\int \mathbf{B}\cdot(\nabla\times\mathbf{A})\,d\tau - \frac{1}{2\mu_0}\oint (\mathbf{A}\times\mathbf{B})\cdot d\mathbf{a},$$

using the divergence theorem on the last term. Take the volume to be a large sphere (Figure 12.35). Far away $\mathbf{A}$ falls off at least as fast as $1/r^2$ and $\mathbf{B}$ as $1/r^3$, the dipole terms, while the area of the sphere grows only as r^2, so the surface term goes to zero. And $\mathbf{B}\cdot(\nabla\times\mathbf{A}) = B^2$.

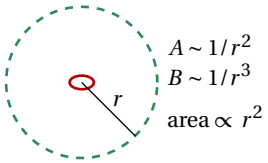


Figure 12.35 The surface term: on a large sphere of radius r around the currents, $\mathbf{A}\times\mathbf{B}$ falls off at least as fast as $1/r^5$, while the area grows only as r^2.

What has not been shown is that U is the work needed to set up the currents, and that cannot be shown without Faraday's law. The reason goes back to Section 10.1.3: magnetic forces do no work. So whatever work builds up the currents must be done against an electric field. While the currents are growing, the magnetic field is changing, and a changing magnetic field induces an electric field; that is Faraday's law. The batteries that drive the currents do work against this induced field, and it is that work that ends up stored as (12.35). Bringing two loops together at fixed currents shows where the energy comes from. If the loops share an axis and their currents circulate the same way, M is positive and the loops attract, so whoever brings one in from far away does a negative amount of work, $-MI_1I_2$, while the field energy increases by $+MI_1I_2$. The difference, $2MI_1I_2$, comes from the batteries, which must work to keep the currents fixed against the EMFs induced as the flux through each loop changes. Exercise P19.11 works through the numbers.

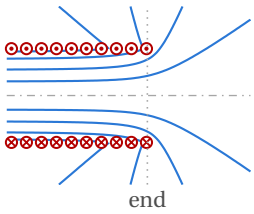
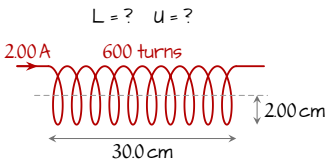


Figure 12.36 Field lines near one end of a solenoid 15 times as long as its radius, computed by adding up the fields of its turns (cut along the axis; turns shown as dots and crosses). Near the end, some lines leave through the side of the coil, so the end turns enclose less flux than $\mu_0 nI\pi R^2$.

Example 12.10

A solenoid 30.0 cm long and 2.00 cm in radius has 600 turns and carries a current of 2.00 A. Treating it as a long solenoid, find its inductance and the energy stored in its magnetic field, and check the energy with (12.35).

Solution:



Before any numbers go in, we sketch the solenoid and label it with what we know and what we want.

The number of turns per unit length is $n = 600/0.300\,\text{m} = 2000\,\text{m}^{-1}$, so the field inside is $B = \mu_0 nI = 5.03\times10^{-3}\,\text{T}$, and by (12.33)

$$L = \mu_0 n^2 \pi R^2 \ell = \left(4\pi\times10^{-7}\,\text{T}\cdot\text{m/A}\right)\left(2000\,\text{m}^{-1}\right)^2 \pi\,(0.0200\,\text{m})^2\,(0.300\,\text{m})$$
$$= 1.89\times10^{-3}\,\text{H}.$$

The energy is $U = \frac{1}{2}LI^2 = 3.79\times10^{-3}\,\text{J}$.

Reality check: The energy density inside is $B^2/2\mu_0 = 10.1\,\text{J/m}^3$, the field is confined to the volume $\pi R^2\ell = 3.77\times10^{-4}\,\text{m}^3$, and their product is 3.79×10^{-3} J again. (Why does only the volume inside the solenoid contribute to the energy? Answer in footnote.[14]) A real solenoid with these proportions, 15 times as long as its radius, has an inductance about 5 percent smaller, because near its ends the field is weaker than $\mu_0 nI$, so less flux passes through the end turns (Figure 12.36). (The 5 percent comes from a numerical calculation of the flux through every turn of a finite solenoid.)

12.8 Summary: Triangle diagram of J, B, A

The magnetic field has no divergence. That one fact makes the field the curl of a vector potential, and together with Ampère's law it makes the current density, the field, and the vector potential three ways of describing the same thing, as ρ, **E**, and V are in electrostatics. Each can be found from either of the others, and the relations among them fit on a triangle diagram like the one in Section 5.11:

[14] In a long solenoid the field is zero outside, (11.7), so the energy density $B^2/2\mu_0$ is zero there too.

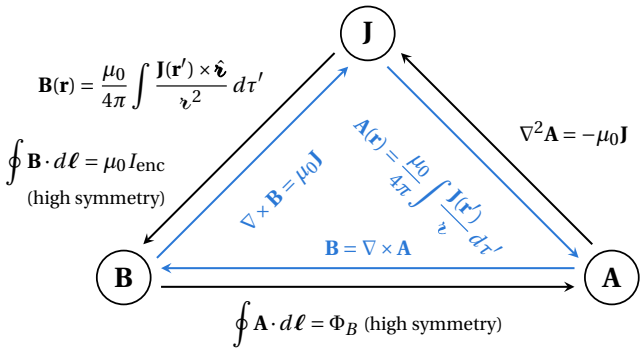


Figure 12.37 The triangle diagram for magnetostatics. Each of **J**, **B**, and **A** can be found from either of the others. The black arrows go counterclockwise around the outside of the triangle, and the blue arrows go clockwise around the inside. Poisson's equation and the script-r integral for **A** assume the Coulomb gauge, and the two volume integrals assume currents confined to a finite region.

If you know any one of the three, you can find the other two. The route from **B** to **A** is the one used least. When the symmetry is too low for the flux method, it is usually easiest to go around the triangle, from **B** to **J** by Ampère's law and from **J** to **A** by the script-r integral.

$\nabla\cdot\mathbf{B} = 0, \quad \mathbf{B} = \nabla\times\mathbf{A}$	the field has no divergence, so it is the curl of a vector potential
$\mathbf{A}' = \mathbf{A} + \nabla\psi, \quad \nabla\cdot\mathbf{A} = 0$	gauge freedom; the Coulomb gauge, used throughout
$\oint \mathbf{A}\cdot d\boldsymbol{\ell} = \Phi_B$	the circulation of **A** is the flux of **B**; the flux method
$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi}\int \frac{\mathbf{J}(\mathbf{r}')}{\imath}\,d\tau'$	**A** from the current, going to zero far away; $I\,d\boldsymbol{\ell}'$ or $\mathbf{K}\,da'$ for wires and sheets
$\nabla^2\mathbf{A} = -\mu_0\mathbf{J}$	Poisson's equation for **A**, one for each Cartesian component
$\mathbf{A}_{\text{above}} = \mathbf{A}_{\text{below}}$	at a surface current, **A** is continuous, and $\partial\mathbf{A}/\partial n$ jumps by $-\mu_0\mathbf{K}$

$\mathbf{m} = I\mathbf{a}, \quad \mathbf{m} = \frac{1}{2}\int \mathbf{r}' \times \mathbf{J}\, d\tau'$	the magnetic dipole moment of a loop, and of a volume current
$\mathbf{A}_{\text{dip}} = \dfrac{\mu_0}{4\pi}\dfrac{\mathbf{m}\times\hat{\mathbf{r}}}{r^2}$	the leading term far from any localized current; there is no monopole term
$\mathbf{B}_{\text{dip}} = \dfrac{\mu_0}{4\pi}\dfrac{3(\mathbf{m}\cdot\hat{\mathbf{r}})\hat{\mathbf{r}} - \mathbf{m}}{r^3}$	the dipole field, the same pattern as the electric one; add $\frac{2}{3}\mu_0\mathbf{m}\delta^3(\mathbf{r})$ for a point dipole
$M = \dfrac{\mu_0}{4\pi}\oint\oint \dfrac{d\boldsymbol{\ell}_1 \cdot d\boldsymbol{\ell}_2}{\mathscr{r}}, \quad \Phi = LI$	mutual inductance (Neumann formula), $M_{12} = M_{21}$; self-inductance
$U = \dfrac{1}{2}\int \mathbf{A}\cdot\mathbf{J}\, d\tau = \dfrac{1}{2\mu_0}\int B^2\, d\tau$	the energy stored in the field, the work needed to set up the currents (Section 19.4.2)
$U = \dfrac{1}{2}LI^2$	the same energy, for a single loop

The components of A behave like V. In the Coulomb gauge each Cartesian component of $\mathbf{A}$ obeys Poisson's equation, with $\mu_0 J_x$, $\mu_0 J_y$, or $\mu_0 J_z$ in place of ρ/ε_0, so every result for V carries over, with $\rho/\varepsilon_0 \to \mu_0 J_x$, $\sigma/\varepsilon_0 \to \mu_0 K_x$, and $\lambda/\varepsilon_0 \to \mu_0 I$. That holds for Cartesian components only. The vector potential points along the currents that produce it, weighted toward the nearby ones.

Choosing a route. If the symmetry allows Ampère's law, find **B** first and then **A** by the flux method: a circle around the axis when **A** circles it, a rectangle in a plane containing the axis when **A** is parallel to it. If not, find **A** with the script-r integral, which has no cross products, and take its curl if the field is wanted. Variable unit vectors must still be written in Cartesian components before they are integrated. For long parallel conductors of any cross section, relax A_z on a grid; its contours are field lines, spaced to show the strength of the field.

Dipoles. Far from any localized current with a nonzero dipole moment, **A** approaches the dipole potential. The dipole moment of a flat loop is the current times the area, along the normal given by the right-hand rule, and it does not depend on the choice of origin. As for electric dipoles, θ is measured from **m**. Near a physical dipole the electric and magnetic fields are very different: inside a current loop **B** points along **m**.

Results worth remembering. An infinite straight wire: $\mathbf{A} = -(\mu_0 I/2\pi)\ln(s/s_0)\,\hat{\mathbf{z}}$. A uniform field: $\mathbf{A} = \frac{1}{2}\mathbf{B}\times\mathbf{r}$. A long solenoid: $\frac{1}{2}\mu_0 nIs\hat{\boldsymbol{\phi}}$ inside and $\mu_0 nIR^2/2s\hat{\boldsymbol{\phi}}$ outside, not zero where the field is zero. A loop of radius a, far away: a dipole with $m = I\pi a^2$. A long solenoid's inductance: $L = \mu_0 n^2 \pi R^2 \ell$.

Exercises

Exercises for 12.1 The magnetic vector potential

Warm-up questions Read Section 12.1, then answer before class

W12.1 A classmate reasons that, since V is defined only up to an added constant, $\mathbf{A}$ must be defined only up to an added constant vector. Do the two potentials of Example 12.1 differ by a constant vector? What, in general, can be added to $\mathbf{A}$ without changing $\mathbf{B}$, and where does a constant vector fit in?

W12.2 This section writes $\mathbf{B} = \nabla \times \mathbf{A}$. Which of Maxwell's equations makes that possible, and why? Why would the discovery of a magnetic monopole be a problem for writing $\mathbf{B}$ that way?

W12.3 Two students describe the same magnetic field with vector potentials in different gauges. Which of these must they agree on, and why: (a) $\mathbf{A}$ at a given point; (b) $\mathbf{B}$ at a given point; (c) $\nabla \cdot \mathbf{A}$ at a given point; (d) $\oint \mathbf{A} \cdot d\boldsymbol{\ell}$ around a given loop?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P12.1 The uniform field $\mathbf{B} = B_0 \hat{\mathbf{z}}$ of Example 12.1 has the vector potentials $\mathbf{A}_1$ and $\mathbf{A}_2$ found there.

(a) Show that $\mathbf{A}_3 = -B_0 y \hat{\mathbf{x}}$ is a third vector potential for the same field, and that it is in the Coulomb gauge. Find a function ψ with $\mathbf{A}_3 = \mathbf{A}_1 + \nabla\psi$.

(b) Calculate $\oint \mathbf{A} \cdot d\boldsymbol{\ell}$ for each of the three potentials around a circle of radius s in the xy plane, centered on the origin and traversed counterclockwise as seen from $+z$. Repeat for the same circle centered on the point $(x_0, y_0, 0)$. Compare your results with the flux through the circle.

(c) $\mathbf{A}_1$ circles the z axis. Does that mean the z axis is special for this field? Explain why the question "where is the center of $\mathbf{A}_1$?" has no physical meaning.

(d) For currents confined to a finite region, the Coulomb gauge together with the condition that $\mathbf{A}$ go to zero far away determines $\mathbf{A}$ uniquely. Why does that not help here?

P12.2 A vector potential is given as $\mathbf{A}_0 = kxy\hat{\mathbf{x}}$, with k a constant.

(a) Find the magnetic field, and the current density that produces it.

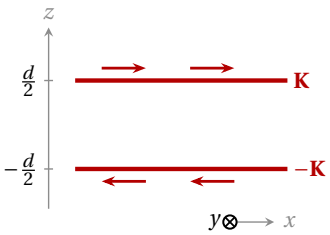


Figure 12.38 The two sheets of Exercise P12.3, edge on.

(b) Show that $\mathbf{A}_0$ is not in the Coulomb gauge. Find a function $\psi(y)$ for which $\mathbf{A} = \mathbf{A}_0 + \nabla\psi$ is, and check that $\mathbf{A}$ gives the same field.

(c) Calculate $\nabla^2\mathbf{A}$ and $\nabla^2\mathbf{A}_0$, and compare each with $-\mu_0\mathbf{J}$. Which one satisfies Poisson's equation (12.7) (Section 12.2), and why?

(d) Could $\mathbf{A}$ have been found from the script-r integral (12.9)? Explain.

P12.3 Two large parallel sheets carry equal and opposite surface currents: $\mathbf{K} = K\hat{\mathbf{x}}$ on the plane $z = d/2$ and $-K\hat{\mathbf{x}}$ on the plane $z = -d/2$ (Figure 12.38).

(a) Find $\mathbf{B}$ everywhere, using (11.6) for each sheet.

(b) Find a vector potential of the form $A_x(z)\hat{\mathbf{x}}$, in the Coulomb gauge, that is zero on the plane $z = 0$. Sketch $A_x(z)$, and show that $\mathbf{A}$ is not zero outside the sheets, where $\mathbf{B}$ is zero.

(c) Check the boundary conditions (12.5) and (12.6) at both sheets.

(d) Show that $A_x(d/2) - A_x(-d/2)$ is the magnetic flux, per unit length in the x direction, between the sheets. Compare this situation with a parallel-plate capacitor.

P12.4 An exam once asked which of these is continuous, in general, as you move past a boundary: all of the vector potential $\mathbf{A}$; not all of $\mathbf{A}$, just its perpendicular component; not all of $\mathbf{A}$, just its parallel component; or nothing is guaranteed to be continuous.

(a) Which answer is right, and why? How does the answer change in the Coulomb gauge?

(b) For the current sheet of Section 12.1.3, take $\mathbf{A} = -\frac{1}{2}\mu_0 K|z|\hat{\mathbf{x}}$ and make the gauge transformation (12.2) with $\psi = c|z|$, where c is a constant. Show that the new potential gives the same field, that its perpendicular component jumps at the sheet, and that it is not in the Coulomb gauge there.

(c) Explain why no gauge transformation can make the parallel component of $\mathbf{A}$ jump without making $\mathbf{A}$ infinite at the surface.

Exercises for 12.2 Poisson's equation for A

Warm-up questions Read Section 12.2, then answer before class

W12.4 A classmate says that, since $\mathbf{B}$ circles a straight wire, $\mathbf{A}$ must circle it too. Which way does the potential (12.9) point near a straight wire, and near a circular loop? Does it always point along the nearby currents? Which potential in Section 12.1.3 does not, and why?

W12.5 Write $\nabla^2 A_y = -\mu_0 J_y$ as a sentence that a classmate who knows electrostatics, but has not read this chapter, would understand. Then say which of the three cautions of Section 12.2.1 rules out the scriptr integral (12.9) for an infinitely long straight wire, and what goes wrong if you use it anyway.

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P12.5 This exercise shows why (12.7) must be read in Cartesian components.

(a) Write $\mathbf{F} = f(s)\,\hat{\boldsymbol{\phi}}$ in Cartesian components, using $\hat{\boldsymbol{\phi}} = (-y\hat{\mathbf{x}} + x\hat{\mathbf{y}})/s$, and take the Laplacian of each component. Show that

$$\nabla^2 \left[f(s)\,\hat{\boldsymbol{\phi}} \right] = \left(\frac{d^2 f}{ds^2} + \frac{1}{s}\frac{df}{ds} - \frac{f}{s^2} \right) \hat{\boldsymbol{\phi}}.$$

In particular, $\nabla^2 \hat{\boldsymbol{\phi}} = -\hat{\boldsymbol{\phi}}/s^2$.

(b) Use this to check that the vector potential of a solenoid, (12.15) (Section 12.4), satisfies $\nabla^2 \mathbf{A} = 0$ both inside and outside.

(c) Show that if you instead took the Laplacian of A_ϕ alone, as a scalar, and multiplied by $\hat{\boldsymbol{\phi}}$, you would conclude that $\nabla^2 \mathbf{A} \neq 0$ outside the solenoid.

(d) Check the result of part (a) with the identity $\nabla^2 \mathbf{F} = \nabla(\nabla \cdot \mathbf{F}) - \nabla \times (\nabla \times \mathbf{F})$, using the formulas for the divergence and curl in cylindrical coordinates.

P12.6 Two long straight wires (Figure 12.39), parallel to the z axis, pass through the points $(\pm d, 0, 0)$. The one at $x = +d$ carries a current I in the $+z$ direction and the one at $x = -d$ carries I in the $-z$ direction, as in part (b) of Exercise P10.11.

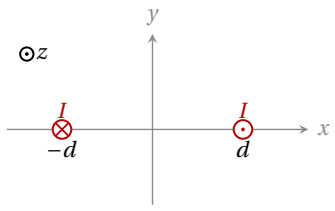


Figure 12.39 The two wires of Exercise P12.6, seen end on.

(a) Use the correspondence of Section 12.2.1, and the potential of an infinite line of charge (Example 5.1), once for each wire, to write down $A_z(x, y)$. Explain why the reference point can be put at infinity here, although it cannot for a single wire.

(b) Show that the curves A_z = constant, which are the field lines (Section 12.5), are circles. Find the center and radius of the circle on which $s_-/s_+ = k$, where $s_\pm$ are the distances from the two wires.

(c) Use $\mathbf{B} = \nabla \times (A_z \hat{\mathbf{z}})$, (12.17), to find $\mathbf{B}$ on the y axis, and check it against the field you get by adding the fields of the two wires.

(d) How does A_z depend on distance far from both wires?

Exercises for 12.3 Solving for A when given J: script-r method

Warm-up questions Read Section 12.3, then answer before class

W12.6 Compare the script-r integral for a thin wire, (12.10), with the Biot–Savart law for the same wire. Name two differences in the integrands that usually make the integral for **A** easier. Which precaution from the Biot–Savart integrals still applies, and why?

W12.7 A wire bent into a semicircle of radius a runs from $(a,0,0)$ through $(0,a,0)$ to $(-a,0,0)$ and carries a current I, as one piece of a closed circuit. Every piece of it is the same distance from the origin. Use that to find its contribution to **A** at the origin from (12.10), without doing an integral. (What do the pieces $d\boldsymbol{\ell}'$ add up to?) Does your answer depend on a?

W12.8 On the axis of the loop of Example 12.4, **A** is zero, yet the field there is not. A classmate suspects a mistake, since $\mathbf{B} = \nabla \times \mathbf{A}$. Resolve the puzzle: what does **B** at a point depend on, if not the value of **A** there? (Example 5.5 met the same puzzle for V and **E**.)

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P12.7 The vector potential (12.12) of a straight segment is not in the Coulomb gauge.

(a) Show that its divergence is $(\mu_0 I/4\pi)(1/\mathscr{r}_1 - 1/\mathscr{r}_2)$, where $\mathscr{r}_1$ and $\mathscr{r}_2$ are the distances from the field point to the end where the current enters and the end where it leaves.

(b) Explain in terms of charge why the divergence should not be zero for a segment on its own.

(c) A closed circuit is made of straight segments, the end of each one joined to the start of the next. Show that the divergences of the segments' potentials cancel, so that their sum is in the Coulomb gauge.

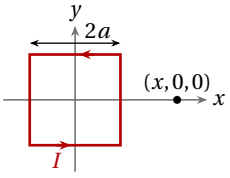


Figure 12.40 The loop of Exercise P12.8 and a field point on the x axis.

P12.8 The square loop of Exercise P10.14, with sides of length $2a$, lies in the xy plane, centered on the origin with its sides parallel to the axes, and carries a current I counterclockwise as seen from $+z$ (Figure 12.40). (Use Mathematica or another computer algebra system if you like.)

(a) Use (12.12) for each side to find **A** at points $(x,0,0)$ on the x axis. Explain why it points along $\hat{\mathbf{y}}$ there, and why it is zero at the origin.

(b) Find **A** at a general point $(x,y,0)$ in the plane of the loop, and use it to compute $B_z = \partial A_y/\partial x - \partial A_x/\partial y$ at points on the x axis. Compare with the result of Exercise P10.14.

(c) Far away, **A** should approach the dipole potential (12.22), with $\mathbf{m} = 4a^2 I\hat{\mathbf{z}}$ (Section 12.6.2). Evaluate $A_y(x,0,0)$ at $x = 5a$ and $x = 10a$ in units of $\mu_0 m/4\pi x^2$. How far away must you be for the dipole approximation to be good to 1 percent along the x axis?

P12.9 This exercise uses the vector potential of a circular loop, (12.14).

(a) Write a program that evaluates (12.14) numerically. Check it against the elliptic-integral form in the footnote to Example 12.4 at a few points.

(b) Compare A_ϕ with the dipole approximation in the plane of the loop ($\theta = 90°$) and along the line $\theta = 45°$. Beyond what distance, in units of a, is the dipole approximation good to 1 percent in each direction?

(c) Divide the loop into N straight pieces, joining points evenly spaced around it, and use the chunk-and-sum method, treating each straight piece as a current element at its midpoint, to find A_ϕ at $(s,z) = (1.5a, 0)$. About how large must N be for an accuracy of 0.1 percent?

Exercises for 12.4 Solving for A when given B: flux method (high symmetry)

Warm-up questions Read Section 12.4, then answer before class

W12.9 To find **A** beside a long straight wire, the flux method uses a rectangle in a plane that contains the wire (Figure 12.18b). Suppose you used a circle centered on the wire, in a plane perpendicular to it, instead. What would each side of $\oint \mathbf{A}\cdot d\boldsymbol{\ell} = \Phi_B$ be, and why would that tell you nothing about **A**?

W12.10 In Example 12.5, suppose the radius R were doubled, with n and I unchanged. By what factor would **A** change at a distance s inside both the old and the new solenoid? At a distance s outside both? Explain each answer with the flux through the loop, not only with the formula.

W12.11 A classmate argues that the vector potential outside a long solenoid, where **B** is zero, can be made zero by a suitable choice of gauge, just as the potential inside a charged spherical shell can be made zero by putting the reference point there. What is wrong with the argument? Which quantity that is the same in every gauge rules it out?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P12.10 A long coaxial cable has a solid inner conductor of radius a, carrying a current I in the $+z$ direction spread uniformly over its cross section,

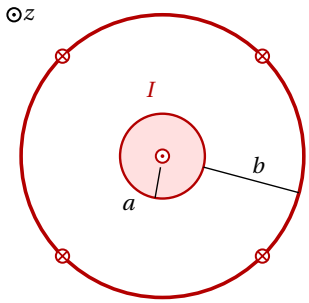


Figure 12.41 The coaxial cable of Exercise P12.10, in cross section.

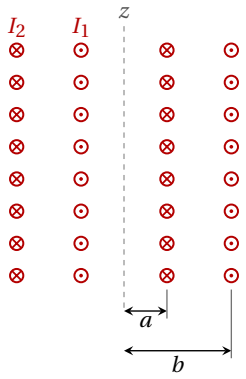


Figure 12.42 The two solenoids of Exercise P12.11, cut along their common axis.

and a thin outer conductor of radius b carrying the same current back (Figure 12.41).

(a) Use the flux method to find $\mathbf{A}$ in the three regions, taking $\mathbf{A} = 0$ outside the cable. Why is that choice possible here?

(b) Check that $\mathbf{A}$ satisfies Poisson's equation in each region, and the boundary conditions (12.5) and (12.6) at $s = a$ and $s = b$.

(c) Find the magnetic flux, per unit length of cable, between the two conductors, and express it in terms of $\mathbf{A}$.

P12.11 Two long solenoids share an axis (Figure 12.42). The inner one has radius a and n_1 turns per unit length and carries a current I_1; the outer one has radius $b > a$ and n_2 turns per unit length and carries a current I_2 in the opposite sense.

(a) Find $\mathbf{B}$ in the three regions $s < a$, $a < s < b$, and $s > b$.

(b) Use the flux method to find $\mathbf{A}$ in the three regions.

(c) Under what condition is $\mathbf{A}$ zero outside both solenoids? What is $\mathbf{B}$ between them then?

P12.12 Section 12.1.2 pointed out that $\nabla \times \mathbf{A} = \mathbf{B}$ and $\nabla \cdot \mathbf{A} = 0$ have the same form as $\nabla \times \mathbf{B} = \mu_0 \mathbf{J}$ and $\nabla \cdot \mathbf{B} = 0$.

(a) Argue that, when $\mathbf{B}$ goes to zero fast enough far away, the Coulomb-gauge vector potential that goes to zero far away is

$$\mathbf{A}(\mathbf{r}) = \frac{1}{4\pi} \int \frac{\mathbf{B}(\mathbf{r}') \times \hat{\boldsymbol{\imath}}}{\imath^2} \, d\tau'.$$

(b) The integral of part (a) does not apply directly to solenoids of infinite length, whose field does not go to zero along their axis, but the analogy behind it still holds; Section 12.4 used it for a single solenoid. Use the field of a long cylinder carrying a uniform current density J_1 along its axis for $s < a$ and a different uniform current density J_2 for $a < s < b$, found by Ampère's law, to write down $\mathbf{A}$ for the two solenoids of Exercise P12.11 without any calculation, and compare with your answer there.

(c) Do the same for the pair of sheets of Exercise P12.3, using the field of a slab of uniform current density.

Exercises for 12.5 Numerical method: relaxation for A in two dimensions

Warm-up questions Read Section 12.5, then answer before class

W12.12 In a two-dimensional problem like those of this section, two field lines are the contours $A_z = 2.0 \times 10^{-6}\,\mathrm{T\cdot m}$ and $A_z = 5.0 \times 10^{-6}\,\mathrm{T\cdot m}$. How much magnetic flux passes between them through a strip 0.50 m long in the z direction? When contours are drawn at equal steps of A_z, why do they crowd together where the field is strong?

W12.13 Suppose every value of A_z on the grid of Example 12.7, edges included, were increased by the same constant. Would the values still satisfy (12.16)? What would happen to **B**, and to the field lines of Figure 12.25? What kind of change to **A** is this, in the language of Section 12.1.1?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P12.13 Write a program that solves (12.16) for two long conductors of square cross section, with sides w, centered at $x = \pm 3w$, carrying currents $+I$ (at $-3w$) and $-I$ (at $+3w$) uniformly over their cross sections (Figure 12.43). Use a grid spacing of $w/5$ or smaller. Relaxation is slow on the larger grids; you may instead solve the grid equations directly with a sparse-matrix solver, as in Example 12.7.

(a) Hold $A_z = 0$ on the edges of a square grid that extends to $\pm L$ in x and y. Find B midway between the conductors for $L = 10w$, $20w$, and $40w$, and compare with the field of two thin wires at the conductors' centers. About how large must L be for an error below 1 percent?

(b) Now hold A_z on the edges at the potential of two thin wires, and repeat with $L = 10w$. Why does this work so much better?

(c) Plot the contours of A_z, and compare them with the circles of Exercise P12.6.

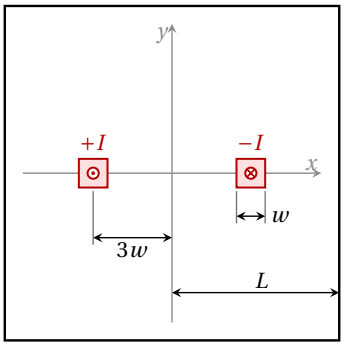


Figure 12.43 The region and the two conductors of Exercise P12.13, not to scale.

P12.14 This exercise continues Example 12.7.

(a) Reproduce the field midway between the bars, $0.149\,\mu_0 I/w$, solving the grid equations as the example does.

(b) Find the force per unit length on each bar, by adding up $\mathbf{J} \times \mathbf{B}$ over its cross section, with **B** from your grid. Compare with the force between two thin wires at the bars' centers, (10.4). Do the bars attract or repel?

(c) During a short circuit the current in bars with $w = 5.0\,\mathrm{mm}$ might reach 20 kA. Find the force on a one-meter length of each bar.

Exercises for 12.6 Multipole expansion of A; the magnetic dipole

Warm-up questions Read Section 12.6, then answer before class

W12.14 The multipole expansion of V can begin with a $1/r$ term, but the expansion of $\mathbf{A}$ never does. What plays the role of the total charge in the expansion of $\mathbf{A}$, for a loop of wire and for a current spread through a volume? Why is it always zero for steady currents?

W12.15 A wire loop runs around the rim of a hemispherical bowl of radius R and carries a current I. One classmate says that the loop's dipole moment has magnitude I times the area of the bowl, $2\pi R^2 I$; another says $\pi R^2 I$. Who is right? Explain with the vector area and Figure 12.27.

W12.16 A wire is bent into a flat figure eight of two equal circles, so that the current goes counterclockwise around one circle and clockwise around the other. What is its magnetic dipole moment? Far away, does its vector potential fall off faster or more slowly than that of a single circular loop, and why?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

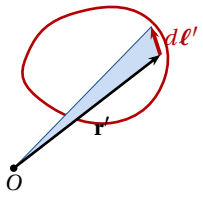


Figure 12.44 Exercise P12.15: the position $\mathbf{r}'$ of a piece $d\boldsymbol{\ell}'$ of a loop, and the thin triangle (shaded) that the two vectors span.

P12.15 This exercise derives formulas stated in Section 12.6.2.

(a) Apply (12.21) with $\mathbf{c} = \hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ in turn, and combine the results to show that $\mathbf{a} = \frac{1}{2}\oint \mathbf{r}' \times d\boldsymbol{\ell}'$ (Figure 12.44). Check it for a circle of radius R in the xy plane.

(b) For a current spread through a volume, the dipole term of (12.18) is $(\mu_0/4\pi r^2)\int(\hat{\mathbf{r}}\cdot\mathbf{r}')\,\mathbf{J}\,d\tau'$. Use the product rule (1.52), as in the proof of (12.20), with $x'_i x'_j \mathbf{J}$ in place of $x'\mathbf{J}$, to show that $\int(x'_i J_j + x'_j J_i)\,d\tau' = 0$, where x'_i and x'_j are any two of x', y', z', or the same one twice. Then show that $\int(\hat{\mathbf{r}}\cdot\mathbf{r}')\,\mathbf{J}\,d\tau' = \mathbf{m}\times\hat{\mathbf{r}}$, with $\mathbf{m}$ given by (12.24), so that the dipole term is (12.22). (Write $\int x'_i J_j\,d\tau'$ as half the sum plus half the difference of $\int x'_i J_j\,d\tau'$ and $\int x'_j J_i\,d\tau'$, and use $(\mathbf{r}'\times\mathbf{J})\times\hat{\mathbf{r}} = (\hat{\mathbf{r}}\cdot\mathbf{r}')\,\mathbf{J} - (\hat{\mathbf{r}}\cdot\mathbf{J})\,\mathbf{r}'$.)

(c) Show that (12.24), applied to a thin loop of wire, gives $\mathbf{m} = I\mathbf{a}$.

(d) A solid sphere of radius R carries a total charge Q spread uniformly through its volume, and spins about the z axis with angular velocity ω, so that $\mathbf{J} = \rho\,\boldsymbol{\omega}\times\mathbf{r}$. Use (12.24) to show that $\mathbf{m} = \frac{1}{5}Q\omega R^2\hat{\mathbf{z}}$, and compare with the spinning shell of Example 12.9.

P12.16 This exercise fills in the steps from the dipole potential to the dipole field.

(a) Take the curl of (12.25) in spherical coordinates to obtain (12.26).

(b) Write $\mathbf{m} = m\hat{\mathbf{z}}$ in terms of $\hat{\mathbf{r}}$ and $\hat{\boldsymbol{\theta}}$, and show that (12.27) agrees with (12.26). Write the electric dipole field (8.15) in the same form.

(c) Show that the dipole field has zero divergence and zero curl everywhere except at the origin.

P12.17 This exercise shows that the field outside the spinning shell of Example 12.9 is exactly a dipole field, using the boundary conditions of Section 12.1.3.

(a) Show that the shell carries the surface current $\mathbf{K} = \sigma\omega R\sin\theta\,\hat{\boldsymbol{\phi}}$.

(b) Show that $\mathbf{A} = Cr\sin\theta\,\hat{\boldsymbol{\phi}}$, with C a constant, gives a uniform field, and that $\mathbf{A} = D\sin\theta\,\hat{\boldsymbol{\phi}}/r^2$ gives a dipole field. Both have zero divergence and zero curl of the curl away from the shell, so they satisfy Ampère's law there.

(c) Use the first form inside the shell and the second outside, and apply (12.5) and (12.6) at $r = R$ to find C and D.

(d) Show that the field outside is exactly the dipole field of the moment found in Example 12.9, and that the field inside is $\frac{2}{3}\mu_0\sigma\omega R\hat{\mathbf{z}}$.

(e) Check the boundary conditions on $\mathbf{B}$, (11.11), at a pole and at the equator.

P12.18 Earth's field at its surface is approximately the field of a magnetic dipole at its center. Take the field at the magnetic equator to be $30\,\mu$T and Earth's radius to be $R_E = 6.371\times10^6$ m.

(a) Find Earth's magnetic dipole moment. A compass needle's north end points roughly toward geographic north; which way does Earth's dipole moment point?

(b) What does the dipole model give for the field at the magnetic poles? Compare with the range of surface fields in the table of Section 10.3.

(c) If the dipole moment were produced by a single loop of current at the radius of Earth's core, 3480 km, what current would it carry?

(d) Find the field at the magnetic equator one Earth radius above the surface.

P12.19 This exercise derives the averages that give the delta-function terms in (12.28) and (12.29). Let a sphere of radius R, centered on the origin, contain all of a distribution of charge or current.

(a) Show that, for any scalar function f and any vector field $\mathbf{F}$, $\int \nabla f \, d\tau = \oint f \, d\mathbf{a}$ and $\int \nabla \times \mathbf{F} \, d\tau = -\oint \mathbf{F} \times d\mathbf{a}$, the integrals being over the ball and its surface. (Apply the divergence theorem to $f\mathbf{c}$ and to $\mathbf{F} \times \mathbf{c}$, where $\mathbf{c}$ is a constant vector.)

(b) On the surface, write V as its multipole expansion. Using $\oint P_n(\hat{\mathbf{r}} \cdot \hat{\mathbf{r}}') \, \hat{\mathbf{r}} \, d\Omega = (4\pi/3) \, \hat{\mathbf{r}}' \, \delta_{n1}$, where $d\Omega$ is the element of solid angle, show that only the dipole term contributes to $\oint V \, d\mathbf{a}$, and that the average of $\mathbf{E}$ over the ball is $-\mathbf{p}/4\pi\varepsilon_0 R^3$.

(c) In the same way, show that only the dipole term of $\mathbf{A}$ contributes to $\oint \mathbf{A} \times d\mathbf{a}$, and that the average of $\mathbf{B}$ over the ball is $(\mu_0/4\pi)(2\mathbf{m}/R^3)$.

(d) Show that the $1/r^3$ terms of (12.28) and (12.29) average to zero over the directions of $\hat{\mathbf{r}}$, and that the delta-function terms give exactly the averages of parts (b) and (c).

Exercises for 12.7 Inductance and the energy stored in magnetic fields

Warm-up questions Read Section 12.7, then answer before class

W12.17 The turns of a long solenoid are wound twice as closely, so that n doubles, while its length, radius, and current stay the same. By what factor does each of these change: the field inside, the flux through one turn, the inductance, and the stored energy? Why does the inductance not simply double?

W12.18 Why would a loop of wire of zero thickness have an infinite self-inductance, although its mutual inductance with a second, separate loop is finite? Answer with the flux close to the wire, and again with the Neumann formula (12.31): what happens to it when both integrals run around the same loop?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

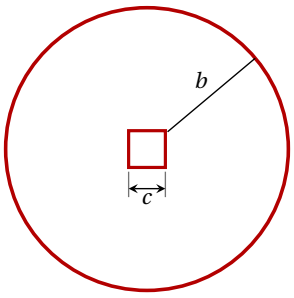


Figure 12.45 The two loops of Exercise P12.20, not to scale.

P12.20 A small square loop with sides c lies at the center of a large circular loop of radius $b \gg c$, in the same plane (Figure 12.45).

(a) Find their mutual inductance, using the field at the center of the large loop.

(b) The symmetry $M_{12} = M_{21}$ says that a current I_s in the small loop sends a flux MI_s through the large one. Check that directly. Treat the small loop as a dipole, and use the fact that the total flux of its field through the whole plane of the loops is zero, so that the flux through the large loop equals the negative of the flux through the part of the plane outside it.

(c) Evaluate M for $b = 10.0\,\text{cm}$ and $c = 1.00\,\text{cm}$.

P12.21 For the coaxial cable of Exercise P12.10:

(a) Use (12.34) and your vector potential to find the energy per unit length, and from it the inductance per unit length. Check the energy with (12.35).

(b) If the inner conductor were a thin tube of radius a instead, with all of its current on its surface, what would the inductance per unit length be? Explain the difference.

(c) Evaluate both for $a = 0.50\,\text{mm}$ and $b = 2.0\,\text{mm}$.

P12.22 The toroid of Exercise P11.10, part (a), has N turns on a core of rectangular cross section, with inner radius a, outer radius b, and height h (Figure 12.46).

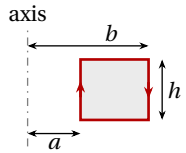


Figure 12.46 The core of the toroid of Exercise P12.22, cut by a plane through its axis. Each turn of the winding runs around the rectangle.

(a) Find its inductance from the flux.

(b) Find the energy stored in its field from (12.35), and check that it equals $\frac{1}{2}LI^2$.

(c) Evaluate L and the energy for the numbers given in Exercise P11.10.

***Exercises for 12.8 Summary: Triangle diagram of* J, B, A**

P12.23 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as the vector potential of a circular loop of current, different gauges for the vector potential of a uniform field, the field lines around long conductors of any cross section found by relaxation, the field of a magnetic dipole compared with the field of a small loop of current, the mutual inductance of two loops of any shape, or the interference of electrons that pass on the two sides of a long solenoid.

Chapter 13

Magnetic Materials



William Gilbert (1544–1603, English) Gilbert was born in Colchester and studied at St John's College, Cambridge, where he earned a doctorate in medicine in 1569. He practiced medicine in London from about 1573, became president of the Royal College of Physicians in 1600, and served as a physician to Queen Elizabeth I and, briefly, to her successor, James I. His book *De Magnete* (1600) reported years of experiments with lodestones. Its central claim was that the Earth itself is a great magnet. He made a small sphere of lodestone, which he called a *terrella*, "little earth," and showed that a small compass needle near its surface lies flat only at its equator and dips toward the nearer pole, as a real compass needle dips more and more as it is carried toward the Earth's poles. In the same book he distinguished the attraction of a magnet from the attraction of rubbed amber, and he called the substances that attract when rubbed *electrica*, from the Greek word for amber, the origin of our word "electric." He died in 1603, probably of the plague. (Wikipedia)

Chapter 9 described what happens to an insulator in an electric field: its atoms and molecules become small electric dipoles, and the material as a whole becomes polarized. This chapter does the same for magnetic fields. Magnetism in materials was known long before electromagnetism was. Lodestones, pieces of a naturally magnetized iron ore, were known to attract iron in antiquity, and the magnetic compass was in use in China and in Europe by the end of the twelfth century. Petrus Peregrinus had described the poles of a lodestone in 1269 (Section 2.4), but the first thorough experimental study of magnetism was William Gilbert's *De Magnete*, published in 1600.

Every material responds to a magnetic field in some way, though in most materials the response is so small that it takes a sensitive experiment to detect it. A few materials, iron above all, respond so strongly that they can stay magnetized after the field is removed, and they are the permanent magnets that Chapter 10 set aside. The plan of the chapter follows Chapter 9 closely. It begins with magnetic dipoles in external fields and with the origin of the magnetic moments of atoms. It then describes a magnetized material by its magnetic dipole moment per unit volume, shows that the field of a magnetized object is the field of a particular distribution of current, called bound current, and develops two ways to find that field. The second uses a new field, $\mathbf{H}$, whose curl depends only on the currents that we put in place ourselves. After a section on the strongly magnetic materials, the chapter ends with boundary conditions, numerical methods, and the energy stored in the field. Almost every equation will have a counterpart in Chapter 9, and almost every one will differ from its counterpart in some small but important way.

13.1 Dipoles again

Section 12.6.3 found the field that a magnetic dipole *produces*. A small loop of wire carrying a current I around the vector area $\mathbf{a}$ has the magnetic dipole moment $\mathbf{m} = I\mathbf{a}$, (12.23), and far from it the vector potential and the field are (12.22) and (12.27). This section is about the other half of the story, as Section 9.1 was for

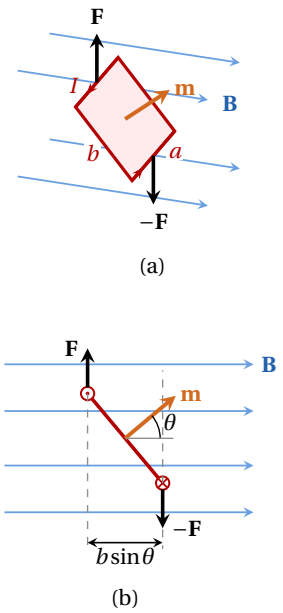


Figure 13.1 A rectangular loop carrying a current I in a uniform field **B**. (a) The forces on the two sides of length a are equal and opposite. (b) Seen along those sides, the two forces form a couple, with lever arm $b\sin\theta$, that turns **m** toward **B**.

electric dipoles: what a field made by other currents does to a magnetic dipole placed in it. Throughout the section, **B** means that external field. As before, a dipole's own field exerts no net force or torque on the dipole itself.

13.1.1 Torque on a dipole

Put a rectangular loop of wire, with sides a and b and a current I, in a uniform field **B** (Figure 13.1). Let the two sides of length a be perpendicular to **B**, and let the loop be free to turn about the axis through its center parallel to those sides. Its dipole moment **m** makes the angle θ with **B**. The force on each side is $I\mathbf{L}\times\mathbf{B}$, (10.3), where **L** is a vector along the side in the direction of the current. The forces on the two sides of length b are equal and opposite, and they act at the midpoints of those sides, which lie on the axis, so they act along the same line, cancel, and exert no torque. The forces on the two sides of length a have the magnitude IaB and are also equal and opposite, so the net force on the loop is zero, as Section 10.4.2 found for any closed loop in a uniform field. But these two forces do not act along the same line. Seen edge on, along the sides of length a (Figure 13.1b), they form a couple: two equal and opposite forces whose lines of action are a distance $b\sin\theta$ apart. The torque of a couple is the same about every point, and its magnitude is the force times that distance,

$$\tau = (IaB)(b\sin\theta) = (Iab)B\sin\theta = mB\sin\theta,$$

since $Iab = m$. The torque turns **m** toward the direction of **B**. A vector with that magnitude and direction is

$$\boldsymbol{\tau} = \mathbf{m}\times\mathbf{B}, \tag{13.1}$$

the magnetic counterpart of $\boldsymbol{\tau} = \mathbf{p}\times\mathbf{E}$, (9.1). The same result holds for a flat or bent loop of any shape in a uniform field (Exercise P13.1), and so for any small distribution of steady current.

This torque is what turns an electric motor, as Section 10.4.2 promised. A coil in the field of a magnet is turned until its moment lines up with the field, and there the torque vanishes. To keep the coil turning, a sliding contact on the shaft, the commutator, reverses the current every half turn, so that the moment is always being pulled around in the same direction.

13.1.2 Energy of a dipole in a field

The energy of a magnetic dipole in a field follows from the torque exactly as it did for an electric dipole in Section 9.1.2. The work needed to turn the dipole slowly from the angle θ_{ref} to θ, against the torque, is $\int_{\theta_{\text{ref}}}^{\theta} mB\sin\theta'\,d\theta'$, and with the reference angle $\theta_{\text{ref}} = \pi/2$ it is $-mB\cos\theta$:

$$U = -\mathbf{m}\cdot\mathbf{B}. \tag{13.2}$$

The energy is lowest when the dipole points along the field and highest when it points the other way. One thing was assumed along the way, and it matters more

here than it did for an electric dipole: the size of **m** stays fixed while the dipole turns. For a loop of wire, that means that the current stays fixed.

Equation (13.2) is therefore the energy that gives the torque and the force on a dipole of fixed moment, and it should be used only for that. It is not the change in the total energy of the system. When a loop with a fixed current turns or moves so that the magnetic force on it does the work W, the energy stored in the magnetic field (Section 12.7.2) does not decrease by W, as it would if $-\mathbf{m}\cdot\mathbf{B}$ were an ordinary potential energy. It increases by W, because the part of the field energy that the loop shares with the other currents is $+\mathbf{m}\cdot\mathbf{B}$, and the batteries that keep the currents fixed supply $2W$. That was the story of the two coaxial loops at the end of Section 12.7.2, and Exercise P19.11 works through it.

☞ *Caution*: $U = -\mathbf{m}\cdot\mathbf{B}$ gives the torque and the force on a dipole of fixed moment. It is not the change in the energy of the whole system: when a current loop moves, the energy stored in the field changes too, and whatever keeps the current fixed does work.

13.1.3 Force on a dipole

In a uniform field the net force on a magnetic dipole is zero. In a nonuniform field it is not, and it can be found from the energy, as $\mathbf{F} = -\nabla U$:

$$\mathbf{F} = \nabla(\mathbf{m}\cdot\mathbf{B}). \tag{13.3}$$

Here **m** is a constant vector, the moment of one particular dipole, and only **B** is differentiated. A dipole that is lined up with the field has $\mathbf{m}\cdot\mathbf{B} = mB$, so it is pulled toward the region where the field is strongest, just as an electric dipole is. A dipole that points against the field is pushed toward the weaker field.

For an electric dipole, Section 9.1.3 found a second form of the force, $(\mathbf{p}\cdot\nabla)\mathbf{E}$, and showed that the two forms agree because $\nabla\times\mathbf{E} = 0$. The magnetic counterpart is not so simple. By the product rule for the gradient of a dot product (Section 1.7), with **m** constant,

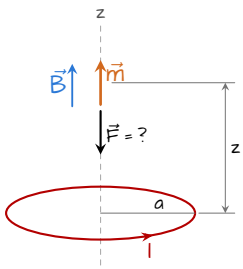
$$\nabla(\mathbf{m}\cdot\mathbf{B}) = (\mathbf{m}\cdot\nabla)\mathbf{B} + \mathbf{m}\times(\nabla\times\mathbf{B}) = (\mathbf{m}\cdot\nabla)\mathbf{B} + \mu_0\,\mathbf{m}\times\mathbf{J},$$

where **J** is the current density of the sources of **B** at the place where the dipole is. So the two forms agree when the dipole is outside the currents that make the field, which is the usual case, and disagree when it is inside them. In that case the gradient form (13.3) is the right one: a direct calculation of the forces on the four sides of a small square loop gives $\nabla(\mathbf{m}\cdot\mathbf{B})$ whether or not the field has a curl (Exercises P13.2 and P13.3).

Example 13.1

A small magnet with dipole moment $\mathbf{m} = m\hat{\mathbf{z}}$ is on the axis of a circular loop of radius a carrying a current I, at a height z above the center of the loop. The current circulates so that the loop's field on its axis points along $+\hat{\mathbf{z}}$. Find the force on the magnet, where it is largest, and its largest value for $a = 5.0\,\text{cm}$, $I = 10\,\text{A}$, and $m = 0.10\,\text{A}\cdot\text{m}^2$.

Solution:



A sketch of the loop and the magnet on its axis shows what is given: the magnet's moment points up, along the loop's field there. The field on the axis is strongest at the center of the loop, so we expect the magnet to be pulled toward the loop.

On the axis the field of the loop is (10.22),

$$B_z(z) = \frac{\mu_0 I a^2}{2\left(z^2+a^2\right)^{3/2}}.$$

Off the axis the field also has a component B_s along $\hat{\mathbf{s}}$, but $\mathbf{m}\cdot\mathbf{B} = mB_z$, and by symmetry the force on a magnet on the axis can only point along the axis. (Why can the force on a magnet on the axis point only along the axis? Answer in footnote.[1]) So

$$F_z = m\frac{\partial B_z}{\partial z} = -\frac{3\mu_0 I a^2 m}{2}\frac{z}{\left(z^2+a^2\right)^{5/2}}.$$

Reality check: For $z>0$ the force is negative: the magnet is pulled toward the loop, as expected. At the center of the loop the force is zero, since the field is largest there.

Setting $dF_z/dz = 0$ gives the largest force at

$$\left(z^2+a^2\right) - 5z^2 = 0, \qquad z = \frac{a}{2},$$

where

$$|F_z| = \frac{3\mu_0 I a^2 m}{2}\frac{a/2}{\left(5a^2/4\right)^{5/2}} = 0.429\,\frac{\mu_0 I m}{a^2} = 2.2\times10^{-4}\,\mathrm{N},$$

about the weight of 22 mg. The magnet is outside the wire, where there is no current, so $(\mathbf{m}\cdot\nabla)\mathbf{B} = m\,\partial\mathbf{B}/\partial z$ gives the same force. If the magnet is turned over, so that it points against the field, the force reverses and the magnet is pushed away from the loop; but then the torque on it, (13.1), is zero only because the moment is exactly antiparallel, and the slightest tilt makes it turn toward the direction of the field.

[1] Turning the loop and the magnet together about the z axis changes nothing, so it cannot change the force. A force with a part perpendicular to the axis would turn with them, so that part must be zero.

13.1.4 Where atomic moments come from

In Chapter 9 the dipole moments of atoms and molecules were made by the field, by shifting their charges, or else they were built into polar molecules. The magnetic moments of atoms have two sources, and both are tied to angular momentum.

The first is the orbital motion of the electrons. A classical picture gives the idea (Figure 13.2). An electron of charge $-e$ moving in a circle of radius r with speed v goes around once in the time $2\pi r/v$, so it is a current $I = ev/2\pi r$ (circulating opposite to the electron's motion) around the area πr^2. Its magnetic moment has the size $I\pi r^2 = evr/2$, and since the electron's angular momentum is $L = m_e vr$, the moment is

$$\mathbf{m} = -\frac{e}{2m_e}\mathbf{L}, \tag{13.4}$$

opposite to **L** because the electron's charge is negative. Quantum mechanics changes the picture, since an electron in an atom does not move along a definite orbit, but it keeps this relation. Angular momentum comes in units of $\hbar$, and an orbital angular momentum of $\hbar$ goes with a magnetic moment of

$$\mu_B = \frac{e\hbar}{2m_e} = 9.274 \times 10^{-24}\,\mathrm{A\cdot m^2}, \tag{13.5}$$

the **Bohr magneton**, the natural unit of atomic magnetic moments.

The second source is the electron's **spin**. Every electron has an intrinsic angular momentum of $\hbar/2$, and with it an intrinsic magnetic moment, almost exactly one Bohr magneton ($1.00116\,\mu_B$), twice what (13.4) would give for an angular momentum of $\hbar/2$. Nothing in an electron is known to circulate; the spin moment is simply a property of the particle, like its charge and its mass. Nuclei have magnetic moments too, but they are about a thousand times smaller, because the masses in the denominator of (13.5) are about two thousand times larger, and they matter only in special experiments such as magnetic resonance imaging.

In most molecules and solids the electrons pair up, two by two, with opposite spins and opposite orbital motions, and their moments cancel. A net moment is left only when some electrons are unpaired, as they are in atoms and ions with partly filled inner shells, such as iron, cobalt, nickel, manganese, and the rare earths, and in a few molecules, such as oxygen. Such atoms and ions have moments of a few Bohr magnetons.

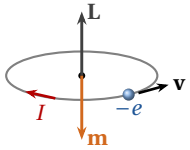


Figure 13.2 A classical picture of an orbiting electron. Its charge is negative, so the current circulates opposite to its motion, and its magnetic moment **m** points opposite to its angular momentum **L**.



Niels Bohr (1885–1962, Danish) Bohr received his doctorate from the University of Copenhagen in 1911 for a thesis on the electron theory of metals. He then worked with J. J. Thomson at Cambridge and, from 1912, with Ernest Rutherford at Manchester. In 1913 he proposed a model of the atom in which the electron moves only in certain allowed orbits and emits or absorbs light when it jumps from one to another; it explained the spectrum of hydrogen. The Bohr radius and the Bohr magneton both come from the smallest of these orbits. He became a professor at Copenhagen in 1916, headed its Institute for Theoretical Physics from 1920 until his death, and received the Nobel Prize in Physics in 1922. In 1939, with John Wheeler, he explained nuclear fission with the liquid-drop model of the nucleus. In 1943 he escaped from occupied Denmark to Sweden, and he spent the rest of the war in Britain and the United States. Afterward he worked for the peaceful use of atomic energy. (Wikipedia)

13.2 Magnetic materials: magnetization and susceptibility

When a dielectric is placed in an electric field, its polarization reduces the field inside it (Section 9.2). A magnetic material in a magnetic field can do either. In some materials the field inside is a little stronger than the applied field, in others a little weaker, and in a few it is enormously stronger. Part of the difference from the electric case comes from the field inside a dipole. Between the two charges

of an electric dipole the field points from the positive charge to the negative one, *opposite* to **p**; inside a current loop the field points *along* **m** (Section 12.6.4). So when the dipoles of a material line up with the applied field, electric dipoles weaken the field inside the material, on average, and magnetic dipoles strengthen it. In some materials, though, the field induces moments that point against it, and those weaken the field (Section 13.2.2). Section 13.6.4 makes this precise.

13.2.1 Magnetization

As for polarization, we describe the dipoles of a material by their density. The **magnetization M** is the magnetic dipole moment per unit volume:

$$\mathbf{M} = \frac{\text{magnetic dipole moment of a small volume}}{\text{volume}}. \tag{13.6}$$



Pierre Curie and Marie Skłodowska-Curie (1859–1906, French; 1867–1934, Polish and French) Pierre Curie and his brother Jacques discovered piezoelectricity in 1880. In his doctoral thesis of 1895, Pierre studied the magnetism of many substances over a wide range of temperatures; it contains the law that bears his name and the temperature above which a ferromagnet loses its ferromagnetism, now called the Curie temperature. Marie Skłodowska came to Paris from Warsaw in 1891 to study physics and mathematics. Her first research project, on the magnetic properties of steels, led to her being introduced to Pierre, an expert on magnetism, and they married in 1895. Together they discovered polonium and radium in 1898, and in 1903 they shared the Nobel Prize in Physics with Henri Becquerel. Pierre was killed in a street accident in 1906. Marie succeeded to his professorship, becoming the Sorbonne's first woman professor, and won a second Nobel Prize, in Chemistry, in 1911. She died in 1934 of aplastic anemia, most likely caused by her long exposure to radiation. (Wikipedia: Pierre, Marie)

If the material has N atoms or molecules per unit volume, each with the same average moment **m**, then $\mathbf{M} = N\mathbf{m}$, and the total moment of an object is

$$\mathbf{m}_{\text{total}} = \int \mathbf{M}\, d\tau. \tag{13.7}$$

The SI unit of **M** is that of a magnetic moment divided by a volume, $\text{A}\cdot\text{m}^2/\text{m}^3 = \text{A/m}$, the unit of a surface current density. Section 13.3 shows that this, too, is not a coincidence.

13.2.2 How a material becomes magnetized

There are three main kinds of response, and a material can show more than one of them.

Paramagnetism. In a material whose atoms, ions, or molecules have net moments, such as oxygen and most compounds of iron, manganese, copper, and the rare earths, the moments point in random directions and the magnetization is zero, until a field is applied. A field exerts a torque on each moment, (13.1), which tends to line it up with the field. Thermal motion works against the lining up, and at room temperature it is far stronger: even in a field of 1 T, the energy $\mu_B B$ is only about 0.2 percent of the thermal energy kT. So only a slight majority of the moments point along the field, and the magnetization is small, proportional to the field, and parallel to it. Paramagnets are weakly attracted toward regions of strong field. Since thermal motion is what opposes the alignment, the magnetization depends on temperature as well:

$$\mathbf{M} \propto \frac{\mathbf{B}}{T}, \tag{13.8}$$

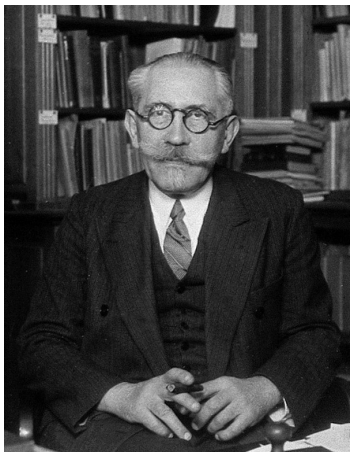
which is **Curie's law**, found by Pierre Curie in 1895. Ten years later Paul Langevin explained it, by applying the statistical mechanics of thermal motion to atomic moments (Exercise P13.4).

Diamagnetism. When a field is applied to any atom, the orbital motion of its electrons changes slightly. While the field is being turned on, the changing magnetic field induces an electric field (this is Faraday's law, the subject of Chapter 15), which speeds up some electrons and slows down others, and the result is a change in the orbital moments that *opposes* the applied field. For an electron in an orbit of radius r perpendicular to the field, the change is

$$\Delta\mathbf{m} = -\frac{e^2 r^2}{4m_e}\mathbf{B} \tag{13.9}$$

(Exercise P13.5). An atom has no resistance, so the change does not die away; it lasts as long as the field is on. It is very small: for $r = a_0$, the Bohr radius, and $B = 1\,\mathrm{T}$, it is about two millionths of a Bohr magneton. Diamagnetism happens in every material, but in materials with unpaired electrons it is usually much smaller than the paramagnetism. In materials whose electrons are all paired, such as water, the noble gases, and most organic compounds, it is all there is: the magnetization is small and points opposite to the field, and the material is weakly repelled from regions of strong field. (Bismuth, the most strongly diamagnetic metal, owes its large susceptibility mostly to its conduction electrons, an effect that needs the quantum theory of solids to explain.)

Ferromagnetism. In iron, cobalt, nickel, gadolinium (below about 19°C), and many alloys and compounds, something else happens. The moments of neighboring atoms line up with one another spontaneously, even with no applied field, over regions called **domains**. A piece of iron that is not magnetized is a patchwork of domains whose magnetizations point in different directions, so that on average they cancel. An applied field lines the domains up, and the magnetization becomes enormous: in a field of 1 T, iron is magnetized about a hundred thousand times more strongly than aluminum, a paramagnet. When the field is removed, some of the alignment can remain, and the iron has become a permanent magnet. Section 13.8 takes up ferromagnets in detail. They are not linear, and most of what follows in Sections 13.5 through 13.7 about linear materials does not apply to them.

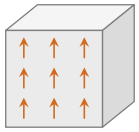


Paul Langevin (1872–1946, French) Langevin studied at the municipal school of physics and chemistry in Paris, where Pierre Curie supervised his laboratory work, and spent a year at the Cavendish Laboratory in Cambridge with J. J. Thomson. His doctoral thesis (1902) was on ionized gases, and he held a chair at the Collège de France from 1909. In 1905 he explained paramagnetism and diamagnetism with the electron theory of atoms and the statistical mechanics of thermal motion. During the First World War he developed a way of detecting submarines with ultrasound, using the piezoelectric effect discovered by the Curie brothers, an early form of sonar. A committed antifascist, he was arrested by the Germans in 1940 and held under house arrest. His son-in-law, the physicist Jacques Solomon, was executed in 1942, and his daughter Hélène was sent to Auschwitz, which she survived. Langevin escaped to Switzerland in 1944. He died in Paris in 1946, and in 1948 his ashes were placed in the Panthéon. (Wikipedia)

Example 13.2

When all of the domains of a piece of iron are lined up, its magnetization reaches its largest possible value, the **saturation magnetization** M_s, which near absolute zero is $1.74 \times 10^6\,\mathrm{A/m}$. The density of iron is $7.87\,\mathrm{g/cm^3}$ and its molar mass is $55.85\,\mathrm{g/mol}$. Find the magnetic moment per atom, in Bohr magnetons.

Solution:



Even a number problem goes better with a sketch of what is given: a cube of iron with every atomic moment lined up. The magnetization is the moment of all the atoms in the cube divided by its volume, so we need the number of atoms per unit volume.

The number of atoms per unit volume is the density divided by the mass of one atom:

$$N = \frac{(7.87\times10^3\,\mathrm{kg/m^3})\,(6.022\times10^{23}\,\mathrm{mol^{-1}})}{0.05585\,\mathrm{kg/mol}} = 8.49\times10^{28}\,\mathrm{m^{-3}}.$$

With $M = Nm$,

$$m = \frac{M_s}{N} = \frac{1.74\times10^6\,\mathrm{A/m}}{8.49\times10^{28}\,\mathrm{m^{-3}}} = 2.05\times10^{-23}\,\mathrm{A\cdot m^2} = 2.21\,\mu_B.$$

(Why does dividing M_s by N give the moment of a single atom? Answer in footnote.[2]) Each iron atom contributes the moment of about two electron spins. It is not a whole number of Bohr magnetons because in a metal some of the electrons are shared among the atoms rather than belonging to any one of them.

Reality check: The value of M_s is from Kittel's table of ferromagnetic crystals, part of which is in Section 13.8.3, and Kittel lists 2.22 Bohr magnetons per atom for iron, close to the value found here.

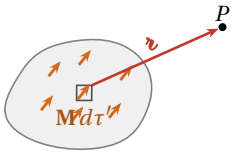


Figure 13.3 A magnetized object. Each small piece of it, of volume $d\tau'$, is a magnetic dipole of moment $\mathbf{M}\,d\tau'$, a distance $\mathscr{r}$ from the field point P.

13.3 Bound currents

Suppose we know the magnetization $\mathbf{M}(\mathbf{r}')$ throughout some object. What field does the object produce? Each small piece of it, of volume $d\tau'$ at $\mathbf{r}'$, is a magnetic dipole with moment $\mathbf{M}\,d\tau'$ (Figure 13.3). The vector potential of a dipole at the origin is (12.22); for a dipole at $\mathbf{r}'$, $\mathbf{r}$ and r are replaced by $\boldsymbol{\mathscr{r}} = \mathbf{r} - \mathbf{r}'$ and $\mathscr{r}$. Adding up the pieces,

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi}\int \frac{\mathbf{M}(\mathbf{r}')\times\hat{\boldsymbol{\mathscr{r}}}}{\mathscr{r}^2}\,d\tau'. \tag{13.10}$$

As in Section 9.3, this integral can be rewritten in a much more useful form. By (9.11), $\hat{\boldsymbol{\mathscr{r}}}/\mathscr{r}^2 = \nabla'(1/\mathscr{r})$, so the integrand is $\mathbf{M}\times\nabla'(1/\mathscr{r})$. The product rule for the curl of a function times a vector (Section 1.7), $\nabla'\times(f\mathbf{M}) = f\,\nabla'\times\mathbf{M} - \mathbf{M}\times\nabla' f$, with $f = 1/\mathscr{r}$, gives

$$\mathbf{M}\times\nabla'\left(\frac{1}{\mathscr{r}}\right) = \frac{1}{\mathscr{r}}\nabla'\times\mathbf{M} - \nabla'\times\left(\frac{\mathbf{M}}{\mathscr{r}}\right),$$

[2] In $M = Nm$, m is the average moment of an atom along $\mathbf{M}$. At saturation near absolute zero, all the domains are lined up and thermal motion does not disturb the moments within them (Section 13.8.3), so every atom's full moment points along $\mathbf{M}$.

so

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi}\int_{\mathcal{V}} \frac{\nabla' \times \mathbf{M}}{\mathscr{r}}\, d\tau' - \frac{\mu_0}{4\pi}\int_{\mathcal{V}} \nabla' \times \left(\frac{\mathbf{M}}{\mathscr{r}}\right) d\tau',$$

where $\mathcal{V}$ is the volume of the object. The second integral is the volume integral of a curl, and a close relative of the divergence theorem turns it into a surface integral:

$$\int_{\mathcal{V}} \nabla \times \mathbf{v}\, d\tau = -\oint_{\mathcal{S}} \mathbf{v} \times d\mathbf{a} \tag{13.11}$$

for any vector field $\mathbf{v}$, where $\mathcal{S}$ is the surface of $\mathcal{V}$ (the derivation is in the box below). With $\mathbf{v} = \mathbf{M}/\mathscr{r}$ and $d\mathbf{a}' = \hat{\mathbf{n}}\, da'$, $\hat{\mathbf{n}}$ the outward unit normal,

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi}\int_{\mathcal{V}} \frac{\nabla' \times \mathbf{M}}{\mathscr{r}}\, d\tau' + \frac{\mu_0}{4\pi}\oint_{\mathcal{S}} \frac{\mathbf{M} \times \hat{\mathbf{n}}}{\mathscr{r}}\, da'. \tag{13.12}$$

The volume integral of a curl

Let $\mathbf{c}$ be any constant vector. By the product rule for the divergence of a cross product (Section 1.7), $\nabla \cdot (\mathbf{v} \times \mathbf{c}) = \mathbf{c} \cdot (\nabla \times \mathbf{v}) - \mathbf{v} \cdot (\nabla \times \mathbf{c}) = \mathbf{c} \cdot (\nabla \times \mathbf{v})$, since $\mathbf{c}$ is constant. Then, by the divergence theorem, (1.36),

$$\mathbf{c} \cdot \int_{\mathcal{V}} \nabla \times \mathbf{v}\, d\tau = \int_{\mathcal{V}} \nabla \cdot (\mathbf{v} \times \mathbf{c})\, d\tau = \oint_{\mathcal{S}} (\mathbf{v} \times \mathbf{c}) \cdot d\mathbf{a} = -\mathbf{c} \cdot \oint_{\mathcal{S}} \mathbf{v} \times d\mathbf{a},$$

where the last step rearranges the triple product, $(\mathbf{v} \times \mathbf{c}) \cdot d\mathbf{a} = \mathbf{c} \cdot (d\mathbf{a} \times \mathbf{v}) = -\mathbf{c} \cdot (\mathbf{v} \times d\mathbf{a})$. This holds for every $\mathbf{c}$, so the two vectors dotted with $\mathbf{c}$ are equal, which is (13.11).

Compare (13.12) with the vector potentials of a volume current and a surface current, (12.9) and (12.11). The first term is the vector potential of the volume current density $\nabla \times \mathbf{M}$, and the second is that of the surface current density $\mathbf{M} \times \hat{\mathbf{n}}$. A magnetized object produces exactly the vector potential, and therefore exactly the field, of the **bound currents**

$$\mathbf{J}_b = \nabla \times \mathbf{M} \tag{13.13}$$

throughout its volume and

$$\mathbf{K}_b = \mathbf{M} \times \hat{\mathbf{n}} \tag{13.14}$$

on its surface. This is why $\mathbf{M}$ has the units of a surface current density. Notice the order of the cross product in (13.14): $\mathbf{M}$ first, then $\hat{\mathbf{n}}$. (The bound charges of Section 9.3 came from a divergence and a dot product, $\rho_b = -\nabla \cdot \mathbf{P}$ and $\sigma_b = \mathbf{P} \cdot \hat{\mathbf{n}}$; the bound currents come from a curl and a cross product, and with no minus sign.) Here too the two formulas are really one: $\mathbf{M}$ stops abruptly at the surface, so $\nabla \times \mathbf{M}$ has a delta function there, of strength $\mathbf{M} \times \hat{\mathbf{n}}$ (Exercise P1.21).

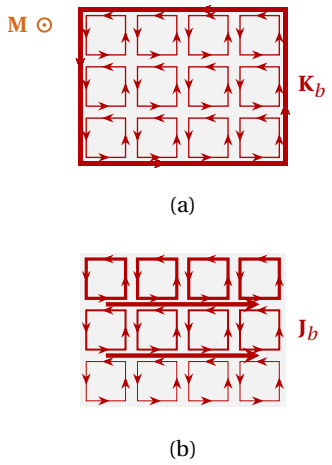


Figure 13.4 Bound currents, with each atomic moment drawn as a small current loop. (a) Uniform magnetization, out of the page. Where two loops touch, their currents cancel; only the outer edges are left, as a current around the surface. (b) Magnetization out of the page that grows toward the top. Where two loops touch, the upper one's current is larger, so a net current is left inside the material, running to the right.

13.3.1 What the bound currents are

Figure 13.4 shows where the bound currents come from, in a picture in which each atomic moment is a tiny loop of current. If the magnetization is uniform (Figure 13.4a), neighboring loops carry the same current, and wherever two of them touch, their currents run in opposite directions and cancel. Inside the material, then, there is no net current, consistent with $\nabla \times \mathbf{M} = 0$ for a constant $\mathbf{M}$. At the surface, though, the outer sides of the outermost loops have no neighbors to cancel them, and together they make a current that runs around the surface: perpendicular to $\mathbf{M}$, and parallel to the surface. That is $\mathbf{K}_b = \mathbf{M} \times \hat{\mathbf{n}}$. The size is right, too. A cylinder of length ℓ and cross-sectional area A, magnetized along its axis, has the dipole moment $MA\ell$; the surface current $K_b = M$ around its wrapper is a current $I = K_b\ell = M\ell$ around the area A, which has the same moment, $(M\ell)A$.

If the magnetization is not uniform, the cancellation inside is not complete. In Figure 13.4b the magnetization points out of the page and grows toward the top of the figure. Each loop's current runs counterclockwise, so where two loops touch, the bottom side of the upper loop carries its current to the right and the top side of the lower loop carries its smaller current to the left, and a net current to the right remains. In terms of $\mathbf{M}$, with $\mathbf{M} = M_z(y)\,\hat{\mathbf{z}}$, $\nabla \times \mathbf{M} = (\partial M_z/\partial y)\,\hat{\mathbf{x}}$, which points to the right when M_z grows with y.

These currents are as real as the bound charges of Chapter 9: they produce exactly the field that the same free currents would. No electron travels around the object to carry them. Each stays with its own atom, and the current at the surface is what is left over when the circulations of neighboring atoms fail to cancel. And many atomic moments, those of electron spins, involve no circulating charge at all. That makes no difference to the field: a spin moment produces the same dipole field as a small current loop, so a magnetized object produces the field of its bound currents whatever the origin of its magnetization.

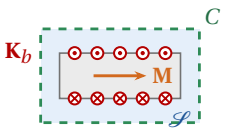


Figure 13.5 A magnetized bar seen from the side, with its bound surface current coming out of the page along the top and going in along the bottom. The loop C (dashed) lies outside the bar; the bar passes through the surface $\mathscr{S}$ (shaded) that C bounds, and the bound current crosses $\mathscr{S}$ once each way.

13.3.2 The total bound current

The total bound charge of a polarized object is zero (Section 9.3.1). The magnetic counterpart is a statement about currents passing through a surface. Let C be any closed loop that lies entirely *outside* the magnetized material, and let $\mathscr{S}$ be any surface bounded by C; the surface may pass through the material. Then the net bound current through $\mathscr{S}$ is zero (Figure 13.5).

The proof is short if we think of the surface current $\mathbf{K}_b$ as the limit of a volume current $\mathbf{J}_b$ in a thin layer at the surface, across which the magnetization drops rapidly, but smoothly, from its value inside to zero. Then all of the bound current is $\nabla \times \mathbf{M}$, and by the curl theorem, (1.37),

$$I_{b,\text{through}} = \int_{\mathscr{S}} (\nabla \times \mathbf{M}) \cdot d\mathbf{a} = \oint_C \mathbf{M} \cdot d\boldsymbol{\ell} = 0, \tag{13.15}$$

because $\mathbf{M}$ is zero everywhere on C. In practice this means that the bound volume current and the bound surface current through any such surface always cancel.

Example 13.5 below shows it happening. Ampère's law then says something useful at once: the circulation of **B** around a loop that lies outside all the magnetized material, and that no other current passes through, is zero.

13.4 Solving for B when given M: bound-current method

Equations (13.13) and (13.14) turn the problem of finding the field of a magnetized object into a problem that Chapters 10–12 have already solved. The **bound-current method** has two steps:

1. Find the bound currents, $\mathbf{K}_b = \mathbf{M} \times \hat{\mathbf{n}}$ on every surface and $\mathbf{J}_b = \nabla \times \mathbf{M}$ inside.
2. Find the field of those currents by any method that suits them: the Biot–Savart law (Section 10.6), Ampère's law if the symmetry allows it, the vector potential, superposition of known results, the dipole approximation far away, or numerical integration.

As in Section 9.4, the magnetization here is taken as given, as it would be in a permanent magnet. If the magnetization was caused by an applied field, the field found by this method is only the field of the magnetized object, and the applied field must be added to it.

Example 13.3

A cylindrical bar magnet of radius R and length L, centered on the origin with its axis along z, is uniformly magnetized along its axis, $\mathbf{M} = M_0 \hat{\mathbf{z}}$. Find the field on its axis, and compare it far away with the field of a dipole.

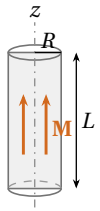
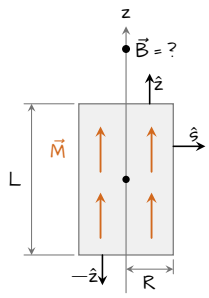


Figure 13.6 The bar magnet of Example 13.3.

Solution:



We start by drawing the magnet in cross section, with **M** along its axis, and marking the field point on the axis. The bound currents need the outward normal on each surface: $\hat{\mathbf{z}}$ on the top, $-\hat{\mathbf{z}}$ on the bottom, and $\hat{\mathbf{s}}$ on the wrapper.

Step 1. Since **M** is constant, $\mathbf{J}_b = 0$. On the top and bottom faces, $\hat{\mathbf{n}} = \pm\hat{\mathbf{z}}$ is parallel to **M**, so $\mathbf{K}_b = 0$. On the wrapper, $\hat{\mathbf{n}} = \hat{\mathbf{s}}$ and

$$\mathbf{K}_b = M_0 \hat{\mathbf{z}} \times \hat{\mathbf{s}} = M_0 \hat{\boldsymbol{\phi}}.$$

The magnet's bound current runs around its wrapper, like the current in the windings of a solenoid.

Step 2. A solenoid with n turns per unit length carrying a current I has the surface current density $nI\hat{\boldsymbol{\phi}}$. (Why is the surface current density of a solenoid nI? Answer in footnote.[3]) The magnet's field is therefore the field of a solenoid of the same size, with nI replaced by M_0, and Example 10.8 found that field on the axis:

$$\mathbf{B} = \frac{\mu_0 M_0}{2}\left[\frac{L/2-z}{\sqrt{R^2+(L/2-z)^2}} + \frac{L/2+z}{\sqrt{R^2+(L/2+z)^2}}\right]\hat{\mathbf{z}}.$$

Figure 10.18 shows it, in units of $\mu_0 M_0$. At the center the field is $\mu_0 M_0 L/\sqrt{L^2+4R^2}$, close to $\mu_0 M_0$ for a long magnet; at the ends it is close to half that.

Reality check: Far away, for z much larger than L and R, a binomial expansion of the two terms gives

$$\mathbf{B} \approx \frac{\mu_0 M_0 R^2 L}{2z^3}\hat{\mathbf{z}} = \frac{\mu_0}{2\pi}\frac{\mathbf{m}}{z^3},$$

with $\mathbf{m} = \pi R^2 L M_0 \hat{\mathbf{z}}$, the magnet's total moment, (13.7). This is the field of a dipole on its axis (Section 12.6.3).

How far away is "far"? It depends on the shape. For a magnet whose length equals its diameter, $L = 2R$, the dipole formula is within 5 percent of the exact field beyond about $z = 1.9R$, less than one length from the center; for a magnet five times as long as its radius, $L = 5R$, it is within 5 percent only beyond about $z = 15R$, three lengths away (Exercise P13.7).

Example 13.4

A sphere of radius R is uniformly magnetized, $\mathbf{M} = M_0\hat{\mathbf{z}}$. Find its field inside and outside.

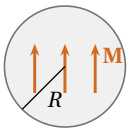
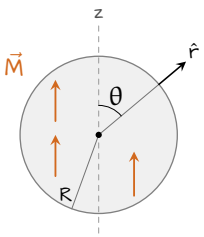


Figure 13.7 The uniformly magnetized sphere of Example 13.4.

Solution:



A cross section of the sphere shows what is given: $\mathbf{M}$ is the same everywhere inside, pointing up. For the bound surface current we will need the outward normal $\hat{\mathbf{r}}$ at the surface, at the angle θ from $\mathbf{M}$.

Step 1. The magnetization is constant, so $\mathbf{J}_b = 0$. On the surface $\hat{\mathbf{n}} = \hat{\mathbf{r}}$, and

$$\mathbf{K}_b = M_0\hat{\mathbf{z}}\times\hat{\mathbf{r}} = M_0\sin\theta\,\hat{\boldsymbol{\phi}}.$$

[3] A length ℓ of the solenoid holds $n\ell$ turns, each carrying the current I around it, so the current around it per unit length is $n\ell I/\ell = nI$.

The bound current circles the z axis, strongest at the equator and zero at the poles.

Step 2. This is exactly the surface current of the spinning charged shell of Example 12.9, $\sigma\omega R\sin\theta\,\hat{\boldsymbol{\phi}}$ (Exercise P12.17), with $\sigma\omega R$ replaced by M_0. (Why does the magnetized sphere have the same field as the spinning shell? Answer in footnote.[4]) That shell's field outside is exactly the field of a dipole, and its field inside is uniform, $\frac{2}{3}\mu_0\sigma\omega R\hat{\mathbf{z}}$. So the magnetized sphere's field inside is

$$\mathbf{B}_{\text{in}} = \frac{2}{3}\mu_0\mathbf{M}, \tag{13.16}$$

and outside it is the dipole field (12.27) of the moment $\frac{4}{3}\pi\sigma\omega R^4\hat{\mathbf{z}} \to \frac{4}{3}\pi R^3\mathbf{M}$, exact everywhere outside the sphere.

Reality check: That moment is the sphere's volume times its magnetization, the total moment that (13.7) gives.

Compare the uniformly polarized sphere of Example 9.3, whose field inside is $\mathbf{E} = -\mathbf{P}/3\varepsilon_0$. The two spheres have exactly the same field outside, a dipole field, with $\mathbf{p}/\varepsilon_0$ in place of $\mu_0\mathbf{m}$. Inside, the electric field points *against* the polarization and the magnetic field points *along* the magnetization, as the fields inside the individual dipoles do (Section 13.2). The field lines of $\mathbf{B}$ are drawn in Figure 13.11a, in Section 13.6.1: they pass straight up through the sphere and close on themselves outside.

Example 13.5

A long cylinder of radius R has the magnetization

$$\mathbf{M} = M_0\frac{s^2}{R^2}\hat{\boldsymbol{\phi}},$$

which circles the axis and grows from zero on the axis to M_0 at the surface. Find the bound currents and the field inside and outside.

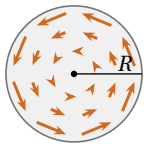
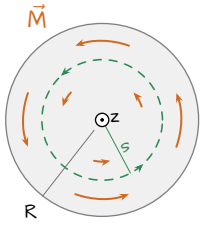


Figure 13.8 The cylinder of Example 13.5, end on: $\mathbf{M}$ circles the axis and grows as s^2.

[4] The Biot–Savart law gives the field from the current density alone, whatever carries the current. The two surface currents are the same function of position, so their fields are the same, inside and outside (Section 13.3.1).

Solution:



We sketch a cross section of the cylinder with arrows for $\mathbf{M}$, circling the axis and longer farther out, and ask what bound currents they leave behind. The currents will run along the axis, so Ampère's law with a circular loop, drawn here, should give the field.

Step 1. By the curl in cylindrical coordinates, (1.27), with only $M_\phi(s)$ nonzero,

$$\begin{aligned}\mathbf{J}_b = \nabla \times \mathbf{M} &= \cancel{\left(\frac{1}{s}\frac{\partial M_z}{\partial \phi} - \frac{\partial M_\phi}{\partial z}\right)}\hat{\mathbf{s}} + \cancel{\left(\frac{\partial M_s}{\partial z} - \frac{\partial M_z}{\partial s}\right)}\hat{\boldsymbol{\phi}} \\ &\quad + \frac{1}{s}\left(\frac{\partial (sM_\phi)}{\partial s} - \cancel{\frac{\partial M_s}{\partial \phi}}\right)\hat{\mathbf{z}} \\ &= \frac{1}{s}\frac{\partial}{\partial s}\left(\frac{M_0 s^3}{R^2}\right)\hat{\mathbf{z}} \\ &= \frac{3M_0 s}{R^2}\hat{\mathbf{z}},\end{aligned}$$

up the cylinder. On the surface, $\hat{\mathbf{n}} = \hat{\mathbf{s}}$, and

$$\mathbf{K}_b = M_0\,\hat{\boldsymbol{\phi}} \times \hat{\mathbf{s}} = -M_0\,\hat{\mathbf{z}},$$

down the cylinder.

Reality check: The total bound current up the cylinder is

$$\int_0^R \frac{3M_0 s}{R^2}\,2\pi s\,ds - M_0(2\pi R) = 2\pi M_0 R - 2\pi M_0 R = 0,$$

as (13.15) requires. (Why must the total bound current along the cylinder be zero? Answer in footnote.[5])

Step 2. The currents run along the axis and depend only on s, as in a long wire, so by the symmetry argument of Example 11.1 the field circles the axis, $\mathbf{B} = B(s)\,\hat{\boldsymbol{\phi}}$, and Ampère's law with a circular loop of radius s gives it. Inside ($s < R$) the loop encloses the bound current

$$I_{\text{enc}} = \int_0^s \frac{3M_0 s'}{R^2}\,2\pi s'\,ds' = \frac{2\pi M_0 s^3}{R^2},$$

so

$$B \cdot 2\pi s = \mu_0 \frac{2\pi M_0 s^3}{R^2}, \qquad \mathbf{B} = \mu_0 M_0 \frac{s^2}{R^2}\,\hat{\boldsymbol{\phi}}.$$

Outside, the loop encloses no net current, so $\mathbf{B} = 0$.

The answer has a curious form: inside, $\mathbf{B} = \mu_0\mathbf{M}$ exactly. That is not a coincidence, and Section 13.7 will find the same answer in two lines.

[5] In the picture of Figure 13.4, each atomic moment is a small loop of current. A cross section of the cylinder that cuts such a loop cuts it twice, once where its current runs up and once where it runs back down, so the bound currents through the cross section add up to zero.

13.5 The H field and permeability

In most problems with magnetic materials the magnetization is not given. What is given, or what we control, is the free current: the current in the windings of a solenoid, or in a wire. The magnetization then depends on the field, and the field depends on the magnetization, through the bound current it produces. Ampère's law, $\nabla \times \mathbf{B} = \mu_0 \mathbf{J}$, holds as always, but the $\mathbf{J}$ in it is the total current density (Figure 13.9),

$$\mathbf{J} = \mathbf{J}_f + \mathbf{J}_b, \tag{13.17}$$

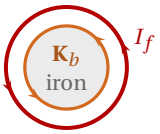


Figure 13.9 A coil wound on an iron rod, end on. The current in the wire is free current, I_f; the magnetized rod carries a bound current $\mathbf{K}_b$ on its surface, in the same sense.

where $\mathbf{J}_f$ is the **free current**: every current that is not bound, such as the current in a wire, which consists of charges that travel all the way around the circuit. The bound part is not known until the problem is solved.

As with $\mathbf{D}$ in Section 9.5, we look for a field whose curl comes from the free currents alone, and find it by moving the bound current to the other side of Ampère's law. The result, which Section 13.6 derives, is the field

$$\mathbf{H} = \frac{\mathbf{B}}{\mu_0} - \mathbf{M}. \tag{13.18}$$

It has no settled name of its own. Physicists usually call it simply the **H field**, or the **magnetic field strength**, and keep "magnetic field" for $\mathbf{B}$; some engineering books call $\mathbf{H}$ the magnetic field and $\mathbf{B}$ the magnetic flux density. In this book the magnetic field is always $\mathbf{B}$. Notice the minus sign, where the definition of $\mathbf{D}$, (9.21), has a plus sign. That is not a mistake, and Section 13.6 shows why the minus sign is the one that makes the curl of $\mathbf{H}$ come out right; Chapter 14 has more to say about it. The factor $1/\mu_0$ makes the units agree: $\mathbf{H}$ has the units of $\mathbf{M}$, amperes per meter. In Gaussian units the unit of $\mathbf{H}$ is the **oersted**, which corresponds to $1000/4\pi = 79.6\,\text{A/m}$. In vacuum, where $\mathbf{M} = 0$, $\mathbf{H}$ is just $\mathbf{B}/\mu_0$, and a field of one gauss in vacuum is a field $\mathbf{H}$ of one oersted.

☞ *Caution*: $\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P}$, but $\mathbf{H} = \mathbf{B}/\mu_0 - \mathbf{M}$. The sign of the magnetization term is different.

Like the definition of $\mathbf{D}$, (13.18) is a definition, so it holds in every material, linear or not, magnetized by an applied field or permanently.

13.5.1 Linear magnetic materials

In paramagnets and diamagnets, as long as the field is not too strong, the magnetization is proportional to the field. By convention the proportionality is written with $\mathbf{H}$:

$$\mathbf{M} = \chi_m \mathbf{H}. \tag{13.19}$$

The dimensionless constant χ_m is the **magnetic susceptibility** of the material, positive for paramagnets and negative for diamagnets. A material that obeys (13.19) is a **linear magnetic material**, and as with dielectrics, "linear" in this book also means isotropic and homogeneous.

The closer analog of $\mathbf{P} = \varepsilon_0 \chi_e \mathbf{E}$ would have been $\mathbf{M} = \chi_m \mathbf{B}/\mu_0$. The definition with $\mathbf{H}$ is the one in use because $\mathbf{H}$, unlike $\mathbf{B}$, is fixed by the free currents in the most common situations (Section 13.7), and those are the currents we control.

For most materials $|\chi_m|$ is so small that the difference between the two definitions is negligible.

In a linear material, (13.18) gives $\mathbf{B} = \mu_0(\mathbf{H}+\mathbf{M}) = \mu_0(1+\chi_m)\mathbf{H}$, so $\mathbf{B}$ is proportional to $\mathbf{H}$,

$$\mathbf{B} = \mu\mathbf{H}, \tag{13.20}$$

where

$$\mu = \mu_0\left(1+\chi_m\right) \tag{13.21}$$

is the **permeability** of the material, and

$$\mu_r = \frac{\mu}{\mu_0} = 1+\chi_m \tag{13.22}$$

is its **relative permeability**. In vacuum $\chi_m = 0$ and $\mu_r = 1$, which is why μ_0 is called the permeability of free space.

Table 13.1 lists the susceptibilities of some materials.[6] They are small: a few parts in 10^5 for typical solids, and a few parts in 10^9 for gases at ordinary pressures, except oxygen, whose molecules have two unpaired electrons. A few paramagnets have much larger susceptibilities, among them liquid oxygen, salts of iron and manganese, and the rare-earth metal gadolinium just above the temperature at which it becomes ferromagnetic. For ferromagnets, as Section 13.8.2 explains, a single value of χ_m means little.

Diamagnetic	
Bismuth	-1.65×10^{-4}
Gold	-3.45×10^{-5}
Silver	-2.39×10^{-5}
Diamond	-2.17×10^{-5}
Copper	-9.67×10^{-6}
Water	-9.01×10^{-6}
Nitrogen gas	-6.27×10^{-9}
Hydrogen gas	-2.08×10^{-9}
Paramagnetic	
Aluminum	2.08×10^{-5}
Tungsten	6.99×10^{-5}
Platinum	2.67×10^{-4}
$CuSO_4\cdot5H_2O$	1.68×10^{-4}
$MnCl_2$	4.27×10^{-3}
Oxygen gas	1.78×10^{-6}
Liquid oxygen, 90 K	3.45×10^{-3}
Gadolinium, 350 K	0.117

Table 13.1 Magnetic susceptibilities χ_m (SI) of some diamagnetic and paramagnetic materials: copper sulfate, $CuSO_4\cdot5H_2O$, and manganese chloride, $MnCl_2$, among them (sources in the footnote).

Since χ_m can have either sign, $\mathbf{M}$ can point either way. In a paramagnet, $\mathbf{M}$, $\mathbf{H}$, and $\mathbf{B}$ all point in the same direction; in a diamagnet, $\mathbf{M}$ points opposite to the other two. In a material with permanent magnetization, which is not a linear material, $\mathbf{H}$ and $\mathbf{M}$ often point in opposite directions, as Section 13.6.1 shows.

Not every equation of this chapter holds in every material. The definition (13.18), the bound currents (13.13) and (13.14), and Ampère's law for $\mathbf{H}$ in the next section hold always. Equations (13.19)–(13.22), and anything derived from them, hold only in linear materials, and never in permanent magnets. The summary at the end of the chapter sorts the equations into these two groups.

13.6 Ampère's law for H

Ampère's law with the total current density (13.17) and $\mathbf{J}_b = \nabla\times\mathbf{M}$ is

$$\frac{1}{\mu_0}\nabla\times\mathbf{B} = \mathbf{J}_f + \mathbf{J}_b = \mathbf{J}_f + \nabla\times\mathbf{M}, \qquad \text{so} \qquad \nabla\times\left(\frac{\mathbf{B}}{\mu_0}-\mathbf{M}\right) = \mathbf{J}_f.$$

[6] Calculated from the molar susceptibilities in "Magnetic Susceptibility of the Elements" and "Magnetic Susceptibility of Inorganic Compounds," in *CRC Handbook of Chemistry and Physics*, 107th Edition (Internet Version 2026), Leah R. McEwen, ed., CRC Press/Taylor & Francis, Boca Raton, FL, with densities from "Melting, Boiling, Triple, and Critical Points and Densities of the Elements" and "Physical Constants of Inorganic Compounds" in the same edition. The handbook gives molar susceptibilities χ_{mol} in the Gaussian units of cm^3/mol; the SI volume susceptibility is $\chi_m = 4\pi\chi_{\text{mol}}\rho/M_{\text{mol}}$, with ρ the density in g/cm^3 and M_{mol} the molar mass in g/mol. The values are at room temperature unless a temperature is given, and those for gases are for 20°C and 1 atmosphere.

The minus sign appears because **M** has to be moved to the left-hand side. By the definition (13.18),

$$\nabla \times \mathbf{H} = \mathbf{J}_f. \tag{13.23}$$

The curl of **H** is the free current density alone. By the curl theorem, the integral form is

$$\oint \mathbf{H} \cdot d\boldsymbol{\ell} = I_{f,\text{enc}}. \tag{13.24}$$

In words, the circulation of **H** around a closed loop is the free current passing through any surface bounded by the loop (Figure 13.10). This is **Ampère's law for H**. Only the definitions of **H** and of the bound current went into it, so it holds in every material.

It is used exactly as Ampère's law was in Section 11.2, and Colton's Third Commandment applies to it: draw the Amperian loop. When the free current, and everything else in the problem, has the symmetry of an infinite straight wire, an infinite plane, an infinite solenoid, or a toroid, the circulation reduces to H times a length, and (13.24) gives H directly from the free current, without any need to know the bound current.

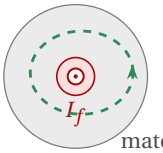


Figure 13.10 A wire carrying the free current I_f (out of the page) through a magnetic material. The circulation of **H** around the loop (dashed) is I_f, whatever the bound currents in the material.

! **Colton's Third Commandment** holds for **H** too: when using Ampère's law for **H**, draw the Amperian loop.

13.6.1 Gauss's law for H

The divergence of **B** is zero, always, (11.1). The divergence of **H** is not:

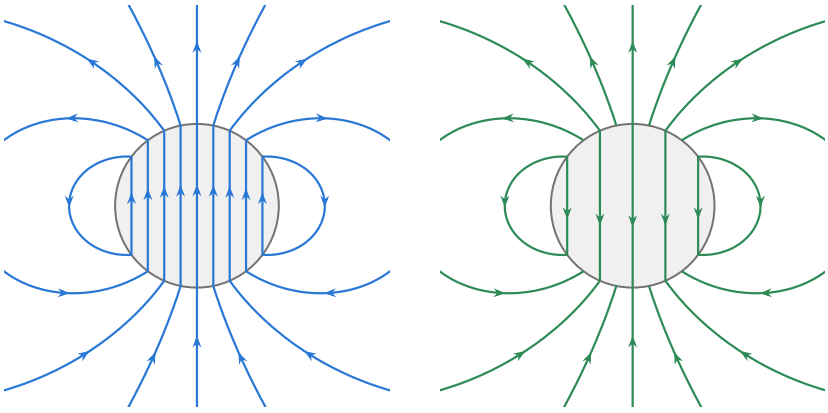
$$\nabla \cdot \mathbf{H} = \frac{1}{\mu_0} \nabla \cdot \mathbf{B} - \nabla \cdot \mathbf{M} = -\nabla \cdot \mathbf{M}, \tag{13.25}$$

or, by the divergence theorem,

$$\oint \mathbf{H} \cdot d\mathbf{a} = -\oint \mathbf{M} \cdot d\mathbf{a}. \tag{13.26}$$

Equation (13.25) has the form of Gauss's law, $\nabla \cdot \mathbf{E} = \rho/\varepsilon_0$, with $-\nabla \cdot \mathbf{M}$ in place of ρ/ε_0, and it deserves the name **Gauss's law for H**. It says that wherever **M** has a divergence, there are sources or sinks of **H** field lines, just as charges are sources and sinks of **E** field lines, but with the direction reversed by the minus sign: **H** lines begin where **M** lines end, and end where **M** lines begin. The usual place is the surface of a magnet, where the component of **M** perpendicular to the surface drops from its value inside to zero outside.

Figure 13.11 shows what this means for the uniformly magnetized sphere of Example 13.4. Its **M** lines run straight up inside the sphere and stop at its surface. On the upper hemisphere, where they stop, the **H** lines begin; on the lower hemisphere, where the **M** lines start, the **H** lines end. Outside, where $\mathbf{M} = 0$, $\mathbf{H} = \mathbf{B}/\mu_0$, and the two sets of lines are the same. Inside, $\mathbf{H} = \mathbf{B}/\mu_0 - \mathbf{M} = \frac{2}{3}\mathbf{M} - \mathbf{M} = -\frac{1}{3}\mathbf{M}$: the **H** lines run straight *down*, against the magnetization, from the upper hemisphere to the lower one. The **B** lines, on the other hand, never begin or end. They run up through the sphere, as strong as $\frac{2}{3}\mu_0 M$, and close on themselves outside.



(a) **B** (b) **H**

Figure 13.11 Field lines of (a) **B** and (b) **H** for the uniformly magnetized sphere of Example 13.4, with **M** pointing up. Outside the sphere the two patterns are the same, the field of a dipole; there the spacing of the lines shows the strength of the field only roughly. Inside, where both fields are uniform, the lines are evenly spaced; **B** points up, along **M**, and **H** points down, against it; the **H** lines begin on the upper hemisphere and end on the lower one, and inside they are half as dense as the **B** lines, since $|\mathbf{H}| = M/3$ while $|\mathbf{B}|/\mu_0 = 2M/3$.

The **H** lines in Figure 13.11b are exactly the field lines of **E** for the uniformly polarized sphere (Figure 9.23b), and that is no accident. With no free current anywhere, **H** has zero curl and divergence $-\nabla\cdot\mathbf{M}$; the field **E** of a polarized object has zero curl and divergence $-\nabla\cdot\mathbf{P}/\varepsilon_0$. So **H** is the field **E** that an object with the polarization $\mathbf{P} = \varepsilon_0\mathbf{M}$ would produce, the field of the "bound charges" $-\nabla\cdot\mathbf{M}$ and $\mathbf{M}\cdot\hat{\mathbf{n}}$ with ε_0 set to 1. Older books call these the magnetic pole densities, and the north and south poles of a magnet are the places where they are positive and negative. The poles are a convenient fiction for calculating **H**, not real magnetic charges: **B** has no divergence, and there are no magnetic monopoles (Section 11.1.1).

13.6.2 H is not determined by the free current alone

It is tempting to conclude from (13.23) that **H** is simply "the field of the free current," related to $\mathbf{J}_f$ the way **B** is related to **J**. It is not, and for the reason Section 9.6.1 gave for **D**. A field is determined by its curl and its divergence together (Section 2.3). The curl of **H** is the free current, but its divergence, $-\nabla\cdot\mathbf{M}$, need not be zero. Wherever the magnetization has a divergence, **H** has a source other than the free current, and then **H** is not zero even where there is no free current anywhere, as inside the magnetized sphere. Ampère's law for **H** is still true in such cases; it is just not useful, because **H** cannot be taken out of the integral. Only symmetry can tell you when it can be.

! **Two sources for H.** $\nabla\times\mathbf{H} = \mathbf{J}_f$, but $\nabla\cdot\mathbf{H} = -\nabla\cdot\mathbf{M}$: wherever **M** has a divergence, **H** has a source besides the free current.

Example 13.6

A cube of side a is a permanent magnet with uniform magnetization $\mathbf{M}$, perpendicular to two of its faces, and there is no free current anywhere. True or false: $\mathbf{H}$ is zero at the center of the cube.

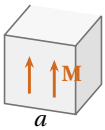
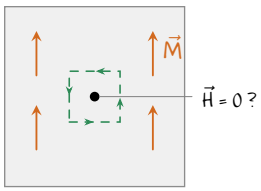


Figure 13.12 The magnetized cube of Example 13.6.

Solution:



To visualize the question, we sketch the cube in cross section, with $\mathbf{M}$ pointing up, perpendicular to the top and bottom faces. A square Amperian loop around the center (dashed) encloses no free current, so the circulation of $\mathbf{H}$ around it is zero. Does that make $\mathbf{H}$ zero at the center?

False. The tempting argument is: draw an Amperian loop around any point inside the cube, like the one in the sketch; it encloses no free current, so $\oint \mathbf{H}\cdot d\boldsymbol{\ell} = 0$, so $H = 0$. The last step is wrong. The circulation of $\mathbf{H}$ around every loop is zero, but that does not make $\mathbf{H}$ itself zero unless symmetry makes $\mathbf{H}$ constant and parallel to the loop along its sides, and a cube does not have that symmetry.

To find the actual value, use the bound currents. Since $\mathbf{M}$ is constant, $\mathbf{J}_b = 0$; the top and bottom faces are perpendicular to $\mathbf{M}$ and carry no bound current, and the four side faces carry $K_b = M$ around the cube, like a square solenoid of length a. The field on the axis of a square loop of side a carrying a current I, a distance u from its center, is

$$B = \frac{\mu_0 I a^2}{2\pi\left(u^2 + a^2/4\right)\sqrt{u^2 + a^2/2}}$$

(Exercise P10.15). The strip of the side faces between u and $u + du$ is such a loop, with $I = M\,du$, so at the center

$$B = \frac{\mu_0 M a^2}{2\pi}\int_{-a/2}^{a/2} \frac{du}{\left(u^2 + a^2/4\right)\sqrt{u^2 + a^2/2}} = \frac{\mu_0 M}{2\pi}\cdot 4\left[\arctan\frac{2u}{\sqrt{4u^2 + 2a^2}}\right]_{-a/2}^{a/2}$$
$$= \frac{\mu_0 M}{2\pi}\cdot 4\cdot\frac{\pi}{3} = \frac{2}{3}\mu_0 M,$$

along $\mathbf{M}$. Then

$$\mathbf{H}(\text{center}) = \frac{\mathbf{B}}{\mu_0} - \mathbf{M} = -\frac{1}{3}\mathbf{M}.$$

Reality check: Gauss's law for $\mathbf{H}$ gives the same answer by another route. The poles of the cube are on its top and bottom faces, $\mathbf{M}\cdot\hat{\mathbf{n}} = \pm M$, so $\mathbf{H}$ is the field of two uniformly "charged" faces, which Example 9.6 found at the center of a cube by a solid-angle argument: $-\mathbf{P}/3\varepsilon_0$ for faces charged $\pm P$, so here $-\mathbf{M}/3$.

13.6.3 Bound current in a linear material

In a linear material that is homogeneous, χ_m is the same everywhere, and

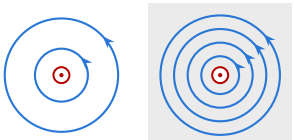
$$\mathbf{J}_b = \nabla \times \mathbf{M} = \chi_m \nabla \times \mathbf{H} = \chi_m \mathbf{J}_f. \tag{13.27}$$

Bound volume current flows only where free current flows: in the same direction in a paramagnet, and in the opposite direction in a diamagnet. The rest of the bound current is on the surfaces of the material. In particular, if a homogeneous linear material carries no free current, there is no bound current inside it either. Inside a homogeneous linear material, whether or not it carries free current, $\nabla \cdot \mathbf{H} = -\nabla \cdot \mathbf{M} = -\chi_m \nabla \cdot \mathbf{H}$, so $(1+\chi_m)\nabla \cdot \mathbf{H} = 0$ and $\nabla \cdot \mathbf{H} = 0$; with no free current, $\nabla \times \mathbf{H} = 0$ as well. Section 13.9.2 uses that.

13.6.4 What the permeability measures

Suppose a homogeneous linear material fills all the space where there is any field, and the free currents are the same as they would be in vacuum. Then $\mathbf{H}$ has the curl $\mathbf{J}_f$ and, by the last subsection, zero divergence everywhere. In vacuum, the field $\mathbf{B}_{\text{vac}}$ of the same free currents has $\nabla \times (\mathbf{B}_{\text{vac}}/\mu_0) = \mathbf{J}_f$ and $\nabla \cdot (\mathbf{B}_{\text{vac}}/\mu_0) = 0$. The two fields $\mathbf{H}$ and $\mathbf{B}_{\text{vac}}/\mu_0$ have the same curl and the same divergence everywhere (and both go to zero far away), so they are the same field (Section 2.3). Therefore

$$\mathbf{B} = \mu \mathbf{H} = \mu_r \mathbf{B}_{\text{vac}}. \tag{13.28}$$



(a) vacuum (b) $\mu_r = 2$

Figure 13.13 The same free current, a long wire, (a) in vacuum and (b) in a linear paramagnet that fills all of space, here with $\mu_r = 2$. $\mathbf{H}$ is the same in both, and $\mathbf{B}$ is twice as strong in (b), drawn with twice as many field lines.

A paramagnet *strengthens* the field of the free currents everywhere by the factor μ_r (Figure 13.13), and a diamagnet weakens it, where a dielectric always weakens the field of the free charges, (9.29). This is the precise form of the remark at the start of Section 13.2. The same rule of thumb applies as for dielectrics: take the answer without the material and replace μ_0 by μ. It is exact under the conditions just stated, and in a few other cases, such as the long solenoid of Example 13.9 below, but not in general when the material fills only part of the space where the field is. For a sphere of linear material in a uniform applied field $\mathbf{B}_0$, for example, the field inside is $3\mu_r \mathbf{B}_0/(\mu_r + 2)$ (Exercise P13.20), not $\mu_r \mathbf{B}_0$.

13.7 Solving for H when given M: Ampère's law for H method (high symmetry)

Ampère's law for $\mathbf{H}$ gives a second way to find the field of a magnetized object, and when it works it is much faster than the bound-current method:

1. Find $\mathbf{H}$ from Ampère's law for $\mathbf{H}$, (13.24), if the symmetry allows it. Draw the Amperian loop.

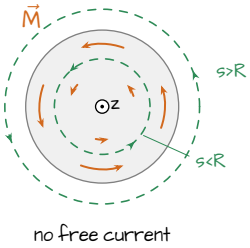
2. Then $\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M})$, from the definition (13.18).

The field found this way is the total field, from all sources. With no free current at all, **H** comes out to be zero (provided the symmetry is there to justify it), and the field is simply $\mathbf{B} = \mu_0\mathbf{M}$. When it is the free current that is given and the material is linear, the steps are: **H** from the free current, then $\mathbf{B} = \mu\mathbf{H}$, then $\mathbf{M} = \chi_m\mathbf{H}$, and finally, if they are wanted, the bound currents.

Example 13.7

Find the field of the cylinder of Example 13.5, with $\mathbf{M} = M_0(s^2/R^2)\,\hat{\boldsymbol{\phi}}$, by Ampère's law for **H**.

Solution:



As before, the first step is a picture of what we are given, now with the Amperian loops we will use: circles centered on the axis, one inside the cylinder and one outside. There is no free current anywhere.

The magnetization depends only on s and circles the axis, so by the same symmetry as in Example 13.5, **H** circles the axis and depends only on s. There is no free current, so for a circle of radius s, inside or outside,

$$\oint \mathbf{H}\cdot d\boldsymbol{\ell} = H\cdot 2\pi s = 0, \qquad \mathbf{H} = 0.$$

Then $\mathbf{B} = \mu_0\mathbf{M} = \mu_0 M_0(s^2/R^2)\,\hat{\boldsymbol{\phi}}$ inside and $\mathbf{B} = 0$ outside, where $\mathbf{M} = 0$.

Reality check: That is the answer of Example 13.5, in two lines instead of a page.

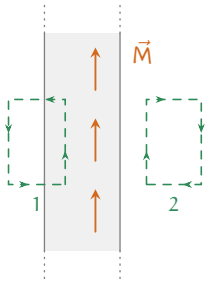
Example 13.8

A very long cylindrical bar magnet has uniform magnetization $\mathbf{M} = M_0\,\hat{\mathbf{z}}$ along its axis, and there is no free current. Find **H** and **B** inside and outside it, far from its ends.



Figure 13.14 The long bar magnet of Example 13.8, cut along its axis.

Solution:



We draw the magnet cut along its axis, with **M** pointing up inside, and two rectangular Amperian loops: one entirely outside, and one with a side inside the magnet and a side outside.

By symmetry, far from the ends **H** points along z and depends only on s, as in the infinite solenoid of Example 11.3. For a rectangle with both long sides outside the magnet, at s_1 and s_2, the short sides contribute nothing, since they are perpendicular to **H**, and no free current passes through the rectangle, so

$$[H(s_1) - H(s_2)]\,\ell = 0.$$

The field outside is the same at every distance, and since it must go to zero far away, it is zero everywhere outside. For a rectangle with one long side inside and one outside, the same argument gives $H_{\text{inside}} = H_{\text{outside}} = 0$. (Why does the bound current on the wrapper, which this rectangle encloses, not change the result? Answer in footnote.[7]) So

$$\mathbf{H} = 0 \quad \text{everywhere}, \qquad \mathbf{B} = \mu_0\mathbf{M} = \mu_0 M_0\,\hat{\mathbf{z}} \text{ inside}, \qquad \mathbf{B} = 0 \text{ outside}.$$

Reality check: The bound-current method agrees: the bound current $M_0\,\hat{\boldsymbol{\phi}}$ on the wrapper is that of an infinite solenoid with $nI = M_0$, whose field is $\mu_0 M_0\,\hat{\mathbf{z}}$ inside and zero outside.

Inside this magnet $\mathbf{H} = 0$ while $\mathbf{B} = \mu_0 M_0\,\hat{\mathbf{z}}$, so **B** is certainly not $\mu\mathbf{H}$ for any permeability. That is no contradiction: $\mathbf{B} = \mu\mathbf{H}$ holds only in linear materials, in which the magnetization is caused by the field and is proportional to it, and a permanent magnet is not one. The answer also depends on the magnet's being infinitely long. Inside a magnet of finite length the symmetry is lost, and **H** is not zero; Tutorial 13.1 takes this up.

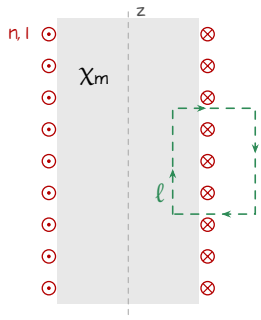
☞ *Caution*: $\mathbf{B} = \mu_0\mu_r\mathbf{H}$ holds only in linear materials. It does not hold in a permanent magnet.

Example 13.9

A long solenoid, with n turns per unit length carrying a current I, is filled with a linear material of susceptibility χ_m. Find **H**, **B**, and **M** inside, and the bound currents.

[7] Ampère's law for **H**, (13.24), counts only the free current through the loop, whatever bound current the loop also encloses (Figure 13.10).

Solution:



The sketch shows the solenoid cut along its axis, with its windings and the material inside, and the Amperian loop of Example 11.3: one long side inside and one outside.

The free current has the symmetry of the infinite solenoid, and $\mathbf{H}$ outside is zero by the argument of Example 13.8. The rectangle encloses the free current $nI\ell$, so

$$H\ell = nI\ell, \qquad \mathbf{H} = nI\hat{\mathbf{z}},$$

the same as without the material. Then

$$\mathbf{B} = \mu_0\left(1+\chi_m\right)nI\hat{\mathbf{z}}, \qquad \mathbf{M} = \chi_m nI\hat{\mathbf{z}}.$$

The material changes the field inside by the factor $\mu_r = 1+\chi_m$: it strengthens it if the material is paramagnetic and weakens it if it is diamagnetic. Filled with liquid oxygen, the solenoid's field is stronger by 0.35 percent; filled with bismuth, it is weaker by 0.017 percent. There is no free current inside the material, so there is no bound volume current, (13.27). On its surface, with $\hat{\mathbf{n}} = \hat{\mathbf{s}}$, the bound current is

$$\mathbf{K}_b = \chi_m nI\hat{\mathbf{z}}\times\hat{\mathbf{s}} = \chi_m nI\hat{\boldsymbol{\phi}},$$

which runs alongside the free current $nI\hat{\boldsymbol{\phi}}$ of the windings for a paramagnet, against it for a diamagnet.

Reality check: The total surface current, free and bound, is $(1+\chi_m)nI$, and a solenoid with that surface current has the field $\mu_0(1+\chi_m)nI$ inside: the same $\mathbf{B}$.

Example 13.10

A long straight aluminum wire of radius R carries a current I along $+\hat{\mathbf{z}}$, spread uniformly over its cross section. Aluminum is a linear paramagnet with $\chi_m = 2.08\times10^{-5}$. Find $\mathbf{H}$, $\mathbf{B}$, and $\mathbf{M}$, and the bound currents. Which way do they flow?

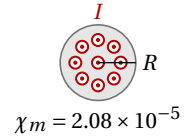
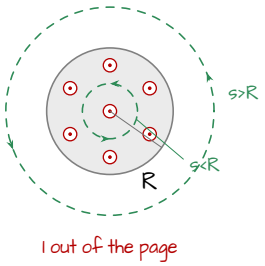


Figure 13.15 The aluminum wire of Example 13.10, end on.

Solution:



We draw the wire end on, with the free current coming out of the page, and two circular Amperian loops, one inside the wire and one outside. The loop inside encloses only part of the free current.

By symmetry $\mathbf{H}$ circles the wire and depends only on s. A circle of radius $s < R$ encloses the free current Is^2/R^2, and a circle outside encloses I, so

$$\mathbf{H} = \frac{Is}{2\pi R^2}\hat{\boldsymbol{\phi}} \quad (s < R), \qquad \mathbf{H} = \frac{I}{2\pi s}\hat{\boldsymbol{\phi}} \quad (s > R).$$

Inside, $\mathbf{B} = \mu_0(1+\chi_m)\mathbf{H}$ and $\mathbf{M} = \chi_m\mathbf{H}$; outside, $\mathbf{B} = \mu_0\mathbf{H}$, the field of a wire in vacuum, and $\mathbf{M} = 0$. The bound volume current is, by (13.27),

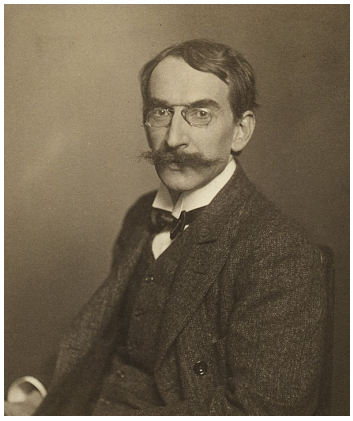
$$\mathbf{J}_b = \chi_m \mathbf{J}_f = \chi_m \frac{I}{\pi R^2}\hat{\mathbf{z}},$$

in the same direction as the free current, and at the surface, with $\hat{\mathbf{n}} = \hat{\mathbf{s}}$,

$$\mathbf{K}_b = \mathbf{M}\times\hat{\mathbf{s}} = \chi_m \frac{I}{2\pi R}\hat{\boldsymbol{\phi}}\times\hat{\mathbf{s}} = -\chi_m\frac{I}{2\pi R}\hat{\mathbf{z}},$$

opposite to it. Both are tiny, about 21 parts per million of the free current.

Reality check: The two add up to zero, $\chi_m I - \chi_m I = 0$, as (13.15) requires, which is why the field outside is unchanged.



Pierre Weiss (1865–1940, French) Weiss was born in Mulhouse and trained first as an engineer, at the polytechnic school in Zurich, where he graduated at the top of his class in 1887. He then studied physics at the École Normale Supérieure in Paris, and his doctoral thesis (1896) was on the magnetization of crystals of magnetite. In 1902 he returned to Zurich as professor of physics and director of the physics institute of the polytechnic, where in 1907 he published his theory of ferromagnetism, with its internal "molecular field" and its regions of spontaneous magnetization, now called Weiss domains. During the First World War he worked on locating enemy artillery by sound. In 1919 he moved to Strasbourg, newly returned to France, to direct its institute of physics, where Louis Néel prepared his thesis under him. He died in Lyon in 1940. (Wikipedia)

The bound surface current in Example 13.10 flows *against* the free current, even though the material is paramagnetic. The surface current is $\mathbf{M}\times\hat{\mathbf{n}}$, and at the surface of the wire $\mathbf{M}$ circles the axis, so its cross product with the outward normal points down the wire.

13.8 Ferromagnets: domains, hysteresis, and the Curie temperature

In a **ferromagnet** the atomic moments line up with one another even when there is no applied field. What lines them up is not the magnetic force between them. Two moments of one Bohr magneton 2.5 Å apart, about the spacing of atoms in iron, have a magnetic interaction energy of order $\mu_0\mu_B^2/4\pi r^3 = 3\times10^{-6}\,\text{eV}$, which thermal motion would overwhelm at any temperature above about 0.04 K. Iron stays ferromagnetic up to 1043 K. In 1907 Pierre Weiss proposed that some much stronger interaction inside the material, which he described as an internal

"molecular field," lines the moments up. Its origin was found in 1928, after the arrival of quantum mechanics, and it is electrical: because of the Pauli exclusion principle, the electric repulsion between the electrons of neighboring atoms depends on whether their spins are parallel or opposite. This effect, the exchange interaction, has no counterpart in classical physics. This book takes the alignment as given and describes its consequences.

13.8.1 Domains

Weiss also proposed that a ferromagnet is divided into small regions, each magnetized to its full spontaneous magnetization, but in different directions, so that an ordinary piece of iron can show little or no magnetization overall. These regions are the domains, and the thin layers between them, across which the direction of magnetization turns, are **domain walls**. Domains form because a single domain would produce a large field outside the material, and that field stores energy; dividing the material into domains that point different ways cancels most of the outside field.

When a field is applied (Figure 13.16), the domains whose magnetization points nearly along the field grow at the expense of the others, as the walls between them move. In stronger fields the magnetization of the remaining domains turns toward the direction of the field. Eventually the whole sample is magnetized along the field, and the magnetization can grow no further: it has reached the saturation magnetization M_s of Example 13.2. For iron at room temperature, $\mu_0 M_s = 2.15\,\mathrm{T}$.

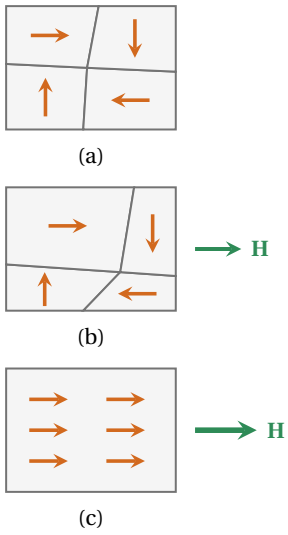


Figure 13.16 Domains in a ferromagnet, each magnetized in the direction of its arrow. (a) With no applied field, the domains point in different directions and their magnetizations nearly cancel. (b) In an applied field **H**, the domains pointing nearly along the field grow at the expense of the others, as their walls move. (c) In a strong field the whole sample is magnetized along the field, and the magnetization has reached its saturation value.

13.8.2 Hysteresis

The domain walls do not move freely. They catch on impurities and imperfections in the crystal, and when the applied field is reduced again, many of them stay where the field pushed them. The magnetization therefore depends on the history of the sample, not just on the field applied now. Figure 13.17 shows a typical result. A sample that starts out unmagnetized, at the origin, follows the curve oa as **H** increases, until it saturates at a. When **H** is then reduced to zero, the magnetization does not go back to zero: it follows ab, and at b the sample retains the **remanent magnetization**, or **remanence**, M_r. It has become a permanent magnet. To bring the magnetization back to zero, **H** must be reversed, to the value $-H_c$ at c; that field is the **coercive field**, or **coercivity**. A still larger reversed field saturates the sample the other way, at d, and making **H** positive again brings it back to a along the lower branch, through e and f. The closed curve $abcdefa$ is the **hysteresis loop**.[8]

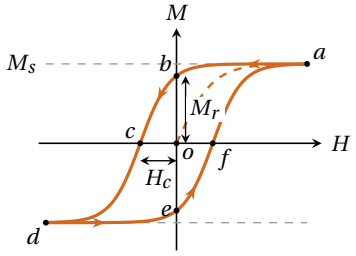
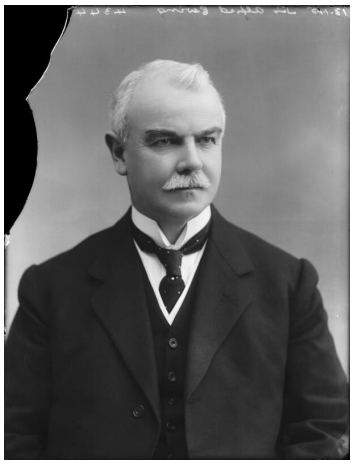


Figure 13.17 A hysteresis loop. An unmagnetized sample follows oa to saturation. With **H** reduced to zero it keeps the remanent magnetization M_r (b); the reversed field H_c brings it to zero (c), and a larger one saturates it the other way (d). The return branch def completes the loop. Animation.

Since the state of a ferromagnet depends on its history, a ferromagnet has no single susceptibility. The ratio M/H is different on the way up and on the way down; it depends on the field; and at b, with $H = 0$ and $M = M_r$, it is infinite. The

[8] The word is James Alfred Ewing's, from a Greek word for a lagging behind, or a falling short.



James Alfred Ewing (1855–1935, Scottish) Ewing was born in Dundee and studied physics under Peter Guthrie Tait and engineering at the University of Edinburgh, and he spent his summers on telegraph-cable expeditions led by William Thomson and Fleeming Jenkin. From 1878 to 1883 he taught mechanical engineering at the University of Tokyo, where he helped develop seismographs and studied how the magnetization of iron lags behind the field that produces it, an effect he named hysteresis. He was a professor at Dundee and then, from 1890, at Cambridge, until in 1903 he became the Admiralty's Director of Naval Education. In the First World War he directed the Admiralty's code-breaking office, known as Room 40. From 1916 he was Principal of the University of Edinburgh. He was knighted in 1911. Portrait: Sir (James) Alfred Ewing by Bassano Ltd, whole-plate glass negative, 13 January 1915. © National Portrait Gallery, London, CC BY-NC-ND 3.0 (https://creativecommons.org/licenses/by-nc-nd/3.0/). (Wikipedia)

equations of Section 13.5.1 apply to a ferromagnet only as rough approximations over a limited range of fields, with a value of M/H read from the curve. The relative permeability quoted for an iron core is a value of this kind, usually the ratio $B/\mu_0 H$ along the curve oa, and it changes with the field. Notice also that in iron M is usually far larger than H, so that $\mathbf{B} = \mu_0(\mathbf{H}+\mathbf{M})$ is very nearly $\mu_0\mathbf{M}$, and a plot of B against H has the same shape as Figure 13.17.

How wide the loop is depends on the material. In **soft** magnetic materials, such as annealed iron and iron–silicon alloys, the domain walls move easily, the coercive field is small, and the loop is narrow. They are used where the magnetization must follow a changing field, as in the cores of electromagnets and transformers (Section 13.11.1 explains why a narrow loop matters there). In **hard** magnetic materials, such as hardened steel and the alloys and compounds used in modern permanent magnets, the coercive field is large and the loop is wide, so that once magnetized they stay magnetized. Inside a magnetized sphere, **H** points against **M** (Section 13.6.1), and that field tends to undo the magnetization; a large coercive field resists it.

There are two common ways to demagnetize a permanent magnet. One is to take it around a series of hysteresis loops, each smaller than the last, by applying an alternating field whose amplitude slowly decreases to zero; the magnetization spirals in toward the origin. The other is to heat it, as the next subsection explains.

13.8.3 The Curie temperature

As a ferromagnet is heated, thermal motion increasingly disturbs the alignment of the moments, and its saturation magnetization decreases, slowly at first and then rapidly, until at the **Curie temperature** T_C it disappears. Above T_C the material is a paramagnet. Pierre Curie studied this transition in his thesis of 1895. Table 13.2 lists the saturation magnetizations and Curie temperatures of some ferromagnets, and of magnetite, which is a ferrimagnet (Section 13.8.4).[9] Gadolinium's Curie temperature is just below room temperature, so at room temperature it is not ferromagnetic.

13.8.4 Antiferromagnets and ferrimagnets

In some materials the interaction between neighboring moments has the opposite effect: it makes them point in *opposite* directions. If the opposite moments are equal, as in manganese oxide, MnO, they cancel, and the material has no spontaneous magnetization at all. Such a material is an **antiferromagnet**. Its order can be detected in other ways, and it disappears above a temperature called the **Néel temperature**, above which the material is an ordinary paramagnet. In a **ferrimagnet**, the opposite moments are unequal, and a net spontaneous magnetization remains. Ferrimagnets behave much like ferromagnets, with domains,

[9] From *Introduction to Solid State Physics*, 8th edition, by Charles Kittel (Wiley, 2005), Table 1 of Chapter 12, "Ferromagnetic crystals," which gives M_s in gauss, as the Gaussian unit of magnetization is often called; for a magnetization, 1 gauss corresponds to 10^3 A/m.

hysteresis, and a Curie temperature, and many of them are electrical insulators, which makes them useful at high frequencies. The ferrites, oxides of iron combined with other metals, are ferrimagnets. So is magnetite, Fe_3O_4, the mineral of the lodestones that began the study of magnetism.

	M_s		T_C
	room	0 K	(K)
Iron	1.707	1.740	1043
Cobalt	1.400	1.446	1388
Nickel	0.485	0.510	627
Gadolinium	—	2.060	292
Fe_3O_4	0.480	—	858
CrO_2	0.515	—	386

Table 13.2 Saturation magnetization M_s, in 10^6 A/m, at room temperature and near 0 K, and the Curie temperature T_C, of some ferromagnets and of magnetite, Fe_3O_4, a ferrimagnet; CrO_2 is chromium dioxide (source in the footnote).

13.9 Boundary conditions at magnetic interfaces

Section 11.3 found the boundary conditions on **B** at a surface carrying a surface current **K**. With magnetic materials there are two more, one for each of the new equations (13.23) and (13.25). As there, "above" is the side of the surface that the unit normal $\hat{\mathbf{n}}$ points toward. Each condition comes from one of the four equations for **B** and **H** (Table 13.3).

Equation	Boundary condition
$\nabla\cdot\mathbf{B}=0$	$B^{\perp}_{\text{above}} = B^{\perp}_{\text{below}}$
$\nabla\times\mathbf{B}=\mu_0\mathbf{J}$	$\mathbf{B}^{\parallel}_{\text{above}} - \mathbf{B}^{\parallel}_{\text{below}} = \mu_0\mathbf{K}\times\hat{\mathbf{n}}$
$\nabla\times\mathbf{H}=\mathbf{J}_f$	$\mathbf{H}^{\parallel}_{\text{above}} - \mathbf{H}^{\parallel}_{\text{below}} = \mathbf{K}_f\times\hat{\mathbf{n}}$
$\nabla\cdot\mathbf{H}=-\nabla\cdot\mathbf{M}$	$H^{\perp}_{\text{above}} - H^{\perp}_{\text{below}} = -\left(M^{\perp}_{\text{above}} - M^{\perp}_{\text{below}}\right)$

Table 13.3 The boundary conditions at a surface in magnetostatics, each with the equation it comes from.

The first two are (11.9) and (11.10). They hold at the surface of a magnetic material just as they do anywhere else, but the **K** in the second is the total surface current, free plus bound, and at the surface of a magnetized object the bound part is $\mathbf{K}_b = \mathbf{M}\times\hat{\mathbf{n}}$. As with Ampère's law, **H** provides a version that involves only the free current. Apply Ampère's law for **H**, (13.24), to a short, squat loop that straddles the surface, as in Figure 11.17. As the loop is made thinner, only its two long sides contribute, and the free current through it is the free surface current crossing the loop. The same steps as in Section 11.3 give

$$\mathbf{H}^{\parallel}_{\text{above}} - \mathbf{H}^{\parallel}_{\text{below}} = \mathbf{K}_f\times\hat{\mathbf{n}}. \tag{13.29}$$

Similarly, a thin pillbox and Gauss's law for **H**, (13.26), give

$$H^{\perp}_{\text{above}} - H^{\perp}_{\text{below}} = -\left(M^{\perp}_{\text{above}} - M^{\perp}_{\text{below}}\right), \tag{13.30}$$

which also follows directly from the continuity of $B^{\perp}$ and the definition of **H**. At the surface of a magnet, where $M^{\perp}$ drops to zero, $H^{\perp}$ jumps; that is where the "poles" of Section 13.6.1 sit.

13.9.1 Interfaces between linear materials

If the materials on both sides are linear, $\mathbf{B}=\mu\mathbf{H}$ on each side, and the two conditions become

$$\mu_{\text{above}}H^{\perp}_{\text{above}} = \mu_{\text{below}}H^{\perp}_{\text{below}}, \qquad \frac{1}{\mu_{\text{above}}}\mathbf{B}^{\parallel}_{\text{above}} - \frac{1}{\mu_{\text{below}}}\mathbf{B}^{\parallel}_{\text{below}} = \mathbf{K}_f\times\hat{\mathbf{n}}. \tag{13.31}$$



Louis Néel (1904–2000, French) Néel was born in Lyon and studied at the École Normale Supérieure in Paris. He prepared his doctoral thesis in Pierre Weiss's laboratory in Strasbourg, where he became a professor in 1937. In 1936 he proposed that in some materials neighboring atomic moments point in opposite directions, the order now called antiferromagnetism; the first evidence for it came from manganese oxide two years later. In the spring of 1940 he found a way to protect the ships of the French navy from German magnetic mines by neutralizing their magnetization. After the defeat of France he moved to Grenoble, which he built after the war into a major center of research. In the late 1940s he explained the magnetism of the ferrites as ferrimagnetism, and he also explained the weak but stable magnetism of some rocks, which records the history of the Earth's magnetic field. He shared the Nobel Prize in Physics for 1970 with Hannes Alfvén. (Wikipedia)

At most interfaces there is no free surface current, and then $\mathbf{H}^{\parallel}$ and $B^{\perp}$ are both continuous. Field lines then change direction at the interface. Let θ be the angle between $\mathbf{B}$ and the normal on each side, so that $\tan\theta = B^{\parallel}/B^{\perp} = \mu H^{\parallel}/B^{\perp}$. Since $H^{\parallel}$ and $B^{\perp}$ are the same on both sides,

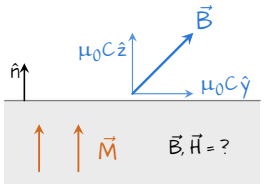
$$\frac{\tan\theta_{\text{above}}}{\tan\theta_{\text{below}}} = \frac{\mu_{\text{above}}}{\mu_{\text{below}}}, \tag{13.32}$$

which has the same form as the rule (9.34) for dielectrics, with μ in place of ε. A material with a large permeability, such as iron, makes the effect dramatic. With μ much larger in the iron than in the air, field lines that run nearly along the surface inside the iron leave it almost perpendicular to the surface (Figure 13.18). That is why field lines leave the pole faces of an iron electromagnet nearly perpendicular to them.

Example 13.11

A large slab of ferromagnetic material lies parallel to the xy plane. It has a frozen-in, uniform magnetization $\mathbf{M} = C\hat{\mathbf{z}}$. Other sources nearby also produce a field, and just above the top surface of the slab the total field is $\mathbf{B} = \mu_0 C(\hat{\mathbf{y}} + \hat{\mathbf{z}})$. (a) Find $\mathbf{B}$ and $\mathbf{H}$ just inside the top surface. (b) A sheet of free current $\mathbf{K} = K_0\hat{\mathbf{x}}$ is now laid on the top surface, and the outside sources are adjusted so that $\mathbf{B}$ just above the surface is unchanged. Find $\mathbf{B}$ and $\mathbf{H}$ just inside the top surface now.

Solution:



We sketch the slab edge on, with $\mathbf{M}$ pointing up inside it and the given field just above it, split into its parts along the surface and perpendicular to it. The normal $\hat{\mathbf{n}}$ points up, so "above" is outside the slab.

(a) The bound current on the top surface is $\mathbf{K}_b = \mathbf{M}\times\hat{\mathbf{n}} = C\hat{\mathbf{z}}\times\hat{\mathbf{z}} = 0$, and there is no free current, so there is no surface current at all, and $\mathbf{B}$ is continuous: just inside,

$$\mathbf{B} = \mu_0 C(\hat{\mathbf{y}} + \hat{\mathbf{z}}), \qquad \mathbf{H} = \frac{\mathbf{B}}{\mu_0} - \mathbf{M} = C(\hat{\mathbf{y}} + \hat{\mathbf{z}}) - C\hat{\mathbf{z}} = C\hat{\mathbf{y}}.$$

Reality check: Equation (13.30) holds: above the surface $H^{\perp} = C$, below it $H^{\perp} = 0$, and the jump, C, equals $-(M^{\perp}_{\text{above}} - M^{\perp}_{\text{below}}) = -(0 - C)$.

(b) $B^{\perp}$ is still continuous, so $B_z = \mu_0 C$ inside. The bound surface current is still zero, so $\mathbf{K} = K_0\hat{\mathbf{x}}$ is the whole surface current. With $\mathbf{K}\times\hat{\mathbf{n}} = K_0\hat{\mathbf{x}}\times\hat{\mathbf{z}} = -K_0\hat{\mathbf{y}}$, (11.10) gives

$$\mathbf{B}^{\parallel}_{\text{below}} = \mathbf{B}^{\parallel}_{\text{above}} - \mu_0\mathbf{K}\times\hat{\mathbf{n}} = \mu_0 C\hat{\mathbf{y}} + \mu_0 K_0\hat{\mathbf{y}},$$

so just inside

$$\mathbf{B} = \mu_0(C + K_0)\hat{\mathbf{y}} + \mu_0 C\hat{\mathbf{z}}, \qquad \mathbf{H} = \frac{\mathbf{B}}{\mu_0} - \mathbf{M} = (C + K_0)\hat{\mathbf{y}}.$$

Reality check: Equation (13.29) holds: $\mathbf{H}^{\parallel}_{\text{above}} - \mathbf{H}^{\parallel}_{\text{below}} = C\hat{\mathbf{y}} - (C + K_0)\hat{\mathbf{y}} = -K_0\hat{\mathbf{y}} = \mathbf{K}_f \times \hat{\mathbf{n}}$.

13.9.2 Boundary-value problems with magnetic materials

Where there is no free current, $\nabla \times \mathbf{H} = 0$, and $\mathbf{H}$ can be written as the gradient of a scalar, the **magnetic scalar potential** ϕ_M:

$$\mathbf{H} = -\nabla\phi_M. \tag{13.33}$$

(The region must be one in which no closed loop goes around a free current. Around a wire, ϕ_M would change by I on each trip around, and would not be a single-valued function.) Inside a homogeneous linear material with no free current, $\mathbf{H}$ also has zero divergence (Section 13.6.3), so ϕ_M satisfies Laplace's equation, $\nabla^2\phi_M = 0$. At an interface with no free surface current, $\mathbf{H}^{\parallel}$ is continuous, which is guaranteed if ϕ_M is continuous, and $B^{\perp} = \mu H^{\perp}$ is continuous, which means that $\mu\partial\phi_M/\partial n$ is continuous. These are exactly the conditions on the potential V at an interface between dielectrics (Section 9.8.4), with ε replaced by μ. So boundary-value problems with linear magnetic materials and no free currents are solved by the methods of Chapter 7, and many of their answers can be read off from the dielectric versions. A sphere of linear material in a uniform applied field $\mathbf{B}_0$, for example, has the uniform field $\mathbf{B} = 3\mu_r\mathbf{B}_0/(\mu_r + 2)$ inside, the counterpart of Example 9.11 (Exercise P13.20). For a permanent magnet with a given $\mathbf{M}$, ϕ_M satisfies Poisson's equation instead, with $-\nabla\cdot\mathbf{M}$ in the role of ρ/ε_0.

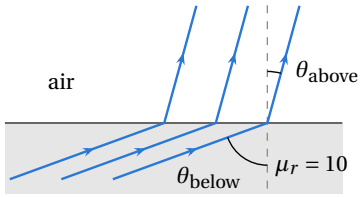


Figure 13.18 Field lines of $\mathbf{B}$ crossing the surface of a linear material with $\mu_r = 10$ (shaded, below) into air. $B^{\perp}$ is the same on both sides and $H^{\parallel} = B^{\parallel}/\mu$ is the same on both sides, so the lines bend toward the normal as they leave the material. For iron, whose μ_r is much larger, the lines in the air would be almost exactly perpendicular to the surface.

Tutorial 13.1 B and H in a bar magnet Printable version

Background

Magnetic materials contain bound currents, which produce fields exactly as free currents do (Section 13.3). We chose the definition of the $\mathbf{H}$ field to give us an Ampère's law that leaves out the bound currents, and that allows an easy calculation of $\mathbf{H}$ in certain high-symmetry situations:

$$\nabla \times \mathbf{H} = \mathbf{J}_f \qquad \longleftrightarrow \qquad \oint \mathbf{H}\cdot d\boldsymbol{\ell} = I_{f,\text{enc}}.$$

Keep in mind that just because $I_{f,\text{enc}}$ might be zero, that does not necessarily mean that $\mathbf{H}$ is zero if the situation is not highly symmetric. Also, the divergence of $\mathbf{H}$ may or may not be zero, unlike the divergence of $\mathbf{B}$, which is always zero. The goal of this tutorial is to use the concept of bound currents, boundary conditions, and more, to be able to sketch the $\mathbf{B}$ and $\mathbf{H}$ fields in and around a finite bar magnet, and hence to develop more intuition about $\mathbf{B}$ and $\mathbf{H}$ fields in general.

Part 1 — Infinite solenoid

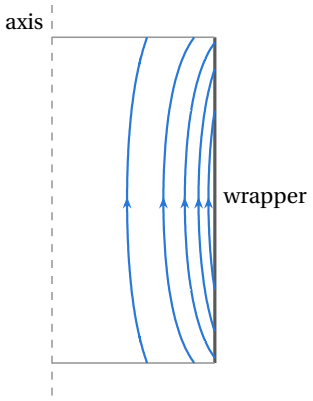


Figure 13.19 The right-hand half of a cross section of a finite solenoid, for Tutorial 13.1, Part 2, with field lines of **B** inside it. Near the ends some of the lines leave through the wrapper.

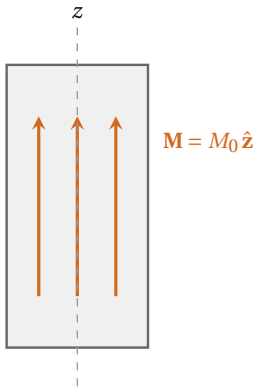


Figure 13.20 The bar magnet of Tutorial 13.1, Part 3, in a cross section through its axis, with its uniform magnetization $\mathbf{M} = M_0\,\hat{\mathbf{z}}$.

(a) Inside an infinite solenoid oriented along the z axis, with a surface current $\mathbf{K}$ in the $\hat{\boldsymbol{\phi}}$ direction, the magnetic field is constant and points in the z direction. Use Ampère's law to determine the field $\mathbf{B}_{\text{inside}}$ in terms of the surface current.

(b) Use the boundary condition $\mathbf{B}^{\parallel}_{\text{above}} - \mathbf{B}^{\parallel}_{\text{below}} = \mu_0 \mathbf{K} \times \hat{\mathbf{n}}$ to show that the field outside the solenoid must be zero. You needed to know that in order to use Ampère's law properly in (a), but it is still an interesting result when it is reached this way. Hint: which region is "above"? Which way does $\hat{\mathbf{n}}$ point?

Part 2 — Finite solenoid

With a finite solenoid, the magnetic field is no longer constant, but decreases near the top and bottom as the field lines spread out. Therefore the field lines inside the solenoid must look like those in Figure 13.19, drawn for the right-hand half of a cross section. Notice how this requires that some of the field lines leave through the wrapper of the solenoid rather than going out the top and bottom.

(a) Focus on one of the lines that leaves through the right-hand side. Write down the boundary conditions for $B^{\perp}$ and $\mathbf{B}^{\parallel}$ and apply them to this situation to determine the approximate direction of the field just outside the solenoid. You don't need an exact angle for the field line outside the solenoid; we're looking for the general behavior. Hint: consider how strong K is relative to the new B_{inside} of the finite solenoid, which is weaker than the field inside an infinite solenoid. The magnitude of K isn't explicitly known, so just decide if $\mu_0 K$ is larger than, smaller than, or equal to the new B_{inside}, and then decide what that means for the field outside.

(b) Field lines of **B** never begin or end, and for a solenoid they close on themselves. How will this field line close?

(c) Now draw the full picture of the magnetic field of a finite solenoid, with the side loops included.

(d) What will the **H** field look like for this situation? How will the **H** field lines compare to the **B** field lines you have just drawn? Hint: consider the definition of **H**.

(e) What are the boundary conditions for $H^{\perp}$ and $\mathbf{H}^{\parallel}$? Describe how they are consistent with your answer to the previous question.

Part 3 — Finite bar magnet

A finite cylindrical bar magnet has the magnetization $\mathbf{M} = M_0\,\hat{\mathbf{z}}$ (Figure 13.20). The goal here is to sketch the **B** and **H** field lines and to make sure that everything makes sense in terms of what we know about these fields. To help you do so, first answer all of these questions:

(a) What currents exist? Volume or surface? Free or bound? Where are they, and in what directions do they flow?

(b) What are the boundary conditions for **B** on each surface (top, bottom, and wrapper)?

(c) How does this situation compare with Part 2 with regard to the magnetic field?

(d) What does the **M** field look like when thought of as a vector field in space?

Make a field-line sketch of **M**. Does **M** have divergence anywhere? If so, where? Hint: how is divergence seen in field lines?

(e) What is **H** at each point in space? Does **H** have divergence anywhere?

(f) Field lines of **B** never begin or end. Why is that?

(g) Do **H** field lines have to close on themselves? Why or why not?

(h) What are the boundary conditions for **H** on each surface? Be specific, that is, apply them to this particular situation. Hint: the perpendicular boundary condition for **H** comes from the condition $B^{\perp}_{\text{above}} = B^{\perp}_{\text{below}}$.

Now sketch the field lines of **B** and, on a separate drawing, those of **H**, in and around the magnet.

Part 4 — Wrap-up

(a) Check that your sketches are reasonable by considering the limiting case of an infinitely long bar magnet.

13.10 Numerical methods: direct integration and relaxation

The numerical methods of Section 9.9 carry over to magnetic materials, one for each kind of problem. When the magnetization is given, the bound currents can be summed numerically, as in Section 10.7. When the free currents are given and the materials are linear, the relaxation method for A_z of Section 12.5 can be adapted to a permeability that varies from place to place.

13.10.1 Direct integration over the bound currents

The bar magnet of Example 13.3 is a good illustration. Its bound current is $M_0\,\hat{\boldsymbol{\phi}}$ on its wrapper, so it is a stack of circular loops, each a ring of current $M_0\,dz'$ (Figure 13.21). Off the axis no simple formula gives the field, but a program can add up the fields of the rings, each of them computed with the elliptic-integral formulas of Exercise P10.22 or by dividing the ring into short pieces. The result on the axis must agree with Example 13.3, and far away with the dipole field; both make good checks on a program. Exercise P13.22 asks you to write one.

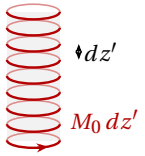


Figure 13.21 The bound current of the uniformly magnetized bar magnet: a stack of rings, each carrying the current $M_0\,dz'$.

13.10.2 Relaxation with magnetic materials

Section 12.5 solved for A_z in problems in which all the currents run along z and nothing depends on z. Its rule, (12.16), came from Ampère's law for **B**, which in a magnetic material involves the unknown bound currents. The rule that works everywhere comes from Ampère's law for **H** instead, in the same way that Section 9.9.2 got its rule from Gauss's law for **D**.

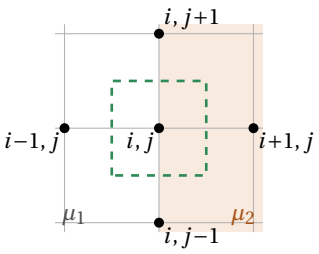


Figure 13.22 The small square (dashed) around the grid point (i, j), with its sides halfway to the four neighboring points; here the point lies on the boundary between materials of permeabilities μ_1 and μ_2.

Draw a small square around the grid point (i, j), with its sides halfway to the four neighboring points (Figure 13.22; compare Figure 9.24), and give each grid cell its own permeability. The field is $\mathbf{B} = (\partial A_z/\partial y)\hat{\mathbf{x}} - (\partial A_z/\partial x)\hat{\mathbf{y}}$, (12.17), and $\mathbf{H} = \mathbf{B}/\mu$. On the right-hand side of the square, $\mathbf{H}$ has the y component $-(1/\mu)\partial A_z/\partial x \approx -(1/\mu_E)(A_{i+1,j} - A_{i,j})/h$, where $1/\mu_E$ is $1/\mu$ for that side; and since half of the side lies in each of the two cells that it passes through, $1/\mu_E$ is the average of $1/\mu$ for those two cells. Going counterclockwise around the square, the right-hand side contributes $-(1/\mu_E)(A_{i+1,j} - A_{i,j})$ to the circulation of $\mathbf{H}$, and the other three sides contribute in the same way. Setting the circulation equal to the free current through the square, $J_f h^2$, and solving for $A_{i,j}$,

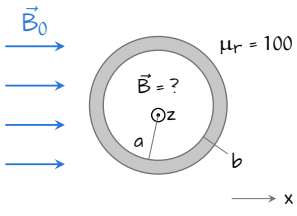
$$A_{i,j} = \frac{w_E A_{i+1,j} + w_W A_{i-1,j} + w_N A_{i,j+1} + w_S A_{i,j-1} + J_{f\,i,j} h^2}{w_E + w_W + w_N + w_S}, \qquad w = \frac{1}{\mu}. \tag{13.34}$$

The new value is a weighted average of the neighbors, each weighted by $1/\mu$ on the side between it and the center: the reciprocal of the permeability, where the dielectric rule (9.38) used the permittivity itself. If $\mu = \mu_0$ everywhere, (13.34) reduces to (12.16). The condition at an interface between materials, the continuity of $\mathbf{H}^{\parallel}$, is built into the rule, since it came from Ampère's law for $\mathbf{H}$. As with large dielectric constants, large ratios of permeability make plain relaxation converge slowly, and a direct solution of the grid equations with a sparse-matrix solver (Section 12.5) is usually the better choice.

Example 13.12

A long iron tube, with inner radius $a = 0.8b$ and outer radius b, is placed in a uniform applied field $\mathbf{B}_0$ perpendicular to its axis. Treating the iron as a linear material with $\mu_r = 100$, find the field inside the tube. How well does the tube shield its inside from the field?

Solution:



We draw the tube in cross section in the applied field, with the axis of the tube along z, out of the page, and the field along x. That is what a magnetic shield is: a shell of high-permeability material around the region we want to protect.

The applied field $B_0\hat{\mathbf{x}}$ is the curl of $A_z = B_0 y$, since $\partial A_z/\partial y = B_0$. Far from the tube the field is just the applied field, so we hold $A_z = B_0 y$ on the edges of a square grid centered on the tube, with half-width $8b$, and give each cell whose center lies in the iron the permeability $100\mu_0$. Solving the grid equations (13.34), with no free current, by a sparse solver, and finding the field at the center from (12.17), gives

Grid spacing h	B_{in}/B_0
$b/20$	0.116
$b/40$	0.108
$b/80$	0.104

The field inside is reduced to about a tenth of the applied field (Figure 13.23). The field lines, which are the contours of A_z (Section 12.5), crowd into the iron and pass around the inside; very few of them cross it.

The results change with h more than in earlier examples: each halving of h roughly halves the change. The cause is the grid itself. A circle drawn with square cells has a staircase edge, and the error that makes is proportional to h, while the error of the averaging rule itself is proportional to h^2. Taking the trend to its limit gives about 0.100. (Why does the trend give 0.100? Answer in footnote.[10])

Reality check: The exact answer for a tube in an infinite region, found by separation of variables with the magnetic scalar potential (Exercise P13.21), is

$$\frac{B_{in}}{B_0} = \frac{4\mu_r b^2}{(\mu_r+1)^2 b^2 - (\mu_r-1)^2 a^2} = 0.102,$$

and most of the remaining difference comes from holding $A_z = B_0 y$ on edges only $8b$ away, where the tube's own field has not quite died out. For a thin tube with $\mu_r(b-a) \gg b$, the exact formula gives $B_{in}/B_0 \approx 2b/\mu_r(b-a)$: the shielding improves in proportion to the permeability and to the thickness of the wall.

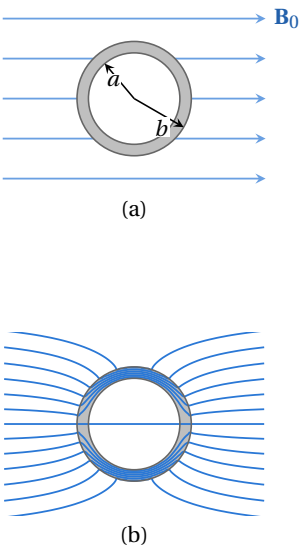


Figure 13.23 (a) A magnetic shield: a long tube of high-permeability material (gray), inner radius a and outer radius b, in a uniform applied field $\mathbf{B}_0$ perpendicular to its axis. (b) Field lines of $\mathbf{B}$ for $a = 0.8b$ and $\mu_r = 100$, contours of A_z from the relaxation equations (13.34) at equal steps. The lines crowd into the iron and avoid the inside, where the field is about a tenth of B_0.

Practical shields are made of alloys whose permeability is far larger than 100, and often of several nested layers. They never shield perfectly, though. The field lines enter the iron because it offers them an easy path, not because anything stops them, and a few always cross to the inside; by the exact formula, the field inside goes to zero only as μ_r goes to infinity.

13.11 Energy stored in the field, inside magnetic materials

Section 12.7.2 stated that the work needed to set up a distribution of steady currents in vacuum is $U = (1/2\mu_0)\int B^2\,d\tau$, (12.35), and explained why showing it needs Faraday's law. The same is true with magnetic materials present. What we do when we set up the currents is drive the free currents; the bound currents appear by themselves, as the material magnetizes. For linear materials the result is

$$U = \frac{1}{2}\int \mathbf{B}\cdot\mathbf{H}\,d\tau, \tag{13.35}$$

[10] Each halving of h halves the change, so the changes still to come, -0.002, -0.001, and so on, add up to the last one, -0.004, and $0.104 - 0.004 = 0.100$.

which in vacuum, where $\mathbf{H} = \mathbf{B}/\mu_0$, is (12.35) again. Written in terms of the free currents instead of the field, it is

$$U = \frac{1}{2}\int \mathbf{A}\cdot\mathbf{J}_f\, d\tau, \tag{13.36}$$

(12.34) with only the free current. That these two are equal needs only Ampère's law for $\mathbf{H}$: put $\mathbf{J}_f = \nabla\times\mathbf{H}$ into (13.36) and move the curl from $\mathbf{H}$ to $\mathbf{A}$ with the product rule, exactly as in Section 12.7.2; then $\mathbf{H}\cdot(\nabla\times\mathbf{A}) = \mathbf{H}\cdot\mathbf{B}$. As with dielectrics, (13.35) holds only for linear materials, and not, for example, for a ferromagnet. The electric and magnetic energies line up as in Table 13.4.

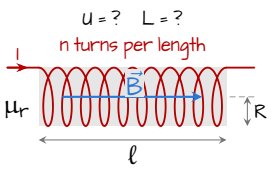
	Electric	Magnetic
Dipole of fixed moment in a field	$U = -\mathbf{p}\cdot\mathbf{E}$	$U = -\mathbf{m}\cdot\mathbf{B}$
Energy stored in the field, vacuum	$\frac{\varepsilon_0}{2}\int E^2\, d\tau$	$\frac{1}{2\mu_0}\int B^2\, d\tau$
Energy stored in the field, linear materials	$\frac{1}{2}\int \mathbf{E}\cdot\mathbf{D}\, d\tau$	$\frac{1}{2}\int \mathbf{B}\cdot\mathbf{H}\, d\tau$

Table 13.4 Electric and magnetic energies side by side.

Example 13.13

The long solenoid of Example 13.9, of radius R and length ℓ, with n turns per unit length carrying the current I, is filled with a linear material of relative permeability μ_r. Find the energy stored in its field and its inductance.

Solution:



Before any formulas go in, we sketch the solenoid with its filling and label it with what we know. Only the inside has a field, and there it is uniform, so the energy is the energy density times the volume.

Example 13.9 found $H = nI$ and $B = \mu_r\mu_0 nI$ inside, and both are zero outside. So

$$U = \frac{1}{2}BH\left(\pi R^2\ell\right) = \frac{1}{2}\mu_r\mu_0 n^2 I^2\pi R^2\ell.$$

(Why is the energy density $\frac{1}{2}BH$, and not $B^2/2\mu_0$ as in vacuum? Answer in footnote.[11]) Comparing with $U = \frac{1}{2}LI^2$, (12.37),

$$L = \mu_r\mu_0 n^2\pi R^2\ell,$$

[11] With a linear material present, the energy stored is the work done to set up the free currents, (13.35). Here $B^2/2\mu_0$ would give μ_r^2 times the energy of the empty solenoid; the flux, below, shows that the factor is μ_r.

μ_r times the inductance of the empty solenoid, (12.33). Filled with liquid oxygen, the solenoid of Example 12.10 would have its inductance raised by 0.35 percent, from 1.89 to 1.90 mH.

Reality check: The flux gives the same answer: the flux through each of the $n\ell$ turns is $\mu_r\mu_0 nI\pi R^2$, so $L = n\ell\,\mu_r\mu_0 nI\pi R^2/I$.

13.11.1 Energy lost to hysteresis

A ferromagnet that is taken around its hysteresis loop does not return the energy that went into magnetizing it (Figure 13.24). The work done per unit volume by whatever drives the free current, in one trip around the loop, is

$$\frac{U_{\text{lost}}}{\text{volume}} = \oint H\,dB, \tag{13.37}$$

the area enclosed by the loop in a plot of B against H, and it ends up as heat. (Showing this, too, needs Faraday's law.) That is why the core of a transformer, whose magnetization goes around its loop sixty times a second on a 60-hertz power line, is made of a soft material such as iron–silicon steel, whose loop is narrow.

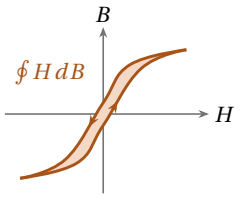


Figure 13.24 Energy lost to hysteresis. The area enclosed by the loop in a plot of B against H, $\oint H\,dB$, is the energy lost per unit volume in each trip around it.

13.11.2 A magnet lifting a nail

Section 10.1.3 promised to come back to a puzzle. Magnetic forces do no work, yet a permanent magnet lifts an iron nail, and the nail gains gravitational potential energy. What does the work?

First consider how the magnet holds the nail at all. The field of the magnet magnetizes the nail: its domains line up along the field (Section 13.8.1), so that the nail becomes a magnet whose moment points along the magnet's field. A dipole lined up with a field is pulled toward the stronger field, (13.3), so the nail is pulled toward the magnet (Figure 13.25), whichever pole of the magnet is nearer.

Now consider the work. The statement that magnetic forces do no work was derived from the force on a moving point charge, $q\mathbf{v}\times\mathbf{B}$, which is always perpendicular to the charge's velocity. If the nail's magnetic moments were tiny loops of circulating charge, the statement would hold for them. As such a loop moves toward the stronger field, the magnetic force on each circulating charge gets, from the loop's own motion, a component along the loop that slows the circulation, just as the force quB along the lifted wire of Section 10.4.2 opposed its current. (In the loop's own frame the field through it is changing, and the same effect appears as an induced electric field, Section 15.2.1.) The kinetic energy that the loop gains as a whole comes out of the kinetic energy of the circulation; the magnetic force only transfers it, as the normal force of the ramp in Section 10.1.3 turned a horizontal push into motion up the ramp. But the magnetism of iron comes mostly from electron spin, and a spin moment is not a loop of circulating

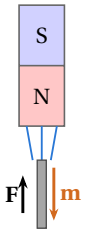


Figure 13.25 A magnet holding a nail. The magnet's field magnetizes the nail along the field, and the nail's moment **m**, lined up with the field, is pulled toward the stronger field near the pole.

charge. Its size is fixed, about one Bohr magneton for each electron, and there is no current in it to slow down. The force $\nabla(\mathbf{m}\cdot\mathbf{B})$ on a spin moment is not of the form $q\mathbf{v}\times\mathbf{B}$, so the argument that magnetic forces do no work does not apply to it. Some physicists conclude that here the magnetic force *does* do work on the nail; others argue that the energy comes from an internal energy of the spinning electron, as it does for the current loop. The question is still debated.[12]

With an electromagnet in place of the permanent magnet, at least the source of the energy is clear: as the nail moves in, the flux through the coil increases, and the source that keeps the coil's current steady does work against the induced voltage, as the batteries did when the two loops of Section 12.7.2 were brought together at fixed currents.

13.12 Summary

The atoms of a material carry magnetic moments, permanent or induced by a field, from the orbital motion and the spin of their electrons. In a field the material magnetizes: weakly along the field (paramagnets), weakly against it (diamagnets), or strongly and with a memory of its history (ferromagnets). The magnetization $\mathbf{M}$, the magnetic dipole moment per unit volume, produces exactly the field of the bound currents $\mathbf{J}_b = \nabla\times\mathbf{M}$ and $\mathbf{K}_b = \mathbf{M}\times\hat{\mathbf{n}}$, and given $\mathbf{M}$ the field can be found from them by any of the methods of Chapters 10–12. The field $\mathbf{H} = \mathbf{B}/\mu_0 - \mathbf{M}$ has the free current as its only curl, so with enough symmetry Ampère's law for $\mathbf{H}$ gives $\mathbf{H}$ directly from the free current; but $\mathbf{H}$ also has a divergence wherever $\mathbf{M}$ does, so it is not determined by the free current alone. As in Chapter 9, some equations hold in any material and others only in linear materials, and none of the second group holds in a permanent magnet. (Chapter 14 sets the electric and magnetic equations side by side.)

In any material

[12] For a classical model in which magnetic forces do work on elementary magnetic dipoles, see "On magnetic forces and work," J. A. Barandes, Found. Phys. 51, 79 (2021). https://doi.org/10.1007/s10701-021-00483-4. For an analysis in which whether the field does work on a spin moment depends on whether the electron has an intrinsic rotational kinetic energy, which it argues is consistent with the Dirac equation, see "Dipole in a magnetic field, work, and quantum spin," R. J. Deissler, Phys. Rev. E 77, 036609 (2008). https://doi.org/10.1103/PhysRevE.77.036609.

$\boldsymbol{\tau} = \mathbf{m} \times \mathbf{B}, \quad U = -\mathbf{m} \cdot \mathbf{B}$	torque and energy of a dipole of fixed moment in a field
$\mathbf{F} = \nabla(\mathbf{m} \cdot \mathbf{B})$	force on a dipole; equal to $(\mathbf{m} \cdot \nabla)\mathbf{B}$ only where $\mathbf{J} = 0$
$\mathbf{M} = N\mathbf{m}, \quad \mathbf{m}_{\text{total}} = \int \mathbf{M}\, d\tau$	magnetization: magnetic dipole moment per unit volume
$\mathbf{J}_b = \nabla \times \mathbf{M}, \quad \mathbf{K}_b = \mathbf{M} \times \hat{\mathbf{n}}$	the bound currents, whose field is the field of the magnetized object
$\mathbf{H} = \mathbf{B}/\mu_0 - \mathbf{M}$	the definition of $\mathbf{H}$; note the minus sign
$\nabla \times \mathbf{H} = \mathbf{J}_f, \quad \oint \mathbf{H} \cdot d\boldsymbol{\ell} = I_{f,\text{enc}}$	Ampère's law for $\mathbf{H}$; useful when symmetry allows
$\nabla \cdot \mathbf{H} = -\nabla \cdot \mathbf{M}$	Gauss's law for $\mathbf{H}$: why $\mathbf{H}$ is not determined by the free current alone
$B^{\perp}_{\text{above}} = B^{\perp}_{\text{below}}$, $\mathbf{H}^{\parallel}_{\text{above}} - \mathbf{H}^{\parallel}_{\text{below}} = \mathbf{K}_f \times \hat{\mathbf{n}}$	boundary conditions at an interface
In linear materials only	
$\mathbf{M} = \chi_m \mathbf{H}$	$\chi_m > 0$ for paramagnets, $\chi_m < 0$ for diamagnets
$\mathbf{B} = \mu\mathbf{H}, \quad \mu = \mu_0\mu_r, \quad \mu_r = 1 + \chi_m$	permeability and relative permeability
$\mathbf{J}_b = \chi_m \mathbf{J}_f$	in a homogeneous material; no free current means no bound current inside
$\mathbf{B} = \mu_r \mathbf{B}_{\text{vac}}$	when the material fills all the space where the field is, with the same free currents
$\mu_{\text{above}} H^{\perp}_{\text{above}} = \mu_{\text{below}} H^{\perp}_{\text{below}}$, $\mathbf{B}^{\parallel}_{\text{above}}/\mu_{\text{above}} - \mathbf{B}^{\parallel}_{\text{below}}/\mu_{\text{below}} = \mathbf{K}_f \times \hat{\mathbf{n}}$	the boundary conditions in terms of μ: the first for the magnetic scalar potential, the second for relaxation in A_z
$U = \frac{1}{2}\int \mathbf{B} \cdot \mathbf{H}\, d\tau$	energy stored, the work done in setting up the free currents

Ferromagnets have none of these properties: their magnetization depends on the history of the field as well as on its present value, and it can remain when the field is removed (the hysteresis loop). Above the Curie temperature a ferromagnet becomes a paramagnet.

Which method? Section 14.3 gathers the methods of magnetostatics in one chart.

Exercises

Exercises for 13.1 Dipoles again

Warm-up questions Read Section 13.1, then answer before class

W13.1 Section 13.1.1 says that the commutator of an electric motor reverses the current in its coil every half turn. Suppose the coil were connected so that its current never reversed. Released from rest with **m** perpendicular to **B**, what would the coil do? At what orientation should the commutator reverse the current, and why there?

W13.2 A small magnet, lined up with the field, sits on the axis of a long solenoid, near its middle, where the field is strongest. A classmate expects the force on the magnet to be largest there. What is the force on it there? Where on the axis is the force largest, and which way does it point?

W13.3 A classmate reasons that every electron carries a magnetic moment of about one Bohr magneton, and a drop of water holds about 10^{22} electrons, so water should be strongly magnetic. What has the classmate left out? Which atoms and molecules named in Section 13.1.4 do have net moments, and why?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

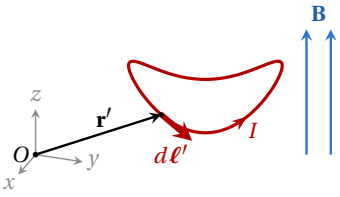


Figure 13.26 The loop of Exercise P13.1: a closed loop of any shape, with a piece $d\boldsymbol{\ell}'$ at $\mathbf{r}'$, in a uniform field **B**.

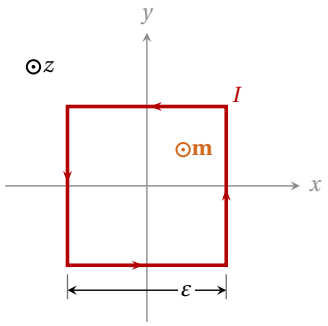


Figure 13.27 The small square loop of Exercise P13.2, of side ε, in the xy plane, with its current counterclockwise as seen from $+z$.

P13.1 A loop of wire of any shape, flat or not, carries a current I in a uniform field **B** (Figure 13.26).

(a) Show that the torque about the origin is $\boldsymbol{\tau} = I\oint \mathbf{r}' \times (d\boldsymbol{\ell}' \times \mathbf{B})$, and use the BAC-CAB rule, (1.49), to write the integrand as $d\boldsymbol{\ell}'\,(\mathbf{r}'\cdot\mathbf{B}) - \mathbf{B}\,(\mathbf{r}'\cdot d\boldsymbol{\ell}')$.

(b) Show that $\oint \mathbf{B}\,(\mathbf{r}'\cdot d\boldsymbol{\ell}') = 0$. (Hint: $\mathbf{r}'\cdot d\boldsymbol{\ell}' = d(r'^2/2)$.)

(c) Use the identity (12.21) to show that $\boldsymbol{\tau} = \mathbf{m}\times\mathbf{B}$, with $\mathbf{m} = I\mathbf{a}$.

(d) A circular coil of 50 turns, each of radius 2.0 cm, carries 0.50 A in a field of 0.20 T. What is the largest torque on it?

P13.2 This exercise shows that the force on a small current loop is $\nabla(\mathbf{m}\cdot\mathbf{B})$ even where the field has a curl. A square loop of side ε lies in the xy plane, centered on the origin, and carries a current I counterclockwise as seen from $+z$ (Figure 13.27), so that $\mathbf{m} = I\varepsilon^2\hat{\mathbf{z}}$. It is in a nonuniform field **B**, which near the origin is $\mathbf{B}(\mathbf{r}) \approx \mathbf{B}_0 + x\,\partial\mathbf{B}/\partial x + y\,\partial\mathbf{B}/\partial y$ in the plane of the loop.

(a) Find the force on each side, to the lowest order in ε that does not cancel, and show that the total force is $\mathbf{F} = m\left(\hat{\mathbf{y}} \times \partial\mathbf{B}/\partial x - \hat{\mathbf{x}} \times \partial\mathbf{B}/\partial y\right)$.

(b) Use $\nabla \cdot \mathbf{B} = 0$ to show that this is $m\nabla B_z = \nabla(\mathbf{m} \cdot \mathbf{B})$. Did you need to assume anything about $\nabla \times \mathbf{B}$?

(c) Show that $(\mathbf{m} \cdot \nabla)\mathbf{B} = m\,\partial\mathbf{B}/\partial z$ differs from this force by $\mathbf{m} \times (\nabla \times \mathbf{B})$.

P13.3 A long straight wire of radius R carries a current I along $+\hat{\mathbf{z}}$, spread uniformly over its cross section (Figure 13.28). A small magnetic dipole $\mathbf{m} = m_0\,\hat{\mathbf{y}}$ sits inside the wire, at the point $(x_0, 0, 0)$ with $0 < x_0 < R$.

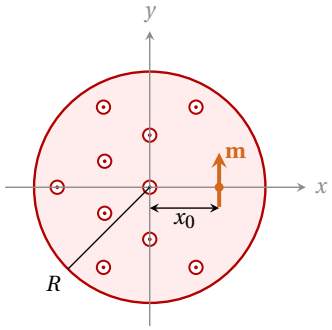


Figure 13.28 The wire of Exercise P13.3, end on, with its current out of the page, and the small dipole $\mathbf{m}$ inside it at $(x_0, 0, 0)$.

(a) Show that inside the wire $\mathbf{B} = (\mu_0 I/2\pi R^2)(-y\,\hat{\mathbf{x}} + x\,\hat{\mathbf{y}})$.

(b) Find $\nabla(\mathbf{m} \cdot \mathbf{B})$ and $(\mathbf{m} \cdot \nabla)\mathbf{B}$ at the dipole. Do they point the same way?

(c) Show that their difference is $\mu_0 \mathbf{m} \times \mathbf{J}$.

(d) If the dipole is a small current loop, which of the two is the force on it (Exercise P13.2)?

(e) Repeat part (b) for a dipole outside the wire, at $(x_0, 0, 0)$ with $x_0 > R$, and show that the two forms agree there.

Exercises for 13.2 Magnetic materials: magnetization and susceptibility

Warm-up questions Read Section 13.2, then answer before class

W13.4 Oxygen gas is paramagnetic, and nitrogen gas is diamagnetic. A sealed bottle of each is heated from 300 K to 600 K in the same applied field. By roughly what factor does the magnetization of the oxygen change? Would you expect that of the nitrogen to change by the same factor? Explain each from the mechanism behind it.

W13.5 A classmate says that diamagnetism is just induction, like the current that a moving magnet induces in a metal ring, and that such currents die away, so diamagnetism must fade once the field stops changing. What does Section 13.2.2 say? And if every material is diamagnetic, why is oxygen called a paramagnet?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P13.4 This exercise derives Curie's law, (13.8), by the method Langevin used in 1905, for a gas of molecules that each have a permanent moment of size m (Figure 13.29).

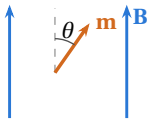


Figure 13.29 A moment $\mathbf{m}$ at the angle θ from the field, as in Exercise P13.4.

(a) In a field $\mathbf{B} = B\hat{\mathbf{z}}$, the energy of a moment at the angle θ from $\hat{\mathbf{z}}$ is $-mB\cos\theta$. In thermal equilibrium at the temperature T, the probability that a moment points within the small solid angle $d\Omega = \sin\theta\, d\theta\, d\phi$ is proportional to $e^{mB\cos\theta/kT}\, d\Omega$. Show that the average of $\cos\theta$ is

$$\langle\cos\theta\rangle = \coth x - \frac{1}{x}, \qquad x = \frac{mB}{kT},$$

the Langevin function.

(b) Show that for $x \ll 1$, $\langle\cos\theta\rangle \approx x/3$, so that a gas with N molecules per unit volume has $M = Nm^2B/3kT$. With $B \approx \mu_0 H$, show that the susceptibility $\chi_m = M/H$ (Section 13.5.1) is $\mu_0 Nm^2/3kT$.

(c) Quantum mechanics replaces m^2 in this result by $g^2S(S+1)\mu_B^2$, where S is the molecule's spin quantum number and $g \approx 2$. For the oxygen molecule $S = 1$. Find χ_m for oxygen gas at 20°C and 1 atmosphere, and compare it with the value in Section 13.5.1.

(d) For a moment of one Bohr magneton at room temperature, how strong would the field have to be to make $x = 1$? Is the approximation of part (b) good for laboratory fields?

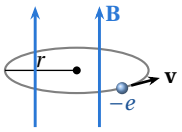


Figure 13.30 The orbit of Exercise P13.5, with the field applied perpendicular to its plane.

P13.5 An electron (charge $-e$, mass m_e) moves at speed v in a circle of radius r around a nucleus, held in its orbit by the electric force (Figure 13.30).

(a) A field B is applied perpendicular to the plane of the orbit. Assume that the radius stays the same, and show that, for a small field, the speed changes by $\Delta v = erB/2m_e$, faster or slower depending on the direction of the electron's motion. (The speed changes while the field is being turned on, as Section 13.2.2 describes; here you need only the final state.)

(b) Show that in either case the orbit's magnetic moment changes by $\Delta\mathbf{m} = -(e^2r^2/4m_e)\mathbf{B}$, (13.9).

(c) In an atom with Z electrons, whose orbits are oriented at random, the average of $x^2 + y^2$ for each electron is $\frac{2}{3}\langle r^2\rangle$. Show that a gas of such atoms, N per unit volume, has the susceptibility (Section 13.5.1) $\chi_m = -\mu_0 NZe^2\langle r^2\rangle/6m_e$.

(d) The simplest quantum-mechanical estimate for helium treats each of its two electrons as a hydrogen-like electron in the ground state, attracted by an effective nuclear charge $Z_{\text{eff}}e$ with $Z_{\text{eff}} = 27/16$, for which $\langle r^2\rangle = 3a_0^2/Z_{\text{eff}}^2$. Estimate χ_m for helium gas at 20°C and 1

atmosphere, and compare it with the best calculated value, -9.88×10^{-10}.[13]

Exercises for 13.3 Bound currents

Warm-up questions Read Section 13.3, then answer before class

W13.6 In Figure 13.4b the magnetization points out of the page and grows toward the top. Suppose instead that it grew toward the right, $\mathbf{M} = M_z(x)\,\hat{\mathbf{z}}$. Which way would the net bound current inside the material run? Find it from the picture of small current loops, and check it with $\nabla \times \mathbf{M}$.

W13.7 A classmate argues that the bound current around the surface of a bar magnet must heat the magnet, as any current heats a wire, so a permanent magnet should slowly run down. Use Section 13.3.1 to explain what the bound current is, and why it does not heat the magnet.

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P13.6 This exercise shows that the magnetic moment of the bound currents of an object, computed with (12.24) and its surface-current version, is the integral of its magnetization:

$$\frac{1}{2}\int_{\mathcal{V}} \mathbf{r} \times \mathbf{J}_b \, d\tau + \frac{1}{2}\oint_{\mathcal{S}} \mathbf{r} \times \mathbf{K}_b \, da = \int_{\mathcal{V}} \mathbf{M} \, d\tau.$$

(a) For a constant vector $\mathbf{c}$, show that $\nabla \times (\mathbf{c} \times \mathbf{r}) = 2\mathbf{c}$.

(b) Show that $\mathbf{c} \cdot [\mathbf{r} \times (\nabla \times \mathbf{M})] = \nabla \cdot [\mathbf{M} \times (\mathbf{c} \times \mathbf{r})] + 2\,\mathbf{c} \cdot \mathbf{M}$, using the product rule for $\nabla \cdot (\mathbf{A} \times \mathbf{B})$.

(c) Integrate over the object, use the divergence theorem, and rearrange the surface term to prove the result.

(d) Check it for a cylinder of radius R and length L, uniformly magnetized along its axis.

[13] Calculated from the susceptibility of a helium-4 atom in "Relativistic treatment of the diamagnetic susceptibility of helium," M. Puchalski, M. Lesiuk, and B. Jeziorski, Phys. Rev. A 108, 042812 (2023), https://doi.org/10.1103/PhysRevA.108.042812, with the ideal-gas density. The measured value in the *CRC Handbook of Chemistry and Physics*, the source of the table in Section 13.5.1, gives -1.06×10^{-9}, about 7% larger in magnitude. The authors of the calculation argue that an error that large in their result is very unlikely, and suggest that a new measurement may resolve the difference.

Exercises for 13.4 Solving for B when given M: bound-current method

Warm-up questions Read Section 13.4, then answer before class

W13.8 Example 13.3 gives the field at the center of a bar magnet of length L and radius R, uniformly magnetized along its axis. What does it approach for a long thin magnet, $L \gg R$, and for a thin disk, $L \ll R$? Explain each limit with the solenoid picture of the bound current, and compare with the field at the center of the polarized cylinder of Example 9.5.

W13.9 Outside, the uniformly magnetized sphere of Example 13.4 and the uniformly polarized sphere of Example 9.3 have fields of the same shape. Inside, **B** points along **M**, but **E** points against **P**. Explain the difference from the sources: where are the bound charges of the one and the bound currents of the other, and which way does each make the field inside point?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

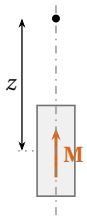


Figure 13.31 The bar magnet of Exercise P13.7 and a point on its axis.

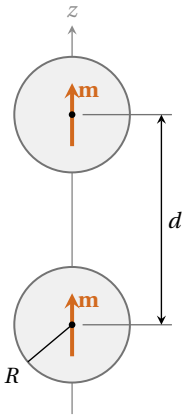


Figure 13.32 The two magnetized spheres of Exercise P13.8, with their moments along the line through their centers, a distance d apart.

P13.7 This exercise asks how far from a bar magnet the dipole approximation can be trusted, for the magnet of Example 13.3 (Figure 13.31).

(a) Expand the field on the axis in powers of $1/z$, and show that

$$B = \frac{\mu_0 M_0 R^2 L}{2z^3}\left[1 + \frac{L^2 - 3R^2}{2z^2} + \cdots\right].$$

(b) For $L = R$, $2R$, $5R$, and $20R$, plot the ratio of the dipole approximation to the exact field on the axis outside the magnet, and find the distance beyond which the dipole approximation is good to 5 percent, and to 1 percent.

(c) Use part (a) to explain why the dipole approximation does better for $L = 2R$ than for $L = R$ or $5R$. For what length does the $1/z^2$ correction vanish?

P13.8 Two spheres of radius R, each uniformly magnetized with moment **m**, lie on the z axis with their centers a distance $d \geq 2R$ apart, and both moments point along $+\hat{\mathbf{z}}$ (Figure 13.32).

(a) Treat each sphere as a point dipole at its center, and use the field on the axis of a dipole, $\mu_0 m/2\pi z^3$, and (13.3) to find the force between them. Do they attract or repel? What if one of them is turned around?

(b) The field outside a uniformly magnetized sphere is exactly a dipole field (Example 13.4). The force on the second sphere is the sum of the forces $\nabla(\mathbf{M}\,d\tau \cdot \mathbf{B}_1)$ on its pieces, where $\mathbf{B}_1$ is the field of the first

sphere. Each Cartesian component of $\mathbf{B}_1$ satisfies Laplace's equation in the second sphere, and so do its derivatives. Use the averaging property of solutions of Laplace's equation (Section 7.1.2) to show that the answer to part (a) is exact. (The average over a ball is an average of averages over concentric spheres.)

(c) Two such spheres, 1.0 cm in diameter and magnetized to $M = 1.0 \times 10^6$ A/m, touch. How large is the force? Compare it with the weight of one sphere, if they are made of a material with density 7.5 g/cm^3.

Exercises for 13.5 The H field and permeability

Warm-up questions Read Section 13.5, then answer before class

W13.10 Classify each of these as free or bound current, with a reason: (a) the current in the copper windings of an electromagnet; (b) the current that runs around the surface of a bar magnet; (c) a beam of electrons in a vacuum tube; (d) the source of the field of a magnetized piece of iron, whose atomic moments are mostly electron spins.

W13.11 In a diamagnet, **M** points opposite to **H**. Use (13.18) and (13.19) to show that **B** still points along **H**. How negative would χ_m have to be for **B** to point against **H**? Does any material in Table 13.1 come close?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

Exercises for 13.6 Ampère's law for H

Warm-up questions Read Section 13.6, then answer before class

W13.12 By Gauss's law for **H**, the poles of a magnet, where $-\nabla\cdot\mathbf{M}$ or $\mathbf{M}\cdot\hat{\mathbf{n}}$ is not zero, are sources of **H**. Where are the poles of (a) the uniformly magnetized sphere; (b) the cylinder of Example 13.5, whose **M** circles its axis; (c) a long cylinder magnetized uniformly across its axis? With no free current anywhere, which of the three must have **H** not zero somewhere?

W13.13 Section 13.6.4 shows that a linear material that fills all the space where there is any field multiplies **B** by μ_r. Yet a sphere of linear material in a uniform applied field $\mathbf{B}_0$ has $\mathbf{B} = 3\mu_r\mathbf{B}_0/(\mu_r+2)$ inside, not $\mu_r\mathbf{B}_0$. Which step of the section's argument fails for the sphere, and why? (Where is the divergence of **H** not zero?)

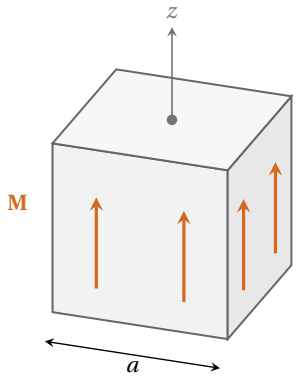


Figure 13.33 The magnetized cube of Exercise P13.9, with the z axis through the centers of its top and bottom faces.

W13.14 The definition $\mathbf{H} = \mathbf{B}/\mu_0 - \mathbf{M}$ has a minus sign where $\mathbf{D} = \varepsilon_0\mathbf{E} + \mathbf{P}$ has a plus sign. Suppose someone defined $\mathbf{H}' = \mathbf{B}/\mu_0 + \mathbf{M}$ instead. Find $\nabla \times \mathbf{H}'$ in terms of $\mathbf{J}_f$ and $\mathbf{J}_b$. Why would $\mathbf{H}'$ be of no help when the bound current is not known?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P13.9 A cube of side a, centered on the origin with its faces perpendicular to the axes, has the uniform magnetization $\mathbf{M} = M_0\hat{\mathbf{z}}$ (Figure 13.33).

(a) Find the bound currents, and sketch them.

(b) What is the field far from the cube?

(c) Use the field on the axis of a square loop, as in Example 13.6, to find $\mathbf{B}$ on the z axis, inside and outside the cube.

(d) Compare your result with the dipole field at $z = a$, $2a$, and $3a$.

(e) Find $\mathbf{H}$ on the z axis inside the cube. Which way does it point?

Exercises for 13.7 Solving for H when given M: Ampère's law for H method (high symmetry)

Warm-up questions Read Section 13.7, then answer before class

W13.15 A long straight wire carrying a current I runs along the axis of a long iron tube, whose bore is wider than the wire; treat the iron as a linear material with relative permeability μ_r. Without calculating, say which of $\mathbf{H}$, $\mathbf{B}$, and $\mathbf{M}$ would be the same without the iron, (a) in the bore, (b) in the iron, and (c) outside the tube. Why is $\mathbf{B}$ outside the tube unchanged?

W13.16 In Example 13.9, $\mathbf{H}$ inside the solenoid is $nI\hat{\mathbf{z}}$ with or without the core. A classmate concludes that a magnetic material never changes $\mathbf{H}$, only $\mathbf{B}$. Give a counterexample from this chapter. What does your counterexample have that the filled solenoid does not?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

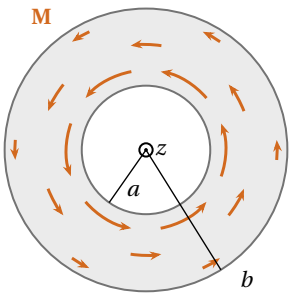


Figure 13.34 Cross section of the tube of Exercise P13.10, with its magnetization circling the axis and weaker farther out.

P13.10 A long thick-walled tube, with inner radius a and outer radius b, has the magnetization $\mathbf{M} = (k/s^2)\hat{\boldsymbol{\phi}}$, where k is a constant (Figure 13.34).

(a) Find the bound currents inside the material and on both of its surfaces, and show that the total bound current along the tube is zero.

(b) Find **B** everywhere from the bound currents and Ampère's law.

(c) Find **B** again with Ampère's law for **H**.

P13.11 A long coaxial cable has an inner conductor of radius a carrying a free current I along $+\hat{\mathbf{z}}$ and a thin outer conductor of radius b carrying the return current (Figure 13.35). The space between them is filled with two linear materials: one with permeability μ_1 in the half $0 < \phi < \pi$, and one with μ_2 in the half $\pi < \phi < 2\pi$.

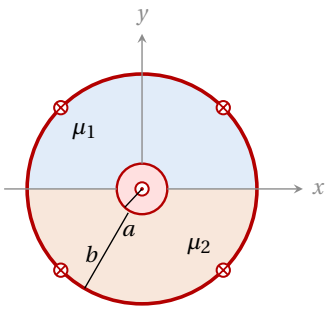


Figure 13.35 Cross section of the cable of Exercise P13.11, with its two fillings.

(a) The two fillings meet on the plane $y = 0$. Use the boundary conditions there to explain why it is **B**, not **H**, that has the same magnitude in the two halves at a given s.

(b) Use Ampère's law for **H** to find **B** and **H** in each half.

(c) Find **M** in each half, and the bound surface current on each half of the inner conductor's surface.

P13.12 A long solenoid of radius R, with n turns per unit length carrying a current I, has a core made of two linear materials: a rod of radius $a < R$ with permeability μ_1 along its axis, inside a tube of permeability μ_2 that fills the rest of the solenoid (Figure 13.36).

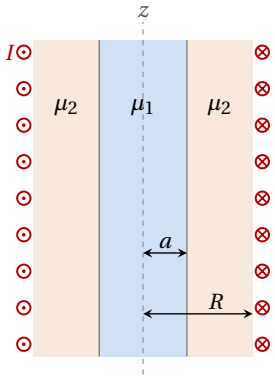


Figure 13.36 The solenoid of Exercise P13.12, cut along its axis, with its two-part core.

(a) Show that $\mathbf{H} = nI\,\hat{\mathbf{z}}$ everywhere inside.

(b) Find **B** and **M** in each material, and the bound surface currents at $s = a$ and $s = R$.

(c) Check that **B** satisfies the boundary conditions of Section 11.3 at $s = a$, with the bound surface current found in part (b).

(d) Find the inductance per unit length.

P13.13 An iron ring of mean circumference ℓ, whose cross section is small compared with its radius, has a narrow gap cut through it, of width g small compared both with ℓ and with the width of the iron (Figure 13.37). It is wound with N turns of wire carrying a current I. Treat the iron as a linear material with relative permeability μ_r.

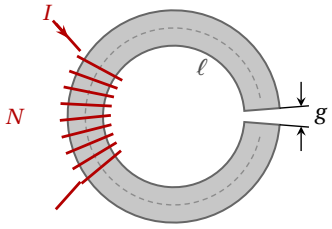


Figure 13.37 The iron ring of Exercise P13.13, with its gap and its winding.

(a) Explain why **B** has nearly the same magnitude in the iron and in the gap.

(b) Use Ampère's law for **H** around the ring to show that

$$B = \frac{\mu_0 N I}{g + (\ell - g)/\mu_r}.$$

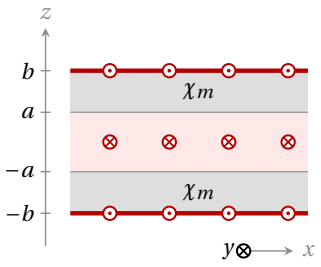


Figure 13.38 The arrangement of Exercise P13.14, edge on: the conductor (center), the two sheets of return current, and the paramagnet (shaded) between them.

(c) Find B for $\mu_r = 1000$, $\ell = 30\,\text{cm}$, $g = 2.0\,\text{mm}$, $N = 500$, and $I = 2.0\,\text{A}$. Compare it with the field the same winding would give with no iron, and with the field it would give with no gap. (What would actually limit the field with no gap?)

(d) What fraction of the free current NI is "used" to drive $\mathbf{H}$ across the gap?

P13.14 A large flat conductor fills the region $-a < z < a$ and carries a uniform free current density $J_0\,\hat{\mathbf{y}}$ (Figure 13.38). Two large thin sheets at $z = \pm b$, with $b > a$, each carry the surface current $-J_0 a\,\hat{\mathbf{y}}$, so that the total current is zero. The regions $a < |z| < b$ are filled with a linear paramagnet of susceptibility χ_m; the conductor itself is not magnetic.

(a) Use symmetry and Ampère's law for $\mathbf{H}$ to find $\mathbf{H}$ everywhere.

(b) Find $\mathbf{B}$ and $\mathbf{M}$ in the paramagnet.

(c) Find the bound currents. Which way do they flow, compared with the free currents next to them?

Exercises for 13.8 Ferromagnets: domains, hysteresis, and the Curie temperature

Warm-up questions Read Section 13.8, then answer before class

W13.17 Would you make each of these from a soft or a hard magnetic material, and why: (a) a refrigerator magnet; (b) the core of the electromagnet on a scrapyard crane, which must drop its load when the current is switched off; (c) a compass needle; (d) the coating of a magnetic tape, which must keep the pattern that a recording head writes on it?

W13.18 Inside a uniformly magnetized sphere, $\mathbf{H} = -\mathbf{M}/3$ (Section 13.6.1). A sphere of ferromagnet is magnetized to saturation, and the applied field is then removed. On the hysteresis loop of Figure 13.17, does the sphere end up at b, or somewhere else? (With no applied field, is $\mathbf{H}$ inside the sphere zero?) Why does a permanent magnet need a large coercive field?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P13.15 This exercise shows why ferromagnetism cannot come from the magnetic forces between atomic moments.

(a) Two moments of one Bohr magneton each are 2.5 Å apart and point along the line joining them. Find their interaction energy, $-\mathbf{m}_2 \cdot \mathbf{B}_1$, in electron volts, and the temperature at which kT equals it.

(b) Compare this with kT at the Curie temperature of iron.

(c) Weiss described the interaction that lines the moments up as an internal "molecular field" B_m. Estimate B_m for iron by setting $\mu_B B_m$ equal to kT_C, and compare it with the fields listed in Section 10.3.

P13.16 Use the table of Section 13.8.3.

(a) As in Example 13.2, find the magnetic moment per atom near absolute zero of cobalt (density 8.86 g/cm^3, molar mass 58.93 g/mol) and of nickel (8.90 g/cm^3, 58.69 g/mol), in Bohr magnetons.

(b) Find $\mu_0 M_s$ for iron at room temperature. A long solenoid with 1000 turns per meter carries a current of 2.0 A. How large is its field without a core? If an iron core could be magnetized to saturation, by what factor would it increase the field inside?

P13.17 A steel needle is magnetized along its length and then heated above the Curie temperature of iron.

(a) What happens to its magnetization?

(b) It is then cooled slowly, lying north–south in the Earth's field. In which direction, if any, is it magnetized when it is cool again? Explain in terms of domains.

(c) Lava that solidifies contains small grains of magnetite. How could the magnetization of old lava flows record the direction of the Earth's field at the time they cooled? (This record is what Néel's theory of rock magnetism explained.)

Exercises for 13.9 Boundary conditions at magnetic interfaces

Warm-up questions Read Section 13.9, then answer before class

W13.19 In part (a) of Example 13.11, **B** is continuous at the top surface and **H** is not. Suppose instead that the slab's frozen-in magnetization were $C\hat{\mathbf{y}}$, parallel to the surface, with **B** just above the surface the same as before. Is **B** continuous now? Is **H**? Find both just inside the slab.

W13.20 Where there is no free current, **H** can be written as $-\nabla\phi_M$. Around a long straight wire carrying a current I, the circulation of **H** around a circle centered on the wire is I. What would the circulation of $-\nabla\phi_M$ around that circle be, if ϕ_M had a single value at each point? So can a single-valued ϕ_M be used around the wire? Can one be used in and around a permanent magnet, where there is no free current anywhere?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P13.18 For the uniformly magnetized sphere of Example 13.4, write $\mathbf{B}$ just inside and just outside the surface, at the polar angle θ, in terms of $\hat{\mathbf{r}}$ and $\hat{\boldsymbol{\theta}}$, and check all four boundary conditions of Section 13.9.

P13.19 The flat surface of a block of iron, treated as a linear material with $\mu_r = 500$, is next to air. Just inside the iron, $\mathbf{B}$ has magnitude 1.2 T and makes an angle of 80° with the normal to the surface.

(a) Find $\mathbf{B}$ just outside, and its angle with the normal.

(b) Find the bound surface current on the surface.

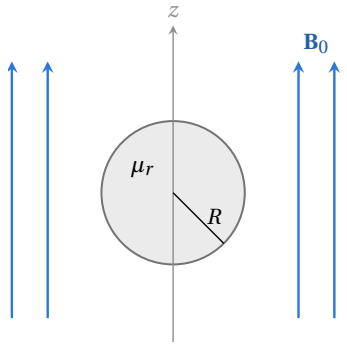


Figure 13.39 The sphere of Exercise P13.20 in a uniform applied field.

P13.20 A sphere of radius R, made of a linear material with relative permeability μ_r, is placed in a uniform applied field $\mathbf{B}_0 = B_0\,\hat{\mathbf{z}}$ (Figure 13.39). There is no free current.

(a) Explain why the magnetic scalar potential can be taken to be $\phi_M = Ar\cos\theta$ inside and $\phi_M = -H_0 r\cos\theta + C\cos\theta/r^2$ outside, with $H_0 = B_0/\mu_0$.

(b) Apply the boundary conditions of Section 13.9.2 to find A and C, and show that the field inside is $\mathbf{B} = 3\mu_r\mathbf{B}_0/(\mu_r+2)$.

(c) Compare with Example 9.11. For a material with $|\chi_m| \ll 1$, show that the field inside differs from $\mathbf{B}_0$ by about $\frac{2}{3}\chi_m\mathbf{B}_0$. By how much is it changed for liquid oxygen? For bismuth?

(d) What is the field inside as $\mu_r \to \infty$? Sketch the field lines of $\mathbf{B}$ in that case.

P13.21 This exercise finds the exact field inside the tube of Example 13.12, in an infinite region.

(a) With no free current, $\mathbf{H} = -\nabla\phi_M$. Using the separation of variables in cylindrical coordinates of Section 7.4.4, explain why ϕ_M can be written as $C_1 s\cos\phi$ inside the tube ($s < a$), $(C_2 s + C_3/s)\cos\phi$ in the wall, and $(-H_0 s + C_4/s)\cos\phi$ outside, with $\mathbf{H}$ along x far away.

(b) Apply the boundary conditions at $s = a$ and $s = b$, and show that $B_{\text{in}}/B_0 = 4\mu_r b^2/[(\mu_r+1)^2 b^2 - (\mu_r-1)^2 a^2]$.

(c) Check the result for $\mu_r = 1$, and show that for a thin wall, with $\mu_r(b-a) \gg b$, it is approximately $2b/\mu_r(b-a)$.

Exercises for 13.10 Numerical methods: direct integration and relaxation

Warm-up questions Read Section 13.10, then answer before class

W13.21 In the rule (13.34) each neighbor is weighted by $1/\mu$, while in the dielectric rule (9.38) each is weighted by ε itself. Explain why: which field does each derivation take around or through the sides of the small square, and how is that field written in terms of the unknown potential?

W13.22 A classmate says that the iron tube of Example 13.12 works the way a wall stops a ball: the iron blocks the field from getting in. What does the text say happens to the field lines instead? Why can a shield of finite permeability never make the field inside exactly zero?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P13.22 Write a program that finds the field of the bar magnet of Example 13.3, with $L = 2R$, at any point, by adding up the fields of a large number of thin rings of current $M_0\,dz'$. Find each ring's field with the elliptic-integral formulas of Exercise P10.22, or by dividing the ring into short pieces.

(a) Check the program on the axis against Example 13.3.

(b) Compare the field at the distance $5R$ from the center, at 90° and at 45° from the axis, with the dipole field (12.27).

(c) Find the field just outside the wrapper, halfway between the ends. Which way does it point?

(d) Make a plot of the field lines in a plane through the axis, inside and outside the magnet.

P13.23 Write a program that solves the grid equations (13.34) for the iron tube of Example 13.12, either by relaxation or with a sparse-matrix solver.

(a) Reproduce the table in the example.

(b) Find the field at the center for $\mu_r = 10$ and $\mu_r = 1000$, with $h = b/40$, and compare with the exact formula of Exercise P13.21.

(c) If you use the Gauss–Seidel method, start with $A_z = B_0 y$ everywhere and count the passes needed to bring the field at the center within 1 percent of its final value, for $\mu_r = 10$ and $\mu_r = 100$. Why does the larger permeability need more passes?

(d) Replace the circular tube by a square one with the same outer width and the same wall thickness (Figure 13.40). Does it shield better or worse? Does the grid error behave differently?

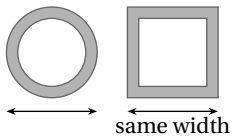


Figure 13.40 The circular tube of Exercise P13.23 and a square one with the same outer width and wall thickness.

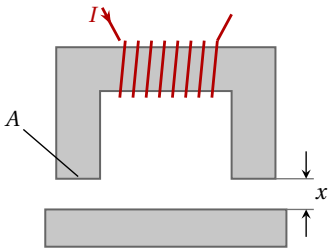


Figure 13.41 The electromagnet of Exercise P13.24, with the bar pulled a small distance x away from its pole faces.

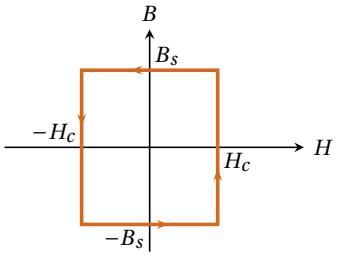


Figure 13.42 An idealized, rectangular hysteresis loop, for Exercise P13.25.

Exercises for 13.11 Energy stored in the field, inside magnetic materials

Warm-up questions Read Section 13.11, then answer before class

W13.23 Example 13.13 shows that filling a solenoid with a linear material multiplies its inductance by μ_r. By what factor would liquid oxygen raise it? An iron core raises it far more. Why, then, is $L = \mu_r \mu_0 n^2 \pi R^2 \ell$ only a rough guide for a coil with an iron core (Section 13.8.2)?

W13.24 Section 13.11.2 says that when an electromagnet lifts a nail, the source of the energy is clear, but when a permanent magnet does, the question is still debated. Where does the energy come from in the first case? Why does the argument that magnetic forces do no work not settle the second?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P13.24 An electromagnet has a U-shaped iron core, and an iron bar lies across its two pole faces, each of area A (Figure 13.41). Suppose the bar is pulled a small distance x away from the faces, leaving two thin air gaps, in each of which the field B is uniform and perpendicular to the faces.

(a) Treat the iron as having a very large permeability, so that the field energy stored in it is negligible. If the flux is held fixed as the bar moves a further dx, by how much does the energy stored in the gaps change? Use this to show that the force holding the bar is $F = B^2 A/\mu_0$, or $B^2/2\mu_0$ per unit area of each face.

(b) How large is the force for $B = 1.0\,\text{T}$ and $A = 4.0\,\text{cm}^2$?

(c) In practice the current, not the flux, is held fixed. Why does that not change the force? (Section 9.10.1 faced the same question.)

P13.25 The hysteresis loop of a soft iron alloy can be approximated by a rectangle in the B–H plane, with corners at $H = \pm H_c$ and $B = \pm B_s$ (Figure 13.42).

(a) Find the energy lost per unit volume in one trip around the loop, for $H_c = 40\,\text{A/m}$ and $B_s = 1.5\,\text{T}$.

(b) The core of a transformer is made of 1.0 liter of this alloy and goes around its loop 60 times a second. How much power is lost?

(c) Repeat part (b) for a material used in permanent magnets, whose loop is close to a rectangle in a plot of $\mu_0 M$ against H, with corners at $H = \pm H_c$ and $\mu_0 M = \pm 1.2\,\text{T}$, where $H_c = 8 \times 10^5\,\text{A/m}$. (Why can its loop in the B–H plane not be a rectangle, and why does it enclose the same area?) Explain why such materials are not used for transformer cores.

Exercises for 13.12 Summary

P13.26 True or false? Explain your reasoning.

(a) A current flows along $+\hat{\mathbf{z}}$ in a long wire of paramagnetic material. The bound currents in the wire also flow along $+\hat{\mathbf{z}}$.

(b) $\mathbf{H}$ inside a cube-shaped permanent magnet is zero.

(c) The magnetization is continuous across the surface of a magnetized cube.

(d) The divergence of $\mathbf{H}$ is always zero.

(e) Field lines of $\mathbf{H}$ never begin or end.

(f) Inside a permanent magnet, $\mathbf{B} = \mu_0\mu_r\mathbf{H}$, with μ_r a constant of the material.

(g) In a homogeneous linear material that carries no free current, there is no bound volume current.

(h) The field inside a long solenoid filled with copper is slightly weaker than without the copper.

(i) A diamagnet is pushed away from regions of strong field.

(j) The magnetic susceptibility of iron is about 10^5.

(k) Above its Curie temperature, iron is paramagnetic.

(l) The torque on a magnetic dipole in a uniform field tends to turn it toward the direction of the field.

P13.27 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as ferromagnetic domains, force between magnets, magnetic fields partially shielded by high-μ materials, the $\mathbf{B}$ and $\mathbf{H}$ fields in and around a bar magnet, the levitation of a diamagnetic object in a strong magnetic field, or the loss of ferromagnetism as a material is heated to its Curie temperature.

Chapter 14

Review of Electro- and Magnetostatics

By this point you have built up two large collections of equations, with experience in how to use them: one set for the electric field and related quantities, and the other for the magnetic field and its related quantities. The equations were developed separately, motivated by different experimental observations and starting points (e.g. Coulomb's law versus the Biot–Savart law), and were discussed several chapters apart. It would be easy to finish magnetostatics with the impression that you have learned two unrelated subjects that just happen to share a textbook.

You have not! When we look at the main equations of both halves side by side, the correspondences are close enough to be startling (Figure 14.1). Nearly every electric equation has a magnetic partner of the same shape; where a partner is missing, the absence itself is informative. This chapter reviews the two subjects together. It puts the two columns next to each other so you can see how similar (or different) the corresponding equations are, discusses what the patterns are telling you (Section 14.2), and gathers the methods of both subjects into two charts that help you think through problem-solving starting points (Section 14.3). The exercises at the end are review problems. Some ask which method you would use, and why; others give a problem from one field and ask for the equivalent problem for the other, the rule that makes the two equivalent, and the key difference between them.

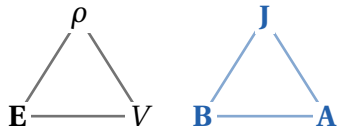


Figure 14.1 The triangle diagrams of Chapters 5 and 12, reduced to their corners: ρ, **E**, V and **J**, **B**, **A** play the same roles.

14.1 The two columns side by side

14.1.1 Statics

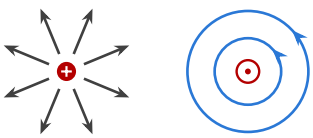
Table 14.1 starts with the sources, the forces, and the fields that given sources produce. The field equations and the potentials that follow from them are in Table 14.2. Then Table 14.3 deals with energy, capacitance, inductance, and the boundary conditions. Solid boxes mark Maxwell's equations in their final form; dashed boxes mark static forms, which Table 14.6 generalizes.

	ELECTRIC	MAGNETIC
1	$q = \int \lambda \, d\ell$ $q = \int \sigma \, da$ $q = \int \rho \, d\tau$	$I = \int K_\perp \, d\ell$ $I = \int \mathbf{J} \cdot d\mathbf{a}$
2	$\mathbf{F} = \frac{qQ}{4\pi\varepsilon_0} \frac{\boldsymbol{\imath}}{\imath^3}$	*No easy parallel.*
3	$\mathbf{F} = Q\mathbf{E}$	$\mathbf{F} = Q(\mathbf{v} \times \mathbf{B})$
4	$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \sum_i \frac{q_i \boldsymbol{\imath}_i}{\imath_i^3}$	
	$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \int \frac{\lambda(\mathbf{r}') \boldsymbol{\imath}}{\imath^3} \, d\ell'$	$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int \frac{I \, d\boldsymbol{\ell}' \times \boldsymbol{\imath}}{\imath^3}$
	$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \int_S \frac{\sigma(\mathbf{r}') \boldsymbol{\imath}}{\imath^3} \, da'$	$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int_S \frac{\mathbf{K}(\mathbf{r}') \times \boldsymbol{\imath}}{\imath^3} \, da'$
	$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \int_V \frac{\rho(\mathbf{r}') \boldsymbol{\imath}}{\imath^3} \, d\tau'$	$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int_V \frac{\mathbf{J}(\mathbf{r}') \times \boldsymbol{\imath}}{\imath^3} \, d\tau'$

Table 14.1 Charge and current, the forces, and the fields of given sources.

	ELECTRIC	MAGNETIC
5	$\boxed{\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0}}$	$\boxed{\nabla \cdot \mathbf{B} = 0}$
6	$\Phi_E = \int_S \mathbf{E} \cdot d\mathbf{a}$	$\Phi_B = \int_S \mathbf{B} \cdot d\mathbf{a}$
7	$\oint_S \mathbf{E} \cdot d\mathbf{a} = \frac{q_{\text{enc}}}{\varepsilon_0}$	$\oint \mathbf{B} \cdot d\boldsymbol{\ell} = \mu_0 I_{\text{enc}}$
8	$\boxed{\nabla \times \mathbf{E} = 0}$ (modified below)	$\boxed{\nabla \times \mathbf{B} = \mu_0 \mathbf{J}}$ (modified below)
9	$\mathbf{E} = -\nabla V$ (modified in Section 15.4)	$\mathbf{B} = \nabla \times \mathbf{A}, \quad \nabla \cdot \mathbf{A} = 0$ (Coulomb gauge)
10	$V(\mathbf{r}) = -\int_{\mathcal{O}}^{\mathbf{r}} \mathbf{E} \cdot d\boldsymbol{\ell}$	$\oint \mathbf{A} \cdot d\boldsymbol{\ell} = \Phi_B$ (the flux method; a partial parallel, Section 14.2)
11	$V(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \sum_i \frac{q_i}{\imath_i}$	
	$V(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \int \frac{\lambda(\mathbf{r}')}{\imath} \, d\ell'$	$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int \frac{I}{\imath} \, d\boldsymbol{\ell}'$
	$V(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \int \frac{\sigma(\mathbf{r}')}{\imath} \, da'$	$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int \frac{\mathbf{K}(\mathbf{r}')}{\imath} \, da'$
	$V(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \int \frac{\rho(\mathbf{r}')}{\imath} \, d\tau'$	$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int \frac{\mathbf{J}(\mathbf{r}')}{\imath} \, d\tau'$

Table 14.2 The field equations and the potentials (Figure 14.2).



(a) **E** (b) **B**

Figure 14.2 (a) The electric field of a charge spreads out from it: divergence, no curl. (b) The magnetic field of a current circles it: curl, no divergence.

	ELECTRIC	MAGNETIC
12	$U = \frac{1}{2}\sum_i q_i V(\mathbf{r}_i)$	
	$U = \frac{1}{2}\int \rho(\mathbf{r}')\, V(\mathbf{r}')\, d\tau'$	$U = \frac{1}{2}\int \mathbf{J}(\mathbf{r}')\cdot\mathbf{A}(\mathbf{r}')\, d\tau'$
	$U = \frac{\varepsilon_0}{2}\int E^2\, d\tau$	$U = \frac{1}{2\mu_0}\int B^2\, d\tau$
13	$C = \frac{Q}{\Delta V}$	$L = \frac{\Phi}{I}$
14	$U = \frac{Q^2}{2C} = \frac{1}{2}C\,\Delta V^2$	$U = \frac{1}{2}LI^2$
15	$E^{\perp}_{\text{above}} - E^{\perp}_{\text{below}} = \frac{\sigma}{\varepsilon_0}$	$B^{\perp}_{\text{above}} = B^{\perp}_{\text{below}}$
16	$\mathbf{E}^{\parallel}_{\text{above}} = \mathbf{E}^{\parallel}_{\text{below}}$	$\mathbf{B}^{\parallel}_{\text{above}} - \mathbf{B}^{\parallel}_{\text{below}} = \mu_0\mathbf{K}\times\hat{\mathbf{n}}$
17	$V_{\text{above}} = V_{\text{below}}$	$\mathbf{A}_{\text{above}} = \mathbf{A}_{\text{below}}$
18	$\frac{\partial V_{\text{above}}}{\partial n} - \frac{\partial V_{\text{below}}}{\partial n} = -\frac{\sigma}{\varepsilon_0}$	$\frac{\partial \mathbf{A}_{\text{above}}}{\partial n} - \frac{\partial \mathbf{A}_{\text{below}}}{\partial n} = -\mu_0\mathbf{K}$
19	$\nabla^2 V = -\frac{\rho}{\varepsilon_0}$	$\nabla^2\mathbf{A} = -\mu_0\mathbf{J}$

Table 14.3 Energy, capacitance and inductance, and the boundary conditions (Figure 14.3). The last three equations for $\mathbf{A}$ hold in the Coulomb gauge; $\mathbf{A}^{\parallel}$ is continuous in any gauge.

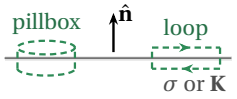


Figure 14.3 The two arguments behind the boundary conditions: a flat pillbox straddling a surface charge or current gives the perpendicular components, and a thin loop gives the parallel ones. The normal $\hat{\mathbf{n}}$ points toward the side called "above."

The multipole expansions are too wide to set side by side, but they correspond term for term, with α the angle between $\mathbf{r}$ and $\mathbf{r}'$ (Sections 8.3.2 and 12.6):

$$V(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\sum_{n=0}^{\infty}\frac{1}{r^{n+1}}\int (r')^n P_n(\cos\alpha)\,\rho(\mathbf{r}')\, d\tau'$$

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi}\sum_{n=0}^{\infty}\frac{1}{r^{n+1}}\int (r')^n P_n(\cos\alpha)\,\mathbf{J}(\mathbf{r}')\, d\tau'$$

For a loop of wire, $\mathbf{J}\,d\tau'$ becomes $I\,d\boldsymbol{\ell}'$. The $n = 0$ term of the second expansion, the magnetic counterpart of the total charge, is always zero for steady currents (Section 12.6.1).

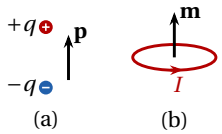


Figure 14.4 (a) An electric dipole, two opposite charges, with $\mathbf{p}$ pointing from $-q$ to $+q$, i.e. the $\hat{\mathbf{z}}$ direction. (b) A magnetic dipole, a small loop of current, with $\mathbf{m} = I\mathbf{a}$ given by the right-hand rule, also pointing in the $\hat{\mathbf{z}}$ direction. The linked animation depicts the fields from the two dipoles, showing how they both give rise to the same far-field dipole pattern despite the differences in V_{dip} and $\mathbf{A}_{\text{dip}}$, and in the two "field from potential" equations.

14.1.2 In materials

Table 14.4 gives the dipoles (Figure 14.4) and the equations that hold in any material. Table 14.5 gives those that hold in linear materials only.

	ELECTRIC	MAGNETIC
20	$\mathbf{p} = \sum_i \mathbf{r}'_i q_i$ $\mathbf{p} = \int \mathbf{r}' \rho(\mathbf{r}')\, d\tau'$	$\mathbf{m} = I \int d\mathbf{a} = I\mathbf{a}$ $\mathbf{m} = \frac{1}{2}\int \mathbf{r}' \times \mathbf{J}(\mathbf{r}')\, d\tau'$
21	$V_{\text{dip}} = \frac{1}{4\pi\varepsilon_0}\frac{\mathbf{p}\cdot\hat{\mathbf{r}}}{r^2}$	$\mathbf{A}_{\text{dip}} = \frac{\mu_0}{4\pi}\frac{\mathbf{m}\times\hat{\mathbf{r}}}{r^2}$
22	$\mathbf{E}_{\text{dip}}(r,\theta) = \frac{p}{4\pi\varepsilon_0 r^3}\left(2\cos\theta\,\hat{\mathbf{r}} + \sin\theta\,\hat{\boldsymbol{\theta}}\right)$	$\mathbf{B}_{\text{dip}}(r,\theta) = \frac{\mu_0 m}{4\pi r^3}\left(2\cos\theta\,\hat{\mathbf{r}} + \sin\theta\,\hat{\boldsymbol{\theta}}\right)$
23	$\boldsymbol{\tau} = \mathbf{p}\times\mathbf{E}$ ($\boldsymbol{\tau}$ is torque here)	$\boldsymbol{\tau} = \mathbf{m}\times\mathbf{B}$
24	$\mathbf{F} = (\mathbf{p}\cdot\nabla)\mathbf{E} = \nabla(\mathbf{p}\cdot\mathbf{E})$	$\mathbf{F} = \nabla(\mathbf{m}\cdot\mathbf{B})$ (equal to $(\mathbf{m}\cdot\nabla)\mathbf{B}$ only where $\mathbf{J}=0$)
25	$U = -\mathbf{p}\cdot\mathbf{E}$	$U = -\mathbf{m}\cdot\mathbf{B}$ ($\mathbf{m}$ held fixed)
26	$\mathbf{P}$ = electric dip. moment/volume	$\mathbf{M}$ = magnetic dip. moment/volume
27	$\mathbf{p} = \int \mathbf{P}(\mathbf{r}')\, d\tau'$	$\mathbf{m} = \int \mathbf{M}(\mathbf{r}')\, d\tau'$
28	$V_{\text{pol. object}} = \frac{1}{4\pi\varepsilon_0}\int \frac{\mathbf{P}(\mathbf{r}')\cdot\boldsymbol{\mathscr{r}}}{\mathscr{r}^3}\, d\tau'$	$\mathbf{A}_{\text{magn. object}} = \frac{\mu_0}{4\pi}\int \frac{\mathbf{M}(\mathbf{r}')\times\boldsymbol{\mathscr{r}}}{\mathscr{r}^3}\, d\tau'$
29	$\sigma_b = \mathbf{P}\cdot\hat{\mathbf{n}}$	$\mathbf{K}_b = \mathbf{M}\times\hat{\mathbf{n}}$
30	$\rho_b = -\nabla\cdot\mathbf{P}$	$\mathbf{J}_b = \nabla\times\mathbf{M}$
31	$\mathbf{D} = \varepsilon_0\mathbf{E} + \mathbf{P}$	$\mathbf{H} = \mathbf{B}/\mu_0 - \mathbf{M}$
32	$\boxed{\nabla\cdot\mathbf{D} = \rho_f}$	$\boxed{\nabla\cdot\mathbf{B} = 0}$ (still)
33	$\oint \mathbf{D}\cdot d\mathbf{a} = q_{f,\text{enc}}$	$\oint \mathbf{H}\cdot d\boldsymbol{\ell} = I_{f,\text{enc}}$
34	$\boxed{\nabla\times\mathbf{E} = 0}$ (still; modified below)	$\boxed{\nabla\times\mathbf{H} = \mathbf{J}_f}$ (modified below)
35	$D^{\perp}_{\text{above}} - D^{\perp}_{\text{below}} = \sigma_f$	$\mathbf{H}^{\parallel}_{\text{above}} - \mathbf{H}^{\parallel}_{\text{below}} = \mathbf{K}_f\times\hat{\mathbf{n}}$

Table 14.4 Dipoles, and the equations that hold in any material (Figure 14.4).

	ELECTRIC	MAGNETIC
36	$\mathbf{P} = \varepsilon_0\chi_e\mathbf{E}$	$\mathbf{M} = \chi_m\mathbf{H}$
37	$\mathbf{D} = \varepsilon\mathbf{E}, \quad \varepsilon = \varepsilon_0\varepsilon_r$	$\mathbf{B} = \mu\mathbf{H}, \quad \mu = \mu_0\mu_r$
38	$\varepsilon_r = 1 + \chi_e$	$\mu_r = 1 + \chi_m$
39	$U = \frac{1}{2}\int \mathbf{D}\cdot\mathbf{E}\, d\tau$	$U = \frac{1}{2}\int \mathbf{B}\cdot\mathbf{H}\, d\tau$

Table 14.5 Equations that hold in linear materials only (Figure 14.5).

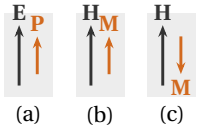


Figure 14.5 In linear materials: (a) in a dielectric **P** points along **E**; (b) in a paramagnet **M** points along **H**; (c) in a diamagnet it points the other way.

14.1.3 Dynamics

The last table, Table 14.6, looks ahead. Faraday's law and the displacement current are the subjects of Chapters 15 and 16 (Sections 15.2 and 16.1). The continuity equation appeared in Section 10.4.1, (10.12), and is treated in full in Section 20.1.

	ELECTRIC	MAGNETIC
40	$\nabla \cdot \mathbf{J} = -\dfrac{\partial \rho}{\partial t}$	$\nabla \cdot \mathbf{J} = -\dfrac{\partial \rho}{\partial t}$
41	$\boxed{\nabla \times \mathbf{E} = -\dfrac{\partial \mathbf{B}}{\partial t}}$	$\boxed{\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \mu_0 \varepsilon_0 \dfrac{\partial \mathbf{E}}{\partial t}}$
		$\boxed{\nabla \times \mathbf{H} = \mathbf{J}_f + \dfrac{\partial \mathbf{D}}{\partial t}}$ (for materials)

Table 14.6 The two curl equations when the fields change. The continuity equation (row 40) is the same equation in both columns — it is what connects charge to current.

14.2 What the pattern is telling you

14.2.1 The integrals are the same integral

Look at the four expressions for **E** in terms of a charge distribution and the three for **B** in terms of a current distribution (row 4). Strip away the constants and they are the same object: a source, divided by the cube of the separation, multiplied by the separation vector $\boldsymbol{\mathscr{r}}$, integrated over wherever the source lives (Figure 14.6). The same is true for V and **A** (row 11), with $\mathscr{r}$ to the first power instead.

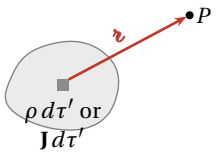


Figure 14.6 The setup shared by every integral in rows 4 and 11: a piece of the source, $\rho\, d\tau'$ or $\mathbf{J}\, d\tau'$, and the separation vector $\boldsymbol{\mathscr{r}}$ from it to the field point P.

This is not a coincidence of notation. It is why the "script-$\mathscr{r}$ method" you learned in electrostatics transferred to magnetostatics with so little retraining. If you can set up the integral for **E** given ρ, you can set up the integral for **B** given **J**, and the only genuinely new thing to keep track of is a cross product. The other methods line up nearly one for one as well, as the two method charts of Section 14.3 show.

That cross product is the first real difference, and it is worth being precise about where it comes from. The fields are really ways of writing the *forces* between charges. Stationary like charges will always repel (Coulomb's law), but *moving* like charges will both repel *and attract*, if they form currents parallel to each other in the same direction (Biot–Savart law). The cross products in the Biot–Savart law and in $\mathbf{F} = Q(\mathbf{v} \times \mathbf{B})$ are the mathematical way to get the correct directions of the forces on moving charges.

14.2.2 Divergence and curl trade places

The deepest structural difference in all of the statics tables is two lines long:

$$\nabla\cdot\mathbf{E}=\frac{\rho}{\varepsilon_0},\qquad \nabla\times\mathbf{E}=0,$$
$$\nabla\cdot\mathbf{B}=0,\qquad \nabla\times\mathbf{B}=\mu_0\mathbf{J}.$$

The static electric field has sources and no circulation; the static magnetic field has circulation and no sources.

! **E** has divergence but no curl. **B** has curl but no divergence. Nearly every difference between the two columns is a consequence of this one swap.

This is why **E** gets a *scalar* potential and **B** gets a *vector* one. A curl-free field can be written as the gradient of a scalar; a divergence-free field can be written as the curl of a vector. The potentials are different kinds of object because, despite their other similarities, the fields have fundamentally different kinds of structure. And yet despite the fundamentally different structure of the fields, in the Coulomb gauge the *potentials* both end up satisfying Poisson's equation so that, aside from constants, V depends on ρ in exactly the same way that the Cartesian components of **A** depend on the Cartesian components of **J**. That is remarkable to me!

It is also why two of the entries in the table have only partial partners (rows 2 and 10). Row 2: There is no simple magnetic Coulomb's law giving the force between two sources — there is no such thing as a steady current element sitting alone in space since steady currents flow in closed loops. And row 10: the potential at a point is a line integral of **E**, $V(\mathbf{r})=-\int_{\mathcal{O}}^{\mathbf{r}}\mathbf{E}\cdot d\boldsymbol{\ell}$, and that works only because the line integral of a curl-free field does not depend on the path. The line integral of **B** *does* depend on the path, so no line integral of **B** gives the value of **A** at a point. What takes its place is the flux method of Section 12.4, $\oint\mathbf{A}\cdot d\boldsymbol{\ell}=\Phi_B$. That gives the circulation of **A** around a loop rather than its value at a point, and so it gives **A** itself only when symmetry makes **A** constant along the loop.

Behind all of this is the experimental fact that magnetic monopoles have never been discovered. If they exist, $\nabla\cdot\mathbf{B}=0$ would gain a right-hand side (magnetic monopole density), as would also $\nabla\times\mathbf{E}=0$ (magnetic monopole current), and the two sets of statics laws would become mathematically equivalent aside from a sign difference in the curl equations.[1]

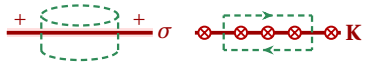
[1] Oliver Heaviside wrote the equations in this symmetric form, with a fictitious magnetic charge and magnetic current beside the electric ones, in "On the forces, stresses, and fluxes of energy in the electromagnetic field," O. Heaviside, Phil. Trans. R. Soc. Lond. A 183, 423–480 (1892), https://doi.org/10.1098/rsta.1892.0011. See also "Faraday's law in the presence of magnetic monopoles," M. Nowakowski and N. G. Kelkar, Europhys. Lett. 71, 346–351 (2005), https://doi.org/10.1209/epl/i2004-10545-2, for a discussion in more modern notation.

14.2.3 Boundary conditions swap

The boundary conditions repeat the same reversal in a form you will use constantly when solving problems:

$$E^{\perp}_{\text{above}} - E^{\perp}_{\text{below}} = \frac{\sigma}{\varepsilon_0}, \qquad \mathbf{E}^{\parallel}_{\text{above}} = \mathbf{E}^{\parallel}_{\text{below}},$$
$$B^{\perp}_{\text{above}} = B^{\perp}_{\text{below}}, \qquad \mathbf{B}^{\parallel}_{\text{above}} - \mathbf{B}^{\parallel}_{\text{below}} = \mu_0 \mathbf{K} \times \hat{\mathbf{n}}.$$

The electric field has a discontinuity in its perpendicular component and is continuous in its parallel one. The magnetic field is exactly the reverse. So if you remember the conditions for **E**, you can practically recover those for **B** by exchanging $\perp$ and $\parallel$. Both pairs come from the same two arguments: a thin pillbox, which tests the divergence, and a thin loop, which tests the curl (Sections 4.4, 5.2, and 11.3; Figure 14.7). The field whose divergence is nonzero at the surface jumps in its perpendicular component, and the field whose curl is nonzero there jumps in its parallel one.



(a) pillbox (b) loop

Figure 14.7 A reminder of the two arguments behind the boundary conditions: (a) a thin pillbox straddling a sheet of charge, and (b) a thin loop straddling a sheet of current.

14.2.4 The one sign that is not symmetric

In materials the correspondence between equations is nearly perfect. Polarization **P** sits opposite magnetization **M**; bound charges sit opposite bound currents; the auxiliary fields **D** and **H** were introduced for the same reason, to let you write a Gauss's or Ampère's law involving only the *free* sources that you actually control rather than bound sources which depend on the materials' responses. Also, although sometimes overlooked, neither **D** nor **H** is actually determined solely by those free sources — since $\nabla \times \mathbf{D} = \nabla \times \mathbf{P}$ (Section 9.6.1) and $\nabla \cdot \mathbf{H} = -\nabla \cdot \mathbf{M}$ (Section 13.6.1), curls of polarization and divergences of magnetization will also serve as sources of **D** and **H**, respectively.

The key difference between **D** and **H** is found in their definitions:

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P}, \qquad \mathbf{H} = \frac{\mathbf{B}}{\mu_0} - \mathbf{M}.$$

One involves the sum of the main field (**E** or **B**) and the material's response (**P** or **M**), whereas the other involves the difference.

The distinction is not a convention. It comes from the bound sources: $\rho_b = -\nabla \cdot \mathbf{P}$ has a minus sign, and $\mathbf{J}_b = \nabla \times \mathbf{M}$ does not. Gauss's law with the bound charge written out is $\nabla \cdot (\varepsilon_0 \mathbf{E}) = \rho_f - \nabla \cdot \mathbf{P}$, so moving the bound charge to the left side puts $+\mathbf{P}$ next to $\varepsilon_0 \mathbf{E}$. Ampère's law with the bound current written out is $\nabla \times (\mathbf{B}/\mu_0) = \mathbf{J}_f + \nabla \times \mathbf{M}$, so moving the bound current to the left side puts $-\mathbf{M}$ next to $\mathbf{B}/\mu_0$. Physically speaking, inside an electric dipole the field points *against* **p**, but inside a magnetic dipole it points *along* **m** (Section 12.6.4).

! $\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P}$ but $\mathbf{H} = \mathbf{B}/\mu_0 - \mathbf{M}$. The signs really are different!

14.2.5 Time dependence: similarities and differences

Table 14.6 is different than the others. In the static tables each column stands on its own: charges make **E**, currents make **B**, and neither field depends on the other.

However, once the fields change in time, that is no longer true. As we will see in Chapters 15 and 16, Faraday's law and Maxwell's correction to Ampère's law say that an **E** or **B** field changing in time gives a curl to the opposite field:

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \qquad \nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \mu_0 \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t}.$$

In empty space, where ρ and **J** are zero, the two curl equations are nearly mirror images of each other. They differ by the factor $\mu_0\varepsilon_0$ and by a sign. Taking the curl of either one, and using the other and the fact that both fields have zero divergence in empty space, gives a wave equation for each field,

$$\nabla^2 \mathbf{E} = \mu_0 \varepsilon_0 \frac{\partial^2 \mathbf{E}}{\partial t^2}, \qquad \nabla^2 \mathbf{B} = \mu_0 \varepsilon_0 \frac{\partial^2 \mathbf{B}}{\partial t^2},$$

whose solutions are waves that travel at $1/\sqrt{\mu_0\varepsilon_0}$, which equals the speed of light c. The coupling of the two columns in row 41 (Figure 14.8) is likely the most consequential line in the entire set of tables, and the fact that Maxwell's equations predict the existence of waves traveling at a speed connected to the constants ε_0 and μ_0 — originally established through static measurements! — is another completely incredible thing to me. Additionally, the change to Ampère's law for materials, $\nabla \times \mathbf{H} = \mathbf{J}_f + \partial \mathbf{D}/\partial t$, leads to the prediction of refraction and damping of electromagnetic waves in insulators and conductors, respectively. Much of the second half of this textbook (Chapters 21–24, and 28) will be devoted to the study of the wave phenomena predicted by these equations.

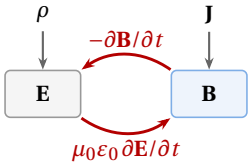


Figure 14.8 In statics, ρ makes **E** and **J** makes **B**, and the two columns stand apart. When the fields change, a changing **B** gives **E** a curl (Faraday's law) and a changing **E** gives **B** one (Maxwell's term), with opposite signs.

14.3 Which method?

Chapters 3–13 introduced the methods one at a time, each in the chapter whose subject needed it. However, physics education research has shown that students often have trouble identifying which method to use when a problem does not announce it — for example, when problems are not grouped by chapter, as on a final exam.[2] To help with those decisions, Figures 14.9 and 14.10 gather the problem-solving methods of electro- and magnetostatics into two flowcharts, illustrating what method to use when.

The two charts have nearly the same shape, as Section 14.2 would lead you to expect. Both ask first about symmetry (Gauss's law or Ampère's law), then about sources that are not known in advance (charge on conductors or bound in a dielectric, current bound in a magnetic material), then about the field far away (the multipole expansion), and both end with the script-$\mathscr{r}$ integrals, done by hand or numerically. Figure 14.10 has no box for the method of images, which the magnetostatics chapters did not employ; Exercise P14.4 shows one place where it does apply. Figure 16.17 presents a similar chart for fields that change in time.

[2] "Analytic framework for students' use of mathematics in upper-division physics," B. R. Wilcox, M. D. Caballero, D. A. Rehn, and S. J. Pollock, Phys. Rev. ST Phys. Educ. Res. 9, 020119 (2013), https://doi.org/10.1103/PhysRevSTPER.9.020119.

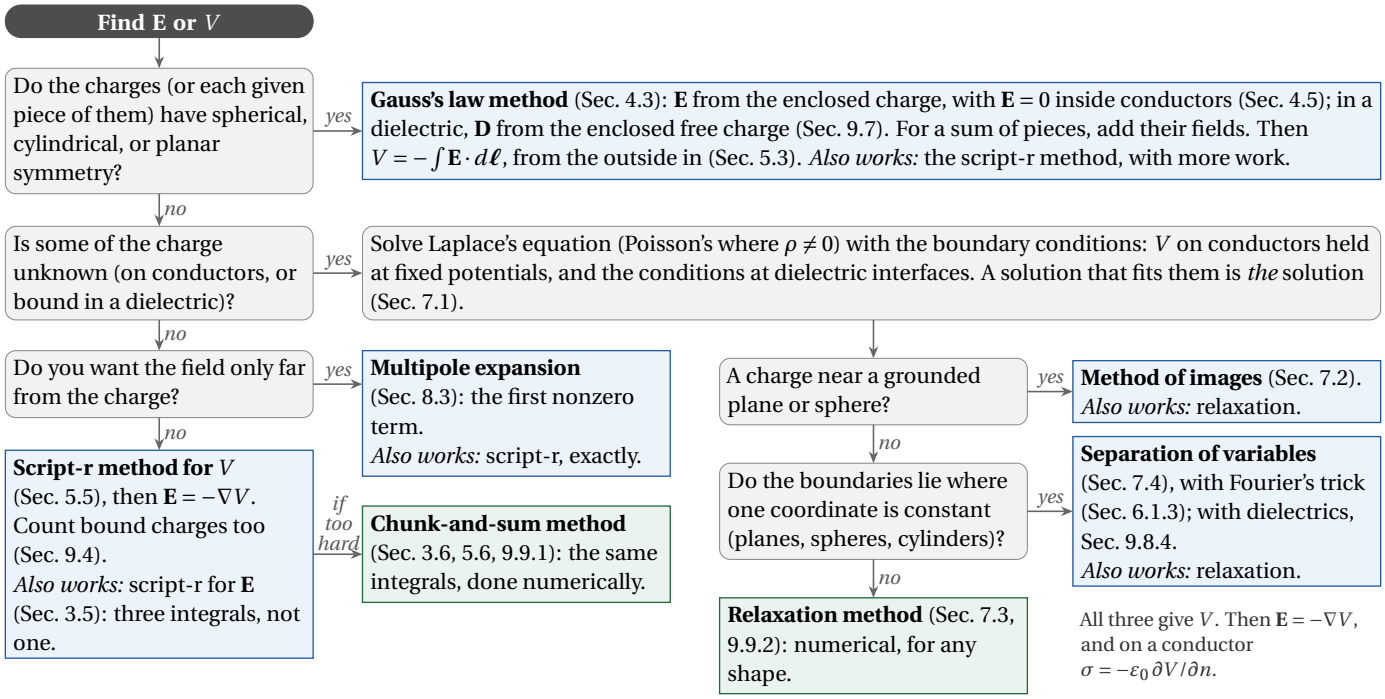


Figure 14.9 Which method? The methods of Chapters 3–9 for finding **E** or V for charges at rest. Gray: questions; blue: methods worked by hand; green: numerical methods. "Also works" names a second method that usually takes longer.

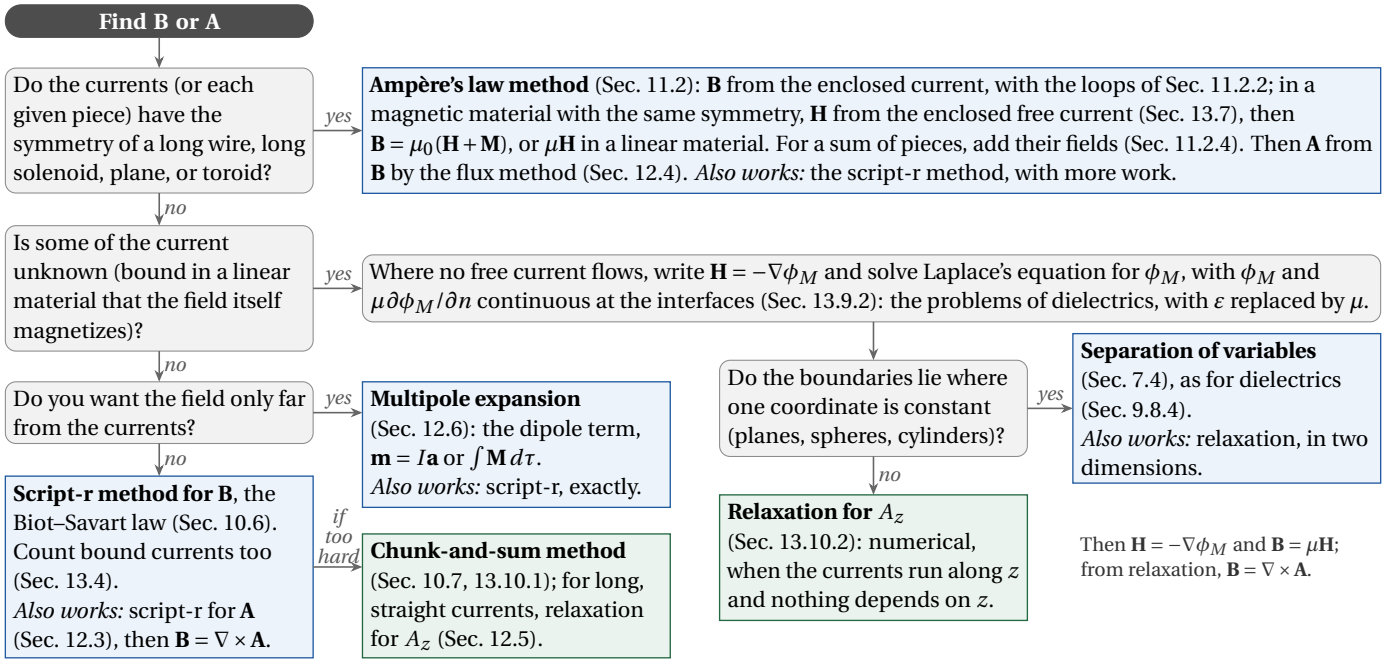


Figure 14.10 Which method? The methods of Chapters 10–13 for finding **B** or **A** in magnetostatics. Gray: questions; blue: methods worked by hand; green: numerical methods. "Also works" names a second method that usually takes longer.

Exercises

Exercises for 14.2 What the pattern is telling you

P14.1 A long, straight copper pipe, with inner radius a and outer radius b, carries a steady current I along its length, spread uniformly over the copper.

(a) Describe an electrostatics problem that is equivalent to finding **B** everywhere for the pipe: give its charge density, and say which law you would use and with what surface.

(b) Name the rule that makes the two problems equivalent, and the one key difference between their fields.

(c) Use the equivalent problem to say, without integrating, what **B** is in the hollow of the pipe, $s < a$, and how **B** outside the pipe compares with the field of a thin straight wire that carries the same current.

(d) Find the electric field of the equivalent problem for $a < s < b$, and translate it into **B** in the copper.

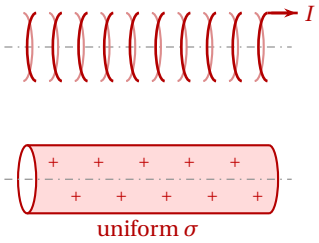


Figure 14.11 The long solenoid (top) and the long, thin, uniformly charged cylindrical shell (bottom) of Exercise P14.2.

P14.2 A classmate says that a long, thin cylindrical shell with a uniform surface charge is the electrostatic equivalent of a long solenoid, since both are long cylinders (Figure 14.11).

(a) Compare the two fields. For each, which way does the field point, and how large is it, inside and outside?

(b) Name the rule that pairs an electrostatics problem with a magnetostatics problem, and use it to explain why this pairing fails.

(c) Use the boundary conditions of Table 14.3 to explain why no single sheet of charge can produce the solenoid's pattern of field: uniform and along the axis inside, zero outside.

P14.3 Table 14.1 has "No easy parallel" beside Coulomb's law (row 2), and the magnetic column has no partner for the field of a set of point charges (the first line of row 4).

(a) Explain why magnetostatics has no point source that could fill these places. Consider three candidates: a magnetic "point charge," a short piece of wire carrying a steady current with nothing attached to its ends, and a single charge that moves at constant velocity. For each, say which equation or requirement of magnetostatics rules it out.

(b) What is the simplest source of a static magnetic field, and which entry of the electric column is its partner?

P14.4 A superconductor keeps magnetic field out of its interior (the Meissner effect); in this exercise, take $\mathbf{B} = 0$ everywhere inside one. A long, straight wire runs parallel to the flat surface of a large superconductor, at the height d above it, and carries a steady current I (Figure 14.12). Take the surface to be the plane $y = 0$, with the superconductor below it, and the wire to run along the line $x = 0$, $y = d$, with the current in the $+z$ direction.

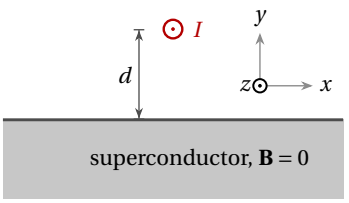


Figure 14.12 The wire and the superconductor of Exercise P14.4, seen end on, with the current out of the page.

(a) What condition does $\mathbf{B}$ satisfy just above the surface? (Use Table 14.3.)

(b) Every current here, including the current that flows on the superconductor's surface, runs along z, so $\mathbf{A} = A_z(x, y)\,\hat{\mathbf{z}}$. Use (12.17) to show that the condition of part (a) makes A_z the same at every point of the surface.

(c) This is the condition that V satisfies on a grounded conducting plane. Use the correspondence of Section 12.2.1 and the method of images (Section 7.2.1) to find A_z above the surface. What is the image?

(d) Find the force per unit length on the wire. Is the wire pulled toward the superconductor or pushed away? Compare the force on a line charge λ at the height d above a grounded conducting plane: do the two forces point the same way? Explain why or why not.

Exercises for 14.3 Which method?

P14.5 For each of the following, sketch the situation, say which method of Figure 14.9 or 14.10 you would use first, and give a one-line reason. Do not solve the problems.

(a) The electric field everywhere, inside and outside, of a solid cylinder of uniform charge density whose length equals its diameter.

(b) The magnetic field 2 m from a small coil of 20 turns wound on a square frame 1 cm on a side.

(c) The electric field between two concentric conducting spherical shells, charged $+Q$ and $-Q$, when the space between them is filled with an insulator whose dielectric constant depends on the distance from the center.

(d) The magnetic field of a circular loop of wire at a point in the plane of the loop, halfway between its center and the wire.

(e) The force per unit length on a long, straight wire, with charge λ per unit length, that runs parallel to a large grounded metal plate at the height d.

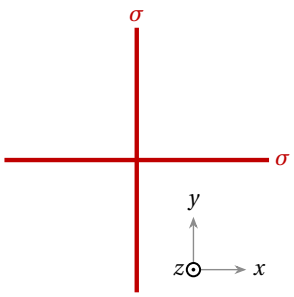


Figure 14.13 The two sheets of charge of Exercise P14.6, seen end on, looking down the z axis.

(f) The magnetic field inside a long cylinder of a linear magnetic material whose axis is perpendicular to a uniform applied field.

P14.6 Two large, thin sheets of charge cross at right angles along the z axis. One lies in the plane $x = 0$ and the other in the plane $y = 0$, and each has the uniform surface charge density σ (Figure 14.13).

(a) Four classmates propose ways to find **E** at a point in the region $x > 0$, $y > 0$: (i) "The charge as a whole has no planar symmetry, so Gauss's law cannot help; use the script-$\mathscr{r}$ method." (ii) "Take one Gaussian surface around the line where the sheets cross, and find E from the total charge inside it." (iii) "Use Gauss's law for each sheet alone, and add the magnitudes of the two fields." (iv) "Use Gauss's law for each sheet alone, and add the two fields as vectors." Which of these work, and which is quickest? Give a one-line reason for each.

(b) Find **E** in each of the four regions between the sheets, and sketch the field lines.

P14.7 For each of the following, sketch the situation, say which method of Figure 14.9 or 14.10 you would use first, and give a one-line reason. Do not solve the problems.

(a) The electric field at a point in the plane of a thin, uniformly charged disk, just outside its rim.

(b) The magnetic field near one edge of a long, straight copper bar, 1 cm by 4 cm in cross section, that carries a steady current.

(c) The potential inside a long, hollow metal pipe of square cross section, far from its ends, when its four walls are insulated from each other, one wall is held at the potential V_0, and the other three are grounded.

(d) The magnetic field inside and outside a long, straight wire of radius R, made of a linear magnetic material, that carries a current I spread uniformly over its cross section.

(e) The electric field 1 m from a plastic cube 1 cm on a side that has a uniform charge density.

(f) The magnetic field at a point 1 mm from a thin wire that is bent into a circle of radius 1 m and carries a steady current.

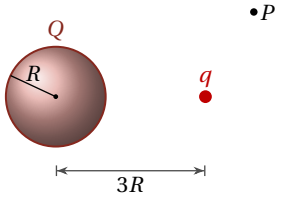


Figure 14.14 The charged ball, the point charge q, and a field point P of Exercise P14.8.

P14.8 A solid ball of radius R has a uniform charge density and the total charge Q. A point charge q sits a distance $3R$ from the ball's center (Figure 14.14). Both charges are positive.

(a) Four classmates propose ways to find **E** at a point P outside the ball: (i) "Draw a Gaussian sphere through P, centered on the ball. If it encloses only the ball, $E = Q/4\pi\varepsilon_0 r^2$ at P." (ii) "Outside the ball, Gauss's law says that the ball's field is the field of a point charge Q at its center. Add that field to the field of q, as vectors." (iii) "Replace the ball and q by a single point charge $Q+q$ at their center of charge, and use Coulomb's law." (iv) "Use the script-$\mathscr{v}$ method for the field of the ball and Coulomb's law for the field of q, and add the two fields." Which of these work, and which is quickest? Give a one-line reason for each.

(b) For which points P does method (iii) give nearly the right answer, and why?

P14.9 For each of the following, sketch the situation, say which method of Figure 14.9 or 14.10 you would use first, and give a one-line reason. Do not solve the problems.

(a) The magnetic field 100 m from two long, straight, parallel cables, 1 m apart, that carry equal steady currents in opposite directions.

(b) The electric field at points on the axis of a small circular hole cut in a large, flat sheet with a uniform surface charge density.

(c) The magnetic field at points on the axis of a flat, ring-shaped permanent magnet, like a thick washer, that is magnetized uniformly along its axis.

(d) The potential between two long metal cylinders, one inside the other, whose axes are parallel but 1 cm apart, when the two are held at different potentials.

(e) The magnetic field 1 m from a small button magnet a few millimeters across.

(f) The electric field just outside a metal object of irregular shape, at a point of its surface where the surface charge density is known.

P14.10 Four classmates propose ways to find **B** inside a long solenoid, far from its ends: (i) "Use Ampère's law with a rectangular loop that has one long side inside the solenoid and the other outside." (ii) "Find **A** by the flux method, then take its curl." (iii) "Add up the fields of all the turns with the Biot–Savart law." (iv) "Solve for A_z on a grid across the solenoid's cross section by the relaxation method of Section 12.5."

(a) Which of these work, and which is quickest? Give a one-line reason for each.

(b) Which of them would still work for a short solenoid, whose length equals its diameter?

P14.11 A long, straight wire carries a steady current I and also has a uniform net charge λ per unit length. A small charge q moves parallel to the wire, in the direction of the current, with the speed v, a distance s from the wire. Take I, λ and q to be positive.

(a) Four classmates propose ways to find the force on q: (i) "Find **E** from Gauss's law and **B** from Ampère's law, then use $\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B})$." (ii) "The charge is moving, so only the magnetic force acts on it." (iii) "Find **E** and **B** with the script-$\mathscr{r}$ integrals, then use the same force law as in (i)." (iv) "Find **E** and **B**, add them, and multiply the sum by q." Which of these work, and which is quickest? Give a one-line reason for each.

(b) Find the force on q, and the speed v at which it is zero.

Part IV

Time-Varying Fields

I am busy just now again on electro-magnetism, and think I have got hold of a good thing, but can't say. It may be a weed instead of a fish that, after all my labour, I may at last pull up.

— Michael Faraday,
letter to Richard Phillips (1831)*

*Michael Faraday to Richard Phillips, 23 September 1831, in H. Bence Jones, *The Life and Letters of Faraday*, 2nd ed., Vol. 2 (Longmans, Green, London, 1870), Chapter 1.

Chapter 15

Faraday's Law

Until now the fields have held still. Charges sat in place or moved in steady currents, and the electric and magnetic fields could be treated as two separate subjects that happen to obey similar mathematics (Chapter 14). This chapter and the next let the fields change in time. Two new effects appear, one in each curl equation. A changing magnetic field produces an electric field, which is Faraday's law, the subject of this chapter. A changing electric field gives the magnetic field a curl, which is Maxwell's displacement current, the subject of Chapter 16. With these two effects the four equations of Chapter 2 are complete, the two fields stop being independent of each other, and the equations predict waves that travel at the speed of light.

The chapter begins with an older and more familiar idea, the electromotive force produced in a wire that moves through a magnetic field. It then describes the experiments that led Faraday to his law, states the law, with Lenz's law for the direction of the induced current, and develops a way of using it, the Faraday-current method, that uses exactly the mathematics of Ampère's law. It also treats the eddy currents that a changing magnetic field drives in a solid piece of metal, and it revises the relation between the fields and the potentials. The last section uses a computer to follow a magnet as it falls through a loop of wire while the current it induces slows its fall. Inductance and the energy stored in a magnetic field, which also rest on Faraday's law, are taken up in Section 19.4.

15.1 Motional EMF

15.1.1 Electromotive force

A steady current around a closed circuit needs something to keep pushing the charges. An electrostatic field cannot do it by itself. Its curl is zero, so its line integral around any closed path is zero (Section 5.1): a charge carried once around the circuit by an electrostatic field gains no energy, while the resistance of the wire takes energy from it on every trip. Some other agent has to do work on the charges (Figure 15.1). In a flashlight it is the chemistry of the battery; in a power

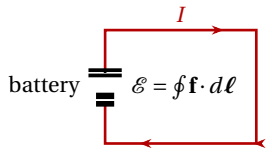


Figure 15.1 A circuit driven by a battery. The electrostatic field does no net work on a charge carried around the circuit; the battery does. The EMF is the line integral of the total force per unit charge **f** once around the circuit.

plant it is a generator; in a solar cell it is light.

To describe all such agents at once, let **f** be the force per unit charge on the charges in a circuit, whatever its cause: it includes the electric field, the chemical forces in a battery, gravity, and anything else that makes charge move. The **electromotive force**, or **EMF**, of the circuit is the line integral of **f** once around it,

$$\mathscr{E} = \oint \mathbf{f} \cdot d\boldsymbol{\ell}. \tag{15.1}$$

With words, it says that the EMF is the work done on a charge, per unit charge, as the charge is carried once around the circuit. Its unit is the joule per coulomb, which is the volt. The name is unfortunate, since an EMF is a work per charge and not a force, but it is an old one and it has stuck.

An EMF has the same units as a potential difference, and in one simple case the two are equal. Inside an ideal battery that is not connected to anything, no current flows, so the chemical force per charge exactly cancels the electric field there, $\mathbf{f}_{\text{chem}} = -\mathbf{E}$. The electrostatic field contributes nothing to (15.1), because its line integral around the closed path is zero, so

$$\mathscr{E} = \int_-^+ \mathbf{f}_{\text{chem}} \cdot d\boldsymbol{\ell} = -\int_-^+ \mathbf{E} \cdot d\boldsymbol{\ell} = V_+ - V_-,$$

where the integral runs through the battery from its negative terminal to its positive one. In general, though, an EMF is not a potential difference. A potential is a field, with a value at every point in space, and an EMF is a single number that belongs to a whole circuit. When the circuit has a resistance R, the EMF drives a current $I = \mathscr{E}/R$, as you saw in your introductory course; Chapter 19 discusses circuits in more detail.

There are three main sources of EMF in practice: batteries and other chemical cells, the motion of a wire through a magnetic field, and a magnetic field that changes in time. The second is the subject of this section, and the third of the next.

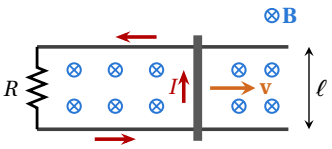


Figure 15.2 A conducting bar slides to the right with speed v along two rails a distance ℓ apart. A resistor R joins the left ends of the rails, and a uniform field **B** points into the page. The magnetic force pushes the positive charges in the bar up the page, so the current I flows counterclockwise.

15.1.2 A conducting bar moving through a field

Two parallel conducting rails, a distance ℓ apart, lie in a uniform magnetic field that points into the page (Figure 15.2). A resistor R joins their left ends, and a conducting bar lies across them and slides to the right with speed v. What happens to the charges in the bar?

Each charge q in the bar moves to the right with the bar, so it feels a magnetic force $q\mathbf{v} \times \mathbf{B}$. With **v** along $+\hat{\mathbf{x}}$ and **B** along $-\hat{\mathbf{z}}$ (into the page),

$$\mathbf{v} \times \mathbf{B} = v\hat{\mathbf{x}} \times (-B\hat{\mathbf{z}}) = vB\hat{\mathbf{y}},$$

so the force pushes positive charges up the bar, along $\hat{\mathbf{y}}$, and electrons down it. In a metal it is the electrons that move, down the bar, but the conventional current I flows up the bar, along the top rail to the left, down through the resistor, and back along the bottom rail: counterclockwise. The moving bar acts just like a battery.

Only the bar contributes to the EMF, since the rails and the resistor are not moving. The force per unit charge in the bar is $\mathbf{f} = \mathbf{v} \times \mathbf{B}$, which has the constant magnitude vB and points along the bar, so

$$\mathscr{E} = \int_{\text{bar}} (\mathbf{v} \times \mathbf{B}) \cdot d\boldsymbol{\ell} = vB\ell. \tag{15.2}$$

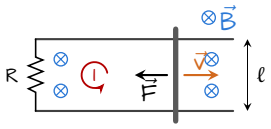
An EMF produced by the motion of a conductor through a magnetic field is called a **motional EMF**.

There is a second contribution to the magnetic force. Once the current flows, the charges also move along the bar, with some drift velocity $\mathbf{u}$, and the magnetic force on them includes a part $q\,\mathbf{u} \times \mathbf{B}$. That part is perpendicular to the bar, so it adds nothing to (15.2). It pushes the charges sideways, against the walls of the bar, and the bar as a whole feels it as the force on a current-carrying wire, $I\boldsymbol{\ell} \times \mathbf{B}$ (Section 10.4.2). Its direction is to the left, against the motion: to keep the bar moving at constant speed, someone has to push it.

Example 15.1

In the arrangement of Figure 15.2, $B = 0.40\,\text{T}$, $\ell = 25\,\text{cm}$, and $R = 0.50\,\Omega$; the rails and the bar have negligible resistance. The bar is pushed so that it slides to the right at a constant 2.0 m/s. Find the current, the force needed to push the bar, and the power that the push delivers, and compare the power with the rate at which the resistor dissipates energy.

Solution:



We start with a sketch of what is given and what is asked: the rails, the resistor, the bar with its velocity, the field into the page, and the current we expect, counterclockwise. The force on the bar is the force on a current in a field, and we expect it to point to the left, against the motion.

The EMF is (15.2),

$$\mathscr{E} = vB\ell = (2.0\,\text{m/s})(0.40\,\text{T})(0.25\,\text{m}) = 0.20\,\text{V},$$

so the current is $I = \mathscr{E}/R = 0.40\,\text{A}$, counterclockwise. The force of the field on the current in the bar is

$$\mathbf{F} = I\boldsymbol{\ell} \times \mathbf{B} = I\ell\,\hat{\mathbf{y}} \times (-B\hat{\mathbf{z}}) = -I\ell B\hat{\mathbf{x}},$$

of magnitude $(0.40\,\text{A})(0.25\,\text{m})(0.40\,\text{T}) = 0.040\,\text{N}$, to the left. (Why does the bar's own motion through the field add nothing to this force? Answer in footnote.[1]) To keep the bar moving at constant speed, the push must be 0.040 N to the right (Figure 15.3). It

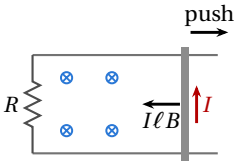


Figure 15.3 The forces on the bar of Example 15.1: the field pushes the current in the bar to the left with the force $I\ell B$, and the push to the right balances it.

[1] The force $q\mathbf{v} \times \mathbf{B}$ from the bar's motion acts on the positive ions as well as the electrons, and the bar is neutral, so these forces cancel. Only the electrons' drift along the bar, the current, gives a net force.

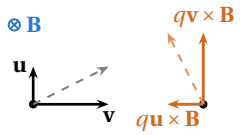


Figure 15.4 A charge in the moving bar has the bar's velocity **v** and a velocity **u** along the bar (left). The magnetic force (right) is perpendicular to the total velocity: its part $q\mathbf{v}\times\mathbf{B}$, along the bar, does positive work, and its part $q\mathbf{u}\times\mathbf{B}$, against the bar's motion, does the same amount of negative work.

delivers the power

$$P_{\text{push}} = Fv = (0.040\,\text{N})(2.0\,\text{m/s}) = 0.080\,\text{W},$$

and the resistor dissipates

$$P_R = I^2R = (0.40\,\text{A})^2(0.50\,\Omega) = 0.080\,\text{W}.$$

Reality check: They are equal, as energy conservation requires. In symbols, $Fv = (I\ell B)v = I(vB\ell) = I\mathcal{E} = I^2R$: every joule of work done by the push ends up as heat in the resistor.

The example settles a question that might otherwise be puzzling. Magnetic forces do no work (Section 10.1.3), yet the current through the resistor is driven by a magnetic force. The two parts of the magnetic force on each charge do equal and opposite amounts of work (Figure 15.4). The part $q\mathbf{v}\times\mathbf{B}$, along the bar, does positive work on the charges as they move up the bar, and that work is the EMF. The part $q\mathbf{u}\times\mathbf{B}$, against the motion of the bar, does exactly the same amount of negative work, since the total force $q(\mathbf{v}+\mathbf{u})\times\mathbf{B}$ is perpendicular to the total velocity $\mathbf{v}+\mathbf{u}$. The energy comes from whoever pushes the bar; the magnetic force only redirects it, as it does when a magnetic force lifts a wire (Section 10.4.2).

15.1.3 The flux rule

The same EMF can be found in a second way. The magnetic flux through the circuit (Section 11.1.1),

$$\Phi_B = \int_S \mathbf{B}\cdot d\mathbf{a},$$

where S is any surface bounded by the circuit, changes as the bar moves. Take the direction around the circuit to be counterclockwise, the direction of the current, so that by the right-hand rule the normal to the circuit's surface is $+\hat{\mathbf{z}}$, out of the page. With the bar a distance x from the resistor, the flux is $\Phi_B = -B\ell x$, negative because **B** points into the page, and it changes at the rate

$$\frac{d\Phi_B}{dt} = -B\ell\frac{dx}{dt} = -B\ell v.$$

This is the negative of the EMF. So

$$\boxed{\mathcal{E} = -\frac{d\Phi_B}{dt}}, \tag{15.3}$$

where (15.1) is taken around the circuit in the direction that makes the normal of S point along $d\mathbf{a}$ by the right-hand rule. We will call (15.3) the **flux rule**. The minus sign carries real information: a positive EMF drives the current in the chosen direction, counterclockwise here, and a negative one drives it the other way.

The flux rule holds for a circuit of any shape that moves, or changes its shape, in a magnetic field that does not change in time. The proof is short.

Proof of the flux rule

Let a closed loop C move in a static field $\mathbf{B}$, each piece $d\boldsymbol{\ell}$ of it with its own velocity $\mathbf{v}$. In a time dt the piece $d\boldsymbol{\ell}$ sweeps out a small parallelogram with sides $\mathbf{v}\,dt$ and $d\boldsymbol{\ell}$ (Figure 15.5), whose vector area is $d\mathbf{a} = (\mathbf{v}\,dt)\times d\boldsymbol{\ell}$, and together these parallelograms make a ribbon that joins the loop at time t to the loop at time $t+dt$. The flux through any surface bounded by a loop is the same for every such surface, because $\nabla\cdot\mathbf{B}=0$ (Section 11.1.1). So for the new loop we may use the old surface plus the ribbon, and the flux increases by the flux through the ribbon:

$$d\Phi_B = \oint \mathbf{B}\cdot[(\mathbf{v}\,dt)\times d\boldsymbol{\ell}]\,.$$

(With $d\boldsymbol{\ell}$ taken around the loop in the positive direction, $(\mathbf{v}\,dt)\times d\boldsymbol{\ell}$ points along the normal of the surface when the loop grows, as it does for the sliding bar, where $\mathbf{v}\times d\boldsymbol{\ell}$ is along $+\hat{\mathbf{z}}$.) A triple product can be rearranged cyclically, $\mathbf{B}\cdot(\mathbf{v}\times d\boldsymbol{\ell}) = d\boldsymbol{\ell}\cdot(\mathbf{B}\times\mathbf{v}) = -(\mathbf{v}\times\mathbf{B})\cdot d\boldsymbol{\ell}$, so

$$\frac{d\Phi_B}{dt} = -\oint (\mathbf{v}\times\mathbf{B})\cdot d\boldsymbol{\ell}.$$

The magnetic force per unit charge on a charge in the wire is $(\mathbf{v}+\mathbf{u})\times\mathbf{B}$, where $\mathbf{u}$ is its drift velocity along the wire. The part $\mathbf{u}\times\mathbf{B}$ is perpendicular to $d\boldsymbol{\ell}$ and contributes nothing to $\oint\mathbf{f}\cdot d\boldsymbol{\ell}$, so the EMF is $\oint(\mathbf{v}\times\mathbf{B})\cdot d\boldsymbol{\ell} = -d\Phi_B/dt$, which is (15.3).

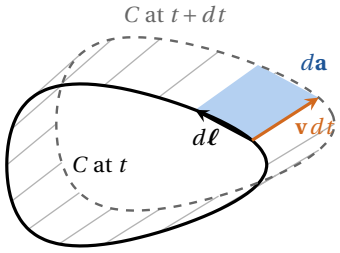


Figure 15.5 A loop moving through a static field. In a time dt each piece $d\boldsymbol{\ell}$ of the loop moves $\mathbf{v}\,dt$ and sweeps out a small parallelogram of vector area $d\mathbf{a} = (\mathbf{v}\,dt)\times d\boldsymbol{\ell}$. Together the parallelograms make a ribbon joining the old loop to the new one.

Nothing new went into the flux rule. It follows from the Lorentz force law and from $\nabla\cdot\mathbf{B}=0$, and it is a convenient way to find a motional EMF without working out the force on each piece of the circuit. It also explains at a glance when there is no EMF. A rectangular loop that slides through a uniform field without turning has two sides moving through the field, and their EMFs, $vB\ell$ each, push the current in opposite directions around the loop and cancel. In the language of the flux rule, the flux through the loop does not change, so $\mathscr{E}=0$.

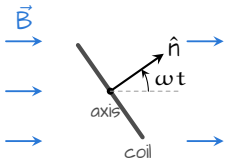
The flux rule needs a well-defined circuit, made of the same pieces of material at every moment.[2] With that proviso, it is the easiest way to find the EMF of a generator.

Example 15.2

A generator consists of a flat coil of N turns, each of area A, that turns at a constant angular velocity ω about an axis in its own plane, in a uniform field $\mathbf{B}$ perpendicular to that axis. Find the EMF. Evaluate its peak value for $N=100$, $A=0.010\,\mathrm{m}^2$, $B=0.50\,\mathrm{T}$, and a rotation rate of 60 turns per second.

[2] When a circuit is closed through a sliding contact, so that different material becomes part of the circuit as time goes on, the rule can fail. Richard Feynman gives two examples, one with an EMF but no change in flux and one with a change in flux but no EMF: R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Vol. II, Section 17-2. https://feynmanlectures.caltech.edu/II_17.html. In every case the force law $\mathbf{F}=q(\mathbf{E}+\mathbf{v}\times\mathbf{B})$ gives the right answer.

Solution:



We draw the coil edge-on, turned at an angle ωt from the position in which its normal points along **B**. The flux through each turn is BA times the cosine of the angle between the normal and the field, so it rises and falls as the coil turns.

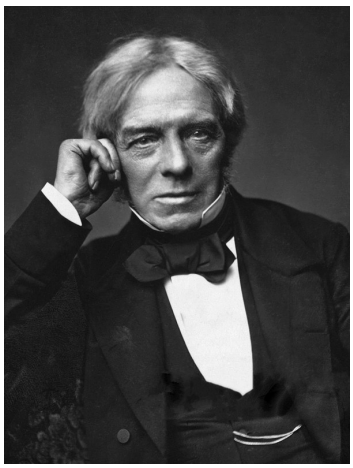
Measure time from an instant when the normal to the coil points along **B**. At time t the angle between the normal and the field is ωt, and the flux through each turn is $BA\cos\omega t$. The N turns are connected in series, so their EMFs add, and the flux rule gives

$$\mathcal{E} = -N\frac{d}{dt}(BA\cos\omega t) = NBA\omega\sin\omega t.$$

The EMF alternates in sign, so the current it drives in a circuit reverses direction twice in every turn: the generator produces an alternating current. Its peak value is

$$\mathcal{E}_{max} = NBA\omega = (100)(0.50\,\mathrm{T})(0.010\,\mathrm{m}^2)(2\pi\times 60\,\mathrm{s}^{-1}) = 190\,\mathrm{V}.$$

Reality check: The EMF is zero when the flux is largest, at $\omega t = 0$, and largest when the flux passes through zero, which is when the two long sides of each turn cut across the field fastest. The units are $\mathrm{T}\cdot\mathrm{m}^2/\mathrm{s} = \mathrm{V}$.



Michael Faraday (1791–1867, English) With only the most basic schooling, Faraday was apprenticed at 14 to a bookbinder and largely educated himself. After attending lectures by Humphry Davy at the Royal Institution in London, he became Davy's assistant there in 1813, and he stayed at the Royal Institution for the rest of his career. In 1821 he built the first devices to turn an electric current into continuous motion; in 1831 he discovered electromagnetic induction; and he found the laws of electrolysis and, in 1845, discovered diamagnetism and the rotation of polarized light by a magnetic field. He thought in terms of lines of force. His mathematics did not extend as far as trigonometry, and it was Maxwell who put his ideas into equations. He declined a knighthood. (Wikipedia)

Most of the world's electricity is made this way, by coils turning in magnetic fields, or by magnets turning inside coils, driven by steam turbines, water, or wind. When the generator delivers a current, the force of the field on that current opposes the rotation, as the force on the sliding bar opposed its motion, so the turbine has to do work to keep the coil turning. A hand-cranked generator connected to a few light bulbs shows this directly: it is noticeably harder to crank with the bulbs connected than with the circuit open.

15.2 Faraday's law and the Faraday current

15.2.1 Faraday's experiments

On 29 August 1831, Michael Faraday wound two coils of insulated wire on opposite sides of an iron ring, connected one coil to a battery, and connected the other to a galvanometer. When he completed the battery circuit, the galvanometer needle jumped, and then settled back to zero while the current in the first coil flowed steadily; when he broke the circuit, the needle jumped again, "but in the contrary direction."[3] Over the following weeks he found that moving a magnet

[3] M. Faraday, *Experimental Researches in Electricity*, First Series, paragraphs 27–32, read to the Royal Society on 24 November 1831. The date of the ring experiment is from Faraday's laboratory notebook, as given by the Royal Institution, where the ring is kept: https://www.rigb.org/explore-science/explore/collection/michael-faradays-ring-coil-apparatus.

into or out of a coil did the same, and he reported the results to the Royal Society that November. Joseph Henry, in Albany, New York, made similar discoveries independently and published them in 1832. (Section 2.4 puts these discoveries in the history of Maxwell's equations.)

In modern terms, all of these observations come to the same thing. Consider three experiments (Figure 15.6). (a) A loop of wire is pulled away from a magnet that is held still. (b) The loop is held still and the magnet is moved away. (c) Nothing moves, but the current in a nearby electromagnet is decreased. In all three, a current flows in the loop while the flux through it changes, and the EMF obeys the same rule, $\mathscr{E} = -d\Phi_B/dt$.

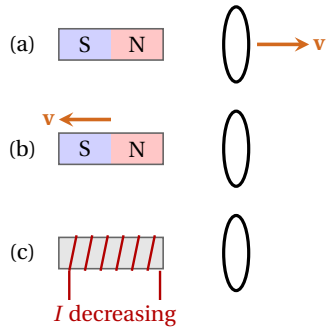


Figure 15.6 Three ways to change the flux through a loop of wire: (a) pull the loop away from a magnet that is held still; (b) hold the loop still and move the magnet away; (c) move nothing, and decrease the current in an electromagnet. In all three the EMF in the loop is $-d\Phi_B/dt$. Animation of a magnet pushed toward a ring and pulled away (Lenz's law, Section 15.2.2).

In experiment (a) this is no surprise: it is the motional EMF of Section 15.1.3. But in (b) and (c) the charges in the loop are at rest, and a magnetic force does not act on charges at rest. Something else pushes them around the loop, and the only candidate is an electric field. A changing magnetic field produces an electric field. For a loop that does not move, the EMF is then the line integral of that electric field, and the experiments say that

$$\boxed{\oint \mathbf{E}\cdot d\boldsymbol{\ell} = -\frac{d\Phi_B}{dt} = -\frac{d}{dt}\int_S \mathbf{B}\cdot d\mathbf{a}}\,, \tag{15.4}$$

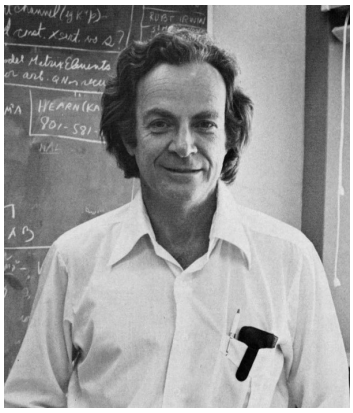
where S is any surface bounded by the loop, with its normal related to the direction of the line integral by the right-hand rule. This is **Faraday's law**. It cannot be derived from anything earlier in this book: like Coulomb's law and the Biot–Savart law, it is an experimental law.

That experiments (b) and (c) give the same rule as experiment (a) is remarkable. If the loop and the magnet move apart at a given speed, the EMF is the same whether we say that the loop moves, so that a magnetic force pushes its charges, or that the magnet moves, so that a changing field produces an electric field that pushes them. Only the relative motion matters. Richard Feynman remarked, "We know of no other place in physics where such a simple and accurate general principle requires for its real understanding an analysis in terms of two different phenomena,"[4] and it was this asymmetry in the description of a magnet and a conductor that Albert Einstein chose as the first example in his 1905 paper on special relativity.[5] Relativity shows that the two descriptions are the same physics seen by observers in different frames of reference (Section 26.2.4).

Because both cases give $\mathscr{E} = -d\Phi_B/dt$, many people call that equation "Faraday's law" whatever the reason the flux changes. In this book Faraday's law means (15.4), for a field that changes, and $\mathscr{E} = -d\Phi_B/dt$ for a moving circuit is the flux rule of Section 15.1.3. When a circuit moves through a field that is also changing,

[4] *The Feynman Lectures on Physics*, Vol. II, Section 17-1 (see the footnote in Section 15.1.3).

[5] "Zur Elektrodynamik bewegter Körper," A. Einstein, Ann. Phys. 17, 891–921 (1905). https://doi.org/10.1002/andp.19053221004. The paper opens: "Take, for example, the reciprocal electrodynamic action of a magnet and a conductor. The observable phenomenon here depends only on the relative motion of the conductor and the magnet," in the translation of W. Perrett and G. B. Jeffery (1923).



Richard P. Feynman (1918–1988, American) Feynman was born in New York City, studied at MIT and Princeton, and spent the Second World War at Los Alamos, on the atomic bomb project. At Cornell, and from the early 1950s at the California Institute of Technology, he developed a new formulation of quantum electrodynamics, with the diagrams now named after him, for which he shared the 1965 Nobel Prize in Physics with Julian Schwinger and Sin-Itiro Tomonaga. The introductory course he taught at Caltech in the early 1960s became *The Feynman Lectures on Physics*, which this chapter cites several times. In 1986, on the commission that investigated the loss of the space shuttle *Challenger*, he showed at a televised hearing that the rubber of the booster rockets' O-ring seals lost its resilience in ice water. (Wikipedia)

the two effects add, and $\mathscr{E} = -d\Phi_B/dt$ still holds, with $d\Phi_B/dt$ the total rate of change of the flux through the moving circuit.

15.2.2 Lenz's law

The minus sign in Faraday's law and in the flux rule fixes the direction of the induced EMF. It can be put into words.

> **Lenz's law:** The induced current flows in the direction that makes its own magnetic field oppose the *change* in the flux.

It is the change that is opposed, not the flux. If the flux through a loop is increasing, the induced current makes a field that points against the existing one, through the loop; if the flux is decreasing, the induced current makes a field that points along it, trying to keep the flux up. The law is a statement of energy conservation. If the induced current reinforced the change that produced it, the change would grow, the current would grow with it, and energy would come from nowhere.

In practice, the direction of an induced current is found in four steps:

1. Find which way **B** points through the loop.
2. Decide whether the flux through the loop is increasing or decreasing.
3. The field of the induced current must point against **B**, through the loop, if the flux is increasing, and along **B** if it is decreasing.
4. The right-hand rule gives the direction of the current that makes that field.

For the sliding bar of Figure 15.2, the flux into the page grows as the bar moves away from the resistor, so the induced current must make a field out of the page inside the circuit, and it flows counterclockwise, as we found from the force on the charges. When the north pole of a magnet is pushed toward a ring, the flux through the ring grows, and the induced current makes the face of the ring nearest the magnet act like a north pole too: the ring repels the magnet, which is how it opposes the change. When the magnet is pulled away, the ring attracts it.

The same rule applies when the changing flux comes from the loop's own current. A current in a coil makes a field that passes through the coil itself, so when the current changes, so does that flux, and the induced EMF opposes the change: it pushes against a current that is increasing and tries to keep a decreasing one going. For circuits that do not move or change shape, the flux through loop 2 due to a current I_1 in loop 1 is MI_1, with the mutual inductance of (12.30), and the flux through a loop due to its own current is LI, with the self-inductance of (12.32). The induced EMFs are therefore

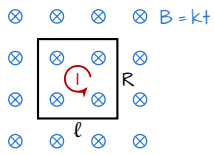
$$\mathscr{E}_2 = -M\frac{dI_1}{dt}, \qquad \mathscr{E} = -L\frac{dI}{dt}. \tag{15.5}$$

The second is called the **back EMF**. Section 19.4 takes up inductance in full.

Example 15.3

A square loop of wire, with sides $\ell = 10\,\text{cm}$ and resistance $R = 0.50\,\Omega$, lies in a uniform magnetic field that points into the page and increases with time as $B = kt$, with $k = 0.20\,\text{T/s}$. Find the direction and size of the induced current and the power it dissipates. Where does the energy come from?

Solution:



We draw the loop in the plane of the page, with the field into the page and growing. By Lenz's law the current must make a field out of the page inside the loop, so we expect it to flow counterclockwise.

The flux through the loop, counted into the page, is $\Phi_B = B\ell^2 = k\ell^2 t$, and it is increasing, so by Lenz's law the induced current flows counterclockwise, making a field out of the page inside the loop. The size of the EMF is

$$|\mathcal{E}| = \frac{d\Phi_B}{dt} = k\ell^2 = (0.20\,\text{T/s})(0.10\,\text{m})^2 = 2.0\,\text{mV},$$

so the current is $I = |\mathcal{E}|/R = 4.0\,\text{mA}$, and it dissipates the power

$$P = I^2 R = \frac{k^2\ell^4}{R} = 8.0\,\mu\text{W},$$

constant in time. (Why is the power constant in time, although the field keeps growing? Answer in footnote.[6]) The loop does not move, so nothing pushes it; the energy comes from whatever is increasing the field. If the field is made by an electromagnet, its power supply has to work harder to increase the current in the electromagnet when the loop is present than when it is not. The induced current makes a field that opposes the increase, and the supply has to overcome it.

Reality check: The units of $k^2\ell^4/R$ are $(\text{T}\cdot\text{m}^2/\text{s})^2/\Omega = \text{V}^2/\Omega = \text{W}$, and the power is also $|\mathcal{E}|\,I = (2.0\,\text{mV})(4.0\,\text{mA}) = 8.0\,\mu\text{W}$.



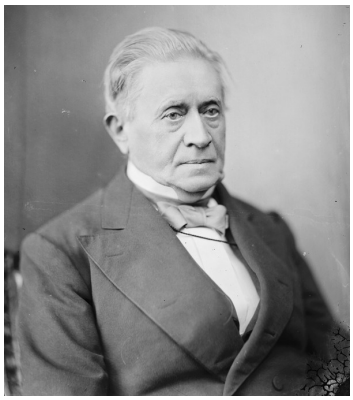
Emil Lenz (1804–1865, Russian, of Baltic German descent) Heinrich Friedrich Emil Lenz was born in Dorpat (now Tartu, Estonia) and studied at its university. From 1823 to 1826 he sailed with Otto von Kotzebue on a voyage around the world, studying the climate and the properties of seawater. He made the rest of his career in St. Petersburg, where he became a professor at the university in 1836, and later dean of its faculty of mathematics and physics and, from 1863, its rector. He took up electromagnetism in 1831, the year of Faraday's discovery of induction. He found his rule for the direction of induced currents in 1833 and published it in 1834, and in 1842, independently of James Joule, he showed that the heat a current produces in a wire is proportional to the square of the current. He died in Rome. (Wikipedia)

The electric field that pushes the charges around the loop is there whether or not the loop is. It fills the region where **B** is changing and the space around it, and the loop only reveals it. It would be a mistake to think of Faraday's law only in terms of the flux through a wire, as if nothing happened where there is no wire.

15.2.3 The differential form

For a loop and a surface that do not move, the time derivative in (15.4) can be taken inside the integral, where it becomes a partial derivative, and the curl

[6] The EMF depends on how fast the flux changes, not on how large the flux is. The flux grows at the steady rate $k\ell^2$, so the EMF, the current, and the power all stay the same.



Joseph Henry (1797–1878, American) Henry studied at the Albany Academy in Albany, New York, and taught mathematics and natural philosophy there from 1826 to 1832. In 1827 he began building electromagnets, winding many turns of insulated wire around an iron core; one that he built for Yale College could lift more than a ton. He discovered induction independently of Faraday, and in 1832 he found self-induction, the EMF that a changing current induces in its own circuit. He was a professor at the College of New Jersey (now Princeton University) from 1832 to 1846, and then the first Secretary of the Smithsonian Institution, from its founding in 1846 until his death. In 1893 an international congress of electricians meeting in Chicago named the unit of inductance (Section 12.7.1) the henry, in his honor. (Wikipedia)

theorem, (1.37), turns the line integral into an integral over the same surface:

$$\int_S (\nabla \times \mathbf{E}) \cdot d\mathbf{a} = -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{a}.$$

This holds for every surface, however small, so the integrands must be equal:

$$\boxed{\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}}. \tag{15.6}$$

This is the differential form of Faraday's law, the third of Maxwell's equations, (2.3). When the magnetic field does not change, it reduces to $\nabla \times \mathbf{E} = 0$, the result of Section 5.1. Electrostatics was the special case all along.

Two consequences are worth stating now. First, an electric field produced by a changing magnetic field is not conservative. Its line integral around a closed path is not zero, so the work it does on a charge carried around the path depends on the path, and it cannot be described by a scalar potential alone. The relation $\mathbf{E} = -\nabla V$ therefore needs revising when the fields change, and Section 15.4 revises it. Second, the boundary condition on the parallel component of $\mathbf{E}$ (Section 5.2) still holds when the fields change. It came from the line integral of $\mathbf{E}$ around a thin loop that straddles the surface, and the flux of $\mathbf{B}$ through such a loop goes to zero as its width does, so $\mathbf{E}^{\parallel}$ is still continuous.

15.2.4 The Faraday current

Compare Faraday's law with Ampère's law (Section 11.1):

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \qquad \nabla \times \mathbf{B} = \mu_0 \mathbf{J}.$$

In each, the left side is the curl of a field and the right side is what produces it. Ampère's law says that currents produce magnetic fields that curl around them. Faraday's law says that a changing magnetic field produces an electric field that curls around $-\partial\mathbf{B}/\partial t$ in exactly the same way. To emphasize this point, I like to call $-\partial\mathbf{B}/\partial t$ the **Faraday current density**,

$$\mathbf{J}_{\text{Far}} = -\frac{\partial \mathbf{B}}{\partial t}, \tag{15.7}$$

even though it is not a current density at all. No one else uses this name, just me, but I am hoping it will catch on. It has units of teslas per second, not amperes per square meter, and nothing flows. But it produces an electric field in exactly the way that a current density produces a magnetic field, with the μ_0 left out:

$$\nabla \times \mathbf{E} = \mathbf{J}_{\text{Far}}.$$

The analogy is complete wherever there are no charges. There $\nabla \cdot \mathbf{E} = 0$, just as $\nabla \cdot \mathbf{B} = 0$ everywhere, and the two fields obey equations of exactly the same form (Table 15.1).

Magnetostatics	Faraday's law, no charges
$\nabla \cdot \mathbf{B} = 0$	$\nabla \cdot \mathbf{E} = 0$
$\nabla \times \mathbf{B} = \mu_0 \mathbf{J}$	$\nabla \times \mathbf{E} = \mathbf{J}_{\text{Far}}$

Table 15.1 The equations of magnetostatics, and those of the induced electric field where there are no charges.

By the Helmholtz theorem (Section 2.3), a field that goes to zero far away is fixed by its divergence and its curl, so the electric field is determined by $\mathbf{J}_{\text{Far}}$ exactly as the magnetic field is determined by $\mu_0\mathbf{J}$. Everything in Chapters 10–12 about finding $\mathbf{B}$ from currents carries over to finding the induced $\mathbf{E}$ from a changing $\mathbf{B}$: the right-hand rule, the Biot–Savart law, and above all Ampère's law, which is the subject of the next section. Where there are charges as well, the electric field they produce is added to the induced one.

The direction of the induced field follows from the right-hand rule, applied as it is to a current: point the thumb of your right hand along $\mathbf{J}_{\text{Far}}$, which is *opposite* to the change in $\mathbf{B}$, and your fingers curl in the direction of $\mathbf{E}$ (Figure 15.7). If $\mathbf{B}$ points along $+\hat{\mathbf{z}}$ and is increasing, $\mathbf{J}_{\text{Far}}$ points along $-\hat{\mathbf{z}}$, and $\mathbf{E}$ circles the z axis clockwise as seen from above, in the $-\hat{\boldsymbol{\phi}}$ direction. If $\mathbf{B}$ along $+\hat{\mathbf{z}}$ is decreasing, $\mathbf{J}_{\text{Far}}$ points along $+\hat{\mathbf{z}}$, and $\mathbf{E}$ circles counterclockwise. This is the same direction that Lenz's law gives for the current in a loop of wire around the axis, as it must be: the field drives the current.

The integral form follows in the same way. The "current" of $\mathbf{J}_{\text{Far}}$ through a surface S bounded by a loop is

$$I_{\text{Far,enc}} = \int_S \mathbf{J}_{\text{Far}} \cdot d\mathbf{a} = -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{a} = -\frac{d\Phi_B}{dt},$$

and Faraday's law (15.4) becomes

$$\oint \mathbf{E} \cdot d\boldsymbol{\ell} = I_{\text{Far,enc}}, \tag{15.8}$$

the counterpart of Ampère's law $\oint \mathbf{B} \cdot d\boldsymbol{\ell} = \mu_0 I_{\text{enc}}$, without the μ_0.

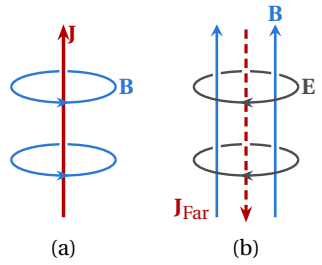


Figure 15.7 (a) A current density $\mathbf{J}$ produces a magnetic field $\mathbf{B}$ that curls around it. (b) A magnetic field that points up and is increasing gives a Faraday current density $\mathbf{J}_{\text{Far}} = -\partial\mathbf{B}/\partial t$ (dashed) pointing down, and the induced electric field $\mathbf{E}$ curls around $\mathbf{J}_{\text{Far}}$ in the same way that $\mathbf{B}$ curls around $\mathbf{J}$: clockwise, as seen from above.

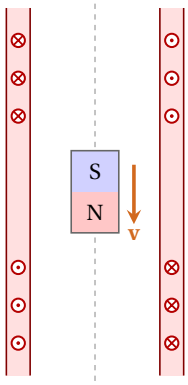


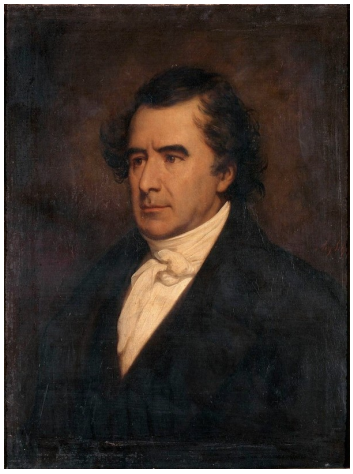
Figure 15.8 A magnet falling down a copper tube, in a cross section through the tube's axis. Below the magnet the flux through each ring of the tube is increasing, and above it the flux is decreasing, so the eddy currents (red) circle in opposite directions below and above the magnet. Both sets oppose the motion: the currents below repel the magnet and those above attract it. Animation.

15.2.5 Eddy currents

In a solid piece of metal, the induced electric field drives currents that circulate within the metal itself. They are called **eddy currents**, and by Lenz's law they oppose the change that produces them, so that a conductor moving near a magnet, or a magnet moving near a conductor, feels a drag. A strong magnet dropped down a vertical copper tube falls slowly, taking several seconds to pass through a tube a meter long, while an unmagnetized piece of metal of the same size drops straight through (Figure 15.8). The eddy currents in the rings of the tube below the magnet repel it, and those in the rings above attract it, and both slow it down. The magnet's own bound currents interact with the eddy currents just as currents in wires do. A magnet dropped onto a thick copper block slows as it approaches,

and the effect is stronger if the copper is first cooled in liquid nitrogen, which makes it a better conductor. A copper plate sliding down a ramp past a magnet slows down, while a plate with slots cut across it barely slows, because the slots break up the paths of the eddy currents.

The first eddy-current effects on record came before Faraday's discovery. In 1824 François Arago reported that the swings of a magnetized needle die out faster when the needle hangs inside a copper ring, and in 1825 that a copper disk spinning below a needle drags the needle around with it.[7] Faraday explained both as induction in 1831. Eddy currents are put to use in the magnetic brakes of some trains and roller coasters, in metal detectors, and in induction cooktops, which heat a pan by inducing currents in its bottom. In transformers and motors they are a nuisance: they heat the iron cores and waste energy. The cores are therefore built of thin insulated sheets, which leave the flux almost unchanged but block the paths of the eddy currents (Exercise P15.8).



François Arago (1786–1853, French) Dominique François Jean Arago entered the École polytechnique at seventeen, and in 1806 he went with Jean-Baptiste Biot to extend the measurement of the meridian through Spain. When war broke out there in 1808 he was arrested as a spy on Mallorca; he escaped to Algiers, was captured at sea, and reached France only in July 1809. That September, at twenty-three, he was elected to the Institut de France, and he taught at the École polytechnique until 1830. In optics he discovered chromatic polarization in 1811 and championed Augustin Fresnel's wave theory of light. In 1820 he showed that an electric current magnetizes iron, and his experiments with spinning metal disks earned him the Royal Society's Copley Medal in 1825. He directed the Paris Observatory, succeeded Fourier as perpetual secretary of the Académie des sciences in 1830, and in 1848, as a minister in the provisional government, signed the decree that abolished slavery in the French colonies. (Wikipedia)

15.3 Solving for E when given a changing B: Faraday-current method (high symmetry)

Faraday's law in the form (15.8), $\oint \mathbf{E}\cdot d\boldsymbol{\ell} = I_{\text{Far,enc}}$, is Ampère's law with $\mathbf{B}$ replaced by $\mathbf{E}$ and $\mu_0 I_{\text{enc}}$ replaced by $I_{\text{Far,enc}}$, and wherever there are no charges, the induced electric field is determined by $\mathbf{J}_{\text{Far}}$ in exactly the way that a magnetic field is determined by $\mu_0\mathbf{J}$ (Section 15.2.4). So the Ampère's law method of Section 11.2 carries over without change: wherever $\mathbf{J}_{\text{Far}}$ has one of the symmetries in which Ampère's law can be solved for $\mathbf{B}$, Faraday's law can be solved for $\mathbf{E}$. Table 15.2 lists what corresponds to what.

	Ampère's law	**Faraday's law**
The source	current density $\mathbf{J}$	Faraday current density $\mathbf{J}_{\text{Far}} = -\partial\mathbf{B}/\partial t$
The field it makes	$\mathbf{B}$	$\mathbf{E}$
Differential form	$\nabla\times\mathbf{B} = \mu_0\mathbf{J}$	$\nabla\times\mathbf{E} = \mathbf{J}_{\text{Far}}$
Integral form	$\oint \mathbf{B}\cdot d\boldsymbol{\ell} = \mu_0 I_{\text{enc}}$	$\oint \mathbf{E}\cdot d\boldsymbol{\ell} = I_{\text{Far,enc}} = -d\Phi_B/dt$
Direction	$\mathbf{B}$ curls around $\mathbf{J}$ by the right-hand rule	$\mathbf{E}$ curls around $\mathbf{J}_{\text{Far}}$ by the right-hand rule

Table 15.2 Ampère's law and Faraday's law, term by term.

The procedure is the same as before, with one step added at the start:

1. Find $\mathbf{B}$ and the Faraday current density $\mathbf{J}_{\text{Far}} = -\partial\mathbf{B}/\partial t$, and sketch $\mathbf{J}_{\text{Far}}$. Remember that it points *opposite* to the change in $\mathbf{B}$.

2. Use the symmetry of $\mathbf{J}_{\text{Far}}$ to determine the direction of $\mathbf{E}$, which curls around $\mathbf{J}_{\text{Far}}$, and which coordinates its magnitude can depend on.

[7] Faraday's explanation is in the First Series of his *Experimental Researches*, paragraphs 80–139: "Explication of Arago's magnetic phenomena."

3. Choose an "Amperian" loop that passes through the point where you want the field and that matches the symmetry. Draw it, list its pieces, and choose the direction in which you will go around it.

4. Calculate $I_{\text{Far,enc}}$, either by integrating $\mathbf{J}_{\text{Far}} \cdot d\mathbf{a}'$ over a surface bounded by the loop or as $-d\Phi_B/dt$, with the normal of the surface related to the direction around the loop by the right-hand rule.

5. Write down $\oint \mathbf{E} \cdot d\boldsymbol{\ell} = I_{\text{Far,enc}}$, simplify the left side to E times a length, solve for E, and give **E** its direction.

The loop is now a path for **E**, not for **B**, but Colton's Third Commandment (Section 11.2.1) applies with full force: thou shalt always draw an Amperian loop. In calling $-\partial\mathbf{B}/\partial t$ the "Faraday current density," I emphasize that it is *not* a current density at all. But it creates an electric field in exactly the same way a regular current density **J** creates a magnetic field, and the right-hand rule gives the direction of that field: point your right thumb along $\mathbf{J}_{\text{Far}}$, and your fingers curl around in the direction of the induced **E** (Figure 15.9).

The geometries are the ones of Section 11.2.2, with $\mathbf{J}_{\text{Far}}$ in place of **J**:

- **A changing field along the axis of a cylinder**, such as the field inside a long solenoid whose current changes, or any field $\mathbf{B} = B(s,t)\,\hat{\mathbf{z}}$. Then $\mathbf{J}_{\text{Far}}$ points along the axis and depends only on s, like the current in a wire, and **E** circles the axis, $\mathbf{E} = E(s)\,\hat{\boldsymbol{\phi}}$. The loop is a coaxial circle.

- **A changing field around the axis of a cylinder**, such as the field of a long straight wire or a coaxial cable whose current changes, $\mathbf{B} = B(s,t)\,\hat{\boldsymbol{\phi}}$. Then $\mathbf{J}_{\text{Far}}$ circles the axis, like the current in a solenoid, and **E** points along the axis, $\mathbf{E} = E(s)\,\hat{\mathbf{z}}$. The loop is a rectangle with two sides parallel to the axis.

- **Planar symmetry**, such as a changing uniform field between two large parallel sheets that carry opposite currents. Then $\mathbf{J}_{\text{Far}}$ is like the current in a slab, and **E** is parallel to the slab and perpendicular to $\mathbf{J}_{\text{Far}}$, in opposite directions on the two sides of the slab's midplane. The loop is a rectangle that straddles the midplane.

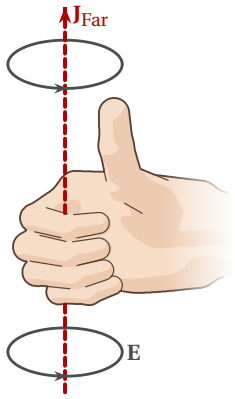


Figure 15.9 The right-hand rule for the Faraday current: with the thumb of your right hand along $\mathbf{J}_{\text{Far}} = -\partial\mathbf{B}/\partial t$ (dashed), your fingers curl around in the direction of the induced electric field **E**.

The fourth geometry, the toroid, has no counterpart. In a toroid whose current changes, $\mathbf{J}_{\text{Far}}$ runs around the ring inside the core, like the current in a loop of wire, and the induced **E**, like the magnetic field of a loop of current, cannot be found from Ampère's law alone. The EMF around a closed path that threads the core can still be found, because it is $-d\Phi_B/dt$ whatever the field looks like; this is how a transformer works.

The analogy with magnetostatics never breaks down for another reason. Because $\nabla \cdot \mathbf{B} = 0$ everywhere and at every instant, $\nabla \cdot \mathbf{J}_{\text{Far}} = -\partial(\nabla \cdot \mathbf{B})/\partial t = 0$: the Faraday current is always "steady" in the sense of Section 10.4.1. It never piles up anywhere, and its lines always close on themselves, as the lines of a steady current do. Every result of magnetostatics therefore has a counterpart for the induced

electric field, including the Biot–Savart law with μ_0 removed. The one condition is the one stated in Section 15.2.4: the method gives the field that the changing **B** produces, and if charges are present, the field they produce must be added to it.

15.3.1 Worked examples

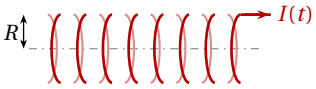
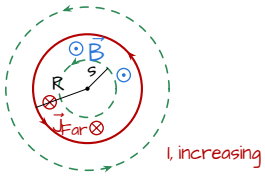


Figure 15.10 Part of the solenoid of Example 15.4, whose current is changing.

Example 15.4

A long solenoid of radius R has n turns of wire per unit length, and the current I in it is changing at the rate dI/dt. Find the induced electric field inside and outside the solenoid, far from its ends. Evaluate it at the surface of a solenoid with $n = 2000\,\mathrm{m}^{-1}$ and $R = 2.0\,\mathrm{cm}$, while its current increases at $100\,\mathrm{A/s}$.

Solution:



As always, the first step is a sketch of what is given and what is asked. We draw the solenoid end-on, with its current going around counterclockwise and increasing, so that **B** points out of the page and grows. The Faraday current then points into the page inside the solenoid. Its loops are circles of radius s, one inside and one outside, which we traverse counterclockwise.

Inside the solenoid the field is $\mathbf{B} = \mu_0 n I\,\hat{\mathbf{z}}$, and outside it is zero, (11.7), so

$$\mathbf{J}_{\text{Far}} = -\frac{\partial \mathbf{B}}{\partial t} = \begin{cases} -\mu_0 n \dfrac{dI}{dt}\hat{\mathbf{z}}, & s < R, \\ 0, & s > R. \end{cases}$$

This is a uniform "current density" filling a cylinder of radius R, like the current in the wire of Example 11.1, and the arguments given there carry over. The field can depend only on s. It has no z component, because by the Biot-Savart law with $\mathbf{J}_{\text{Far}}$ in place of $\mu_0\mathbf{J}$, each piece of the Faraday current contributes a field perpendicular to itself. It has no s component either: there are no charges, so $\nabla\cdot\mathbf{E} = 0$, and a closed cylinder coaxial with the solenoid can have no net flux of **E** through its curved side. So $\mathbf{E} = E(s)\,\hat{\boldsymbol{\phi}}$.

Going counterclockwise around a circle of radius s, the normal by the right-hand rule is $\hat{\mathbf{z}}$, and $\oint \mathbf{E}\cdot d\boldsymbol{\ell} = 2\pi s E_\phi$. Inside, the circle encloses the Faraday current through a disk of radius s,

$$I_{\text{Far,enc}} = -\mu_0 n \frac{dI}{dt}\pi s^2,$$

which is $-d\Phi_B/dt$ for that disk. So $2\pi s E_\phi = -\mu_0 n \pi s^2\, dI/dt$. Outside, the circle encloses all of the Faraday current, $-\mu_0 n \pi R^2\, dI/dt$, however large the circle is. To-

gether,

$$\mathbf{E} = \begin{cases} -\dfrac{\mu_0 n}{2}\dfrac{dI}{dt}\, s\hat{\boldsymbol{\phi}}, & s < R, \\ -\dfrac{\mu_0 n R^2}{2s}\dfrac{dI}{dt}\hat{\boldsymbol{\phi}}, & s > R. \end{cases} \tag{15.9}$$

When the current increases, the induced field points in the $-\hat{\boldsymbol{\phi}}$ direction, against the current in the windings, and opposes the increase, as Lenz's law requires. Outside the solenoid there is no magnetic field at all, but there is an electric field (Figure 15.11), and a charge placed there feels it.

For the given solenoid, $\mu_0 n\, dI/dt = (4\pi\times10^{-7}\,\mathrm{T\cdot m/A})(2000\,\mathrm{m^{-1}})(100\,\mathrm{A/s}) = 0.25\,\mathrm{T/s}$, and at the surface, $s = R$,

$$E = \frac{\mu_0 nR}{2}\frac{dI}{dt} = \frac{(0.25\,\mathrm{T/s})(0.020\,\mathrm{m})}{2} = 2.5\times10^{-3}\,\mathrm{V/m},$$

a weak field. A loop of wire around the solenoid, of any size and shape, has the EMF $\mu_0 n\pi R^2\, dI/dt = 0.32\,\mathrm{mV}$.

Reality check: The units are right: $\mathrm{T\cdot m/s} = \mathrm{V/m}$. The field is continuous at $s = R$, as it should be. (Why must the field be continuous at $s = R$? Answer in footnote.[8]) The curl in cylindrical coordinates is $(1/s)\,\partial(sE_\phi)/\partial s\,\hat{\mathbf{z}}$, which is $-\mu_0 n\,(dI/dt)\,\hat{\mathbf{z}} = \mathbf{J}_{\mathrm{Far}}$ inside and zero outside. And around one turn of the solenoid's own winding, at $s = R$, the EMF is $2\pi R E_\phi = -\mu_0 n\pi R^2\, dI/dt$, so around the $n\ell$ turns of a length ℓ it is $-\mu_0 n^2\pi R^2\ell\, dI/dt = -L\, dI/dt$, with the inductance L of (12.33). That is the back EMF of (15.5).

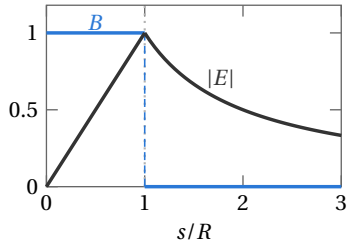


Figure 15.11 The fields of the solenoid against s/R, each in units of its value just inside the windings: **B** (blue) is uniform inside and zero outside, while the induced **E** (black) grows as s inside and falls as $1/s$ outside.

Example 15.5

A long coaxial cable has a solid inner conductor of radius a and a thin outer conductor of radius b. A current I flows up the inner conductor, spread uniformly over its cross section, and back down the outer one, and it is changing at the rate dI/dt. Find the electric field induced by the changing current, everywhere.

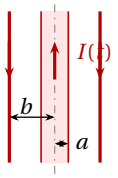
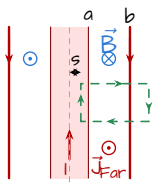


Figure 15.12 The coaxial cable of Example 15.5, cut along its axis.

[8] It is parallel to the surface $s = R$, and the parallel component of **E** is continuous at every surface, even when **B** changes (Section 15.2.3). In the language of the analogy, $\mathbf{J}_{\mathrm{Far}}$ has no surface part there, just as **B** is continuous where there is no surface current.

Solution:



We first sketch the cable cut along its axis, with the current increasing up the inner conductor and down the outer one. The field $\mathbf{B}$ circles the axis inside the outer conductor, into the page on the right of the axis, so $\mathbf{J}_{\text{Far}}$ circles it the other way, out of the page on the right, like the current in a solenoid. So we expect $\mathbf{E}$ to point along the axis, and we use a rectangle in the plane of the page, to the right of the axis.

By Ampère's law (Example 11.1), with the outer conductor adding nothing inside itself and canceling the inner conductor's field outside,

$$\mathbf{B} = \begin{cases} \dfrac{\mu_0 I s}{2\pi a^2}\,\hat{\boldsymbol{\phi}}, & s < a, \\ \dfrac{\mu_0 I}{2\pi s}\,\hat{\boldsymbol{\phi}}, & a < s < b, \\ 0, & s > b, \end{cases}$$

and $\mathbf{J}_{\text{Far}} = -\partial\mathbf{B}/\partial t$ is the same with I replaced by $-dI/dt$. It circles the axis and depends only on s, as the current in a solenoid does, and the arguments of Example 11.3 carry over: $\mathbf{E}$ has no ϕ component, because a circle around the axis encloses no Faraday current, and no s component, because $\nabla\cdot\mathbf{E} = 0$. So $\mathbf{E} = E(s)\,\hat{\mathbf{z}}$.

Outside the cable, a rectangle with both long sides outside encloses no Faraday current, so E is the same at every distance outside; as for the field outside a long solenoid, it is zero. Now take a rectangle of height ℓ with its inner side at a distance s from the axis and its outer side outside the cable, and go up the inner side. That is clockwise as drawn, so the normal points into the page, along $+\hat{\boldsymbol{\phi}}$. The inner side contributes $E(s)\,\ell$, the short sides are perpendicular to $\mathbf{E}$, and the outer side lies in zero field. Between the conductors, $a < s < b$, the Faraday current through the rectangle is

$$I_{\text{Far,enc}} = \int_s^b \left(-\frac{\mu_0}{2\pi\bar{s}}\frac{dI}{dt}\right)\ell\,d\bar{s} = -\frac{\mu_0\ell}{2\pi}\frac{dI}{dt}\ln\frac{b}{s},$$

which is $-d\Phi_B/dt$ for the rectangle. Inside the inner conductor, $s < a$, the rectangle also catches the Faraday current between s and a,

$$I_{\text{Far,enc}} = -\frac{\mu_0\ell}{2\pi}\frac{dI}{dt}\left[\ln\frac{b}{a} + \int_s^a \frac{\bar{s}}{a^2}\,d\bar{s}\right] = -\frac{\mu_0\ell}{2\pi}\frac{dI}{dt}\left[\ln\frac{b}{a} + \frac{a^2 - s^2}{2a^2}\right].$$

Dividing by ℓ,

$$\mathbf{E} = \begin{cases} -\dfrac{\mu_0}{2\pi}\dfrac{dI}{dt}\left[\ln\dfrac{b}{a} + \dfrac{a^2 - s^2}{2a^2}\right]\hat{\mathbf{z}}, & s < a, \\ -\dfrac{\mu_0}{2\pi}\dfrac{dI}{dt}\ln\dfrac{b}{s}\,\hat{\mathbf{z}}, & a < s < b, \\ 0, & s > b. \end{cases} \tag{15.10}$$

When the current increases, the induced field in the inner conductor points down, against the current (Figure 15.13), and the source that drives the current has to push against it. This is the self-inductance of the cable at work. The field is strongest on the axis and weakest at the surface of the inner conductor, so it opposes a change in the current most strongly at the center of the wire. When the current alternates rapidly, this pushes the current toward the surface of the wire, an effect called the skin effect (Section 22.2.2).

Reality check: The field is continuous at $s = a$ and goes to zero at $s = b$, from both sides. Its curl is $-(\partial E_z/\partial s)\,\hat{\boldsymbol{\phi}}$, which is $-(\mu_0 s/2\pi a^2)(dI/dt)\,\hat{\boldsymbol{\phi}} = \mathbf{J}_{\text{Far}}$ inside the inner conductor and $-(\mu_0/2\pi s)(dI/dt)\,\hat{\boldsymbol{\phi}} = \mathbf{J}_{\text{Far}}$ between the conductors.

The calculation ignores the field produced by charges on the conductors, which a real cable also has: a radial field between the conductors, wherever there is a voltage between them, and a small field along each conductor that drives the current through its resistance. Those fields add to this one.

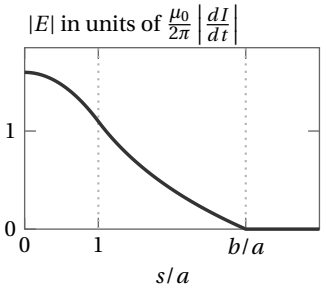


Figure 15.13 The size of the induced field of Example 15.5 against s/a, for $b = 3a$: largest on the axis, falling as $\ln(b/s)$ between the conductors, and zero outside.

15.4 Fields from potentials: time dependence

The potentials of Chapters 5 and 12 were introduced for static fields: $\mathbf{E} = -\nabla V$ because $\nabla \times \mathbf{E} = 0$, and $\mathbf{B} = \nabla \times \mathbf{A}$ because $\nabla \cdot \mathbf{B} = 0$. The second reason still holds when the fields change, since $\nabla \cdot \mathbf{B} = 0$ at every instant, so $\mathbf{B}$ can still be written as the curl of a vector potential. The first reason does not: by Faraday's law, the curl of $\mathbf{E}$ is no longer zero, and $\mathbf{E}$ is no longer the gradient of anything. Put $\mathbf{B} = \nabla \times \mathbf{A}$ into Faraday's law,

$$\nabla \times \mathbf{E} = -\frac{\partial}{\partial t}(\nabla \times \mathbf{A}) = -\nabla \times \frac{\partial \mathbf{A}}{\partial t},$$

and move the right side to the left:

$$\nabla \times \left(\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t}\right) = 0.$$

The combination in parentheses has no curl, so it is the gradient of a scalar function (Section 5.3.1), which we write as $-V$. Then

$$\boxed{\mathbf{E} = -\nabla V - \frac{\partial \mathbf{A}}{\partial t}, \qquad \mathbf{B} = \nabla \times \mathbf{A}}. \tag{15.11}$$

This is the revision promised in Section 5.3.1. When nothing changes, the new term is zero and $\mathbf{E} = -\nabla V$, (5.6), as before.

In the Coulomb gauge the two terms are the field that the charges would produce by Coulomb's law and the induced field of Section 15.2.4 (Figure 15.14). The term $-\partial \mathbf{A}/\partial t$ is the induced field. When the currents change slowly enough

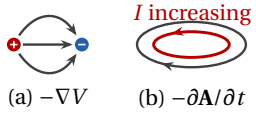


Figure 15.14 The two parts of $\mathbf{E} = -\nabla V - \partial \mathbf{A}/\partial t$: (a) $-\nabla V$, the Coulomb field of the charges; (b) $-\partial \mathbf{A}/\partial t$, induced by changing currents, which for a growing current in a loop circles opposite to the current.

(the quasistatic approximation of Section 16.4.1), **A** at each instant is given by the currents at that instant through (12.9), so the induced field is

$$-\frac{\partial \mathbf{A}}{\partial t} = -\frac{\mu_0}{4\pi}\int \frac{\partial \mathbf{J}/\partial t}{\mathscr{r}}\, d\tau',$$

and it points, roughly, opposite to the change in the currents nearby, which is Lenz's law again.

The term $-\nabla V$ is the charges' Coulomb field. Gauss's law, with (15.11), reads

$$\nabla\cdot\mathbf{E} = -\nabla^2 V - \frac{\partial}{\partial t}(\nabla\cdot\mathbf{A}) = \frac{\rho}{\varepsilon_0},$$

and in the Coulomb gauge of Section 12.1.1, $\nabla\cdot\mathbf{A} = 0$, so

$$\nabla^2 V = -\frac{\rho}{\varepsilon_0}$$

at every instant, even when the charges move. In the Coulomb gauge, then, the scalar potential at each instant is the potential that Coulomb's law gives for the charges where they are at that instant, (5.16). That does not mean that a change in the charges is felt instantly far away. The potential is not something an experiment measures; the field is, and when the charges move, part of $\partial\mathbf{A}/\partial t$ cancels the instantaneous change in ∇V, so that **E** changes only after light has had time to arrive. Chapter 24 takes up potentials that account for the travel time directly, in another gauge.

Therefore, if the electric potential V does not vary from place to place, **E** will equal $-\partial\mathbf{A}/\partial t$; in the Coulomb gauge that is the case whenever there are no charges anywhere, since then $V = 0$. As a direct example of this, refer back to the solenoid of Example 15.4: $-\partial\mathbf{A}/\partial t$, with the vector potential (12.15) that the flux method of Section 12.4 gave, is exactly the field (15.9), inside and outside. (For the coaxial cable of Example 15.5, compare (15.10) with your answer to Exercise P12.10.) The two methods are the same calculation, loop for loop: in the flux method the circulation of **A** around a loop is the flux of **B** through it, and in the Faraday-current method the circulation of **E** is the negative of the rate of change of that flux.

Gauge freedom survives in a modified form. Adding $\nabla\psi$ to **A** leaves **B** unchanged, as in Section 12.1.1, but it changes $-\partial\mathbf{A}/\partial t$ by $-\nabla(\partial\psi/\partial t)$, so to leave **E** unchanged, V must change at the same time:

$$\mathbf{A} \to \mathbf{A} + \nabla\psi, \qquad V \to V - \frac{\partial\psi}{\partial t}. \tag{15.12}$$

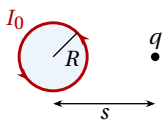
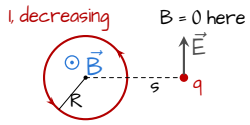


Figure 15.15 The solenoid of Example 15.6, end on, and the charge outside it.

Example 15.6

A particle with charge q sits at rest at a distance s from the axis of a long solenoid of radius $R < s$, with n turns per unit length and current I_0. The current is switched off. Find the momentum that the particle gains, assuming that it moves only a negligible distance while the current falls.

Solution:



As always, the first step is a sketch of what is given and what is asked: the solenoid end-on, with its current going around counterclockwise and decreasing, and the charge outside it, where there is no magnetic field. While the current falls, the induced field there circles the axis in the same sense as the current, and so does the push on a positive charge.

The particle is outside the solenoid, where $\mathbf{B}$ is zero before, during, and after the change, so no magnetic force ever acts on it. The induced field outside is (15.9),

$$\mathbf{E} = -\frac{\mu_0 n R^2}{2s}\frac{dI}{dt}\hat{\boldsymbol{\phi}},$$

which points along $+\hat{\boldsymbol{\phi}}$, the direction of the current, while dI/dt is negative. The momentum the particle gains is the integral of the force $q\mathbf{E}$ over the time the current takes to fall:

$$\Delta\mathbf{p} = \int q\mathbf{E}\,dt = -\frac{q\mu_0 n R^2}{2s}\hat{\boldsymbol{\phi}}\int_{I_0}^{0} dI = \frac{q\mu_0 n R^2 I_0}{2s}\hat{\boldsymbol{\phi}}.$$

It does not depend on how quickly the current falls. A fast switch gives a large force for a short time, a slow one a small force for a long time, and the product is the same. Compare the vector potential outside the solenoid, (12.15), before the switch: $\mathbf{A} = (\mu_0 n I_0 R^2/2s)\,\hat{\boldsymbol{\phi}}$. The momentum the particle gains is $q\mathbf{A}$.

That is no coincidence. Here there are no charges (we ignore the particle's own field, which exerts no force on it), so $V = 0$ and $\mathbf{E} = -\partial\mathbf{A}/\partial t$, and

$$\Delta\mathbf{p} = -q\int \frac{\partial\mathbf{A}}{\partial t}\,dt = q\left(\mathbf{A}_{\text{before}} - \mathbf{A}_{\text{after}}\right) = q\mathbf{A}_{\text{before}},$$

with $\mathbf{A}$ in the Coulomb gauge, going to zero far away, so that $\mathbf{A}_{\text{after}} = 0$. This is the result promised in the footnote in Section 12.1, and the reason that Maxwell called $\mathbf{A}$ the electromagnetic momentum.

Reality check: The units of $q\mathbf{A}$ are $\mathrm{C}\cdot\mathrm{T}\cdot\mathrm{m} = \mathrm{C}\cdot(\mathrm{N}/(\mathrm{A}\cdot\mathrm{m}))\cdot\mathrm{m} = \mathrm{N}\cdot\mathrm{s}$, those of momentum, and the direction agrees with Lenz's law: the induced field pushes charges the way that would keep the current, and the flux, from falling.

15.5 Numerical method: a magnet falling through a coil

A favorite demonstration of Faraday's law, and a common laboratory exercise, is to drop a small magnet through a coil of wire and record the EMF as it falls. The EMF depends on how fast the magnet moves, and if the coil is part of a closed

circuit, so that current flows, the force of that current on the magnet changes how fast it moves. The equation of motion that results can rarely be solved by hand, but it can be solved numerically by the time stepping of Section 10.1.5.

15.5.1 The model

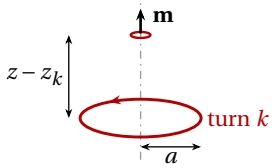


Figure 15.16 One turn of the coil and the magnet, a tiny loop of moment **m** on the axis. The flux of the magnet's field through the turn is m times the field that the turn would produce at the magnet per unit current, because $M_{12} = M_{21}$.

Treat the magnet as a point dipole $\mathbf{m} = m\hat{\mathbf{z}}$ of mass M, falling along the z axis, and the coil as N circular turns of radius a, centered on the axis at heights $z_1, \ldots, z_N$, connected in a circuit of total resistance R. The calculation has three pieces.

The flux. The flux of the magnet's field through a turn is hard to calculate directly, because the dipole field is complicated and the surface is large. It is easy to calculate the other way around, because the mutual inductance of two loops is the same both ways, as the Neumann formula (12.31) shows (Figure 15.16). Think of the magnet as a tiny loop of area A_m carrying a current I_m, so that $m = I_m A_m$. The flux through the turn per unit current in the tiny loop equals the flux through the tiny loop per unit current in the turn, which is $B_{\text{turn}} A_m / I_{\text{turn}}$, with B_{turn} the field of the turn at the magnet. Multiplying by I_m, the flux through turn k is m times the field that turn k would produce at the magnet, per unit current in it, which is (10.22):

$$\Phi_k(z) = \frac{\mu_0 m a^2}{2\left[a^2 + (z - z_k)^2\right]^{3/2}}. \tag{15.13}$$

The flux through the coil is the sum, $\Phi(z) = \sum_k \Phi_k(z)$, and it depends on the magnet's height z.

The EMF and the current. By the chain rule, the EMF is

$$\mathscr{E} = -\frac{d\Phi}{dt} = -\Phi'(z)\, v,$$

where $v = dz/dt$ is the magnet's velocity (negative while it falls) and $\Phi' = d\Phi/dz$. If the coil's own inductance is small enough to ignore (Exercise P15.13), the current is $I = \mathscr{E}/R$, counted positive in the sense whose normal is $+\hat{\mathbf{z}}$.

The force. The current in the coil produces a field at the magnet, and the force on a dipole is $\mathbf{F} = \nabla(\mathbf{m} \cdot \mathbf{B})$, (13.3). By the reciprocity argument above, m times the coil's field at the magnet is $I\Phi(z)$, so the force is

$$F_z = I\Phi'(z) = -\frac{\left[\Phi'(z)\right]^2}{R}\, v. \tag{15.14}$$

It is always opposite to v, as Lenz's law requires, and the power it takes from the magnet, $-F_z v = I^2 R$, is exactly the power dissipated in the resistance.

The equation of motion is $M\, dv/dt = -Mg + I\Phi'(z)$. Each time step evaluates the current and the force at the present position and velocity, and updates the

velocity and then the position:

$$I_n = -\frac{\Phi'(z_n)\, v_n}{R},$$
$$v_{n+1} = v_n + \left[-g + \frac{I_n \Phi'(z_n)}{M}\right]\Delta t,$$
$$z_{n+1} = z_n + v_{n+1}\,\Delta t.$$

Three checks are worth building in. Repeat the run with a smaller Δt until the results stop changing. Check that energy is conserved: the potential energy the magnet loses, $Mg(z_0 - z)$, should equal its kinetic energy plus the energy dissipated so far, $\sum I_n^2 R\,\Delta t$. And with the circuit open ($R \to \infty$) the magnet should fall freely, as it does in the absence of the coil, with $z = z_0 - \frac{1}{2}gt^2$.

15.5.2 Tips for Mathematica and Python

- Write $\Phi'(z)$ as a sum over the turns. In Python, with the heights z_k in a NumPy array, `np.sum` over an expression in `z - zk` evaluates it without a loop; in Mathematica, `Total` over a list does the same.
- The general-purpose solvers, `solve_ivp` in Python's `scipy.integrate` and `NDSolve` in Mathematica, handle this equation well. Write it as two first-order equations, for z and v, and stop the run when the magnet is well below the coil (an "event" in `solve_ivp`, `WhenEvent` in `NDSolve`).
- Plot the EMF against time rather than against position: that is what an oscilloscope or a data logger records.

Example 15.7

A neodymium magnet about 1 cm on a side has the dipole moment $m = 1.0\,\mathrm{A\cdot m^2}$ and the mass $M = 7.5\,\mathrm{g}$. It is dropped from rest, with its north pole up, from a height of 10 cm above the middle of a coil of 500 turns of radius 1.2 cm, wound evenly over a length of 2.0 cm, and falls through the coil along its axis. Find the EMF in the coil as a function of time, (a) with the coil connected to a voltmeter, so that no current flows, and (b) with the coil connected in a circuit whose total resistance is 7.0 Ω. How much energy is dissipated in case (b)?

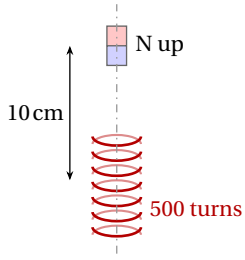


Figure 15.17 The magnet and the coil of Example 15.7, not to scale.

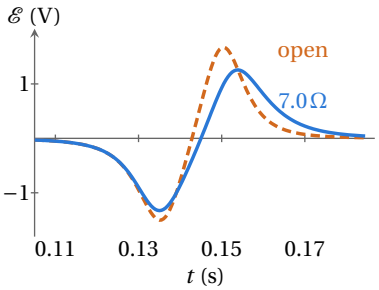
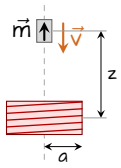


Figure 15.18 The EMF in the coil of Example 15.7 as the magnet falls through it, with the coil open (dashed) and connected in a circuit of resistance $7.0\,\Omega$ (solid). Time is measured from the release; the magnet passes the middle of the coil at about 0.14 s. Animation of both falls, in slow motion.

Solution:



We first sketch what is given: the coil from the side, with its turns spread over its length, the magnet above it on the axis with its moment pointing up, and the height z of the magnet measured from the middle of the coil. As the magnet approaches, the upward flux through the coil grows, so we expect the EMF to be negative at first and positive after the magnet passes.

Put the turns at 500 evenly spaced heights between $z = -1.0\,\text{cm}$ and $z = +1.0\,\text{cm}$, start the magnet at $z_0 = 10\,\text{cm}$ with $v = 0$, and step until it is 10 cm below the coil. Steps of $\Delta t = 10^{-4}\,\text{s}$ give results within 0.1% of those with $\Delta t = 10^{-5}\,\text{s}$, and even $10^{-3}\,\text{s}$ is within about 1%.

(a) With no current, the magnet falls freely, and the EMF is two pulses of opposite sign (Figure 15.18). As the magnet approaches, the upward flux through the coil grows, and the EMF is negative, reaching $-1.50\,\text{V}$ at $t = 0.135\,\text{s}$. As it leaves, the flux falls and the EMF is positive, reaching $+1.67\,\text{V}$ at $t = 0.150\,\text{s}$. The second pulse is taller and narrower than the first because the magnet is moving faster: it passes the middle of the coil at $1.40\,\text{m/s}$ and is still speeding up. The area under each pulse, $\int \mathcal{E}\,dt = -\Delta\Phi$, is the same in size, $0.020\,\text{Wb}$, because the flux rises from nearly zero to its maximum and falls back again. That area does not depend on how the magnet moves, so it is the same in (b).

(b) With current flowing, the coil brakes the magnet. It passes the middle of the coil at $1.13\,\text{m/s}$ instead of $1.40\,\text{m/s}$, and the pulses are $-1.33\,\text{V}$ at $0.135\,\text{s}$ and $+1.26\,\text{V}$ at $0.154\,\text{s}$: now the second pulse is the smaller one, because the magnet leaves the coil more slowly than it would have in free fall. By the time it is 10 cm below the coil it is moving at $1.62\,\text{m/s}$ instead of $1.98\,\text{m/s}$. The energy dissipated in the circuit is $4.9\,\text{mJ}$, a third of the $14.7\,\text{mJ}$ of potential energy the magnet loses in falling 20 cm.

Reality check: The rest is the magnet's kinetic energy, $\frac{1}{2}(7.5\,\text{g})(1.62\,\text{m/s})^2 = 9.8\,\text{mJ}$: the energy check of Section 15.5.1 holds, to four figures in the program.

The model is rough near the coil, where the magnet is not small compared with its distance from the wire, but it reproduces what is seen in the laboratory: two pulses of opposite sign, the second sharper when the coil is open, and a magnet that slows when the coil is closed.

15.6 Summary

When the magnetic field changes in time, the curl equation for the electric field gains a term: a changing magnetic field produces an electric field that curls around $-\partial\mathbf{B}/\partial t$ in exactly the way that the magnetic field curls around a current.

The next chapter does the same for the curl of the magnetic field, which completes Maxwell's equations.

EMF and induction

$\mathcal{E} = \oint \mathbf{f} \cdot d\boldsymbol{\ell}$	the EMF of a circuit: the work per unit charge done once around it, by whatever pushes the charges; in volts
$\mathcal{E} = vB\ell$	the motional EMF of a bar of length ℓ moving at speed v across a field B
$\mathcal{E} = -\dfrac{d\Phi_B}{dt}$	the flux rule, for a circuit that moves through a static field; Faraday's law, for a fixed circuit in a changing field; both together when both happen
Lenz's law	the induced current flows in the direction that makes its own field oppose the *change* in the flux
$\mathcal{E}_2 = -M\dfrac{dI_1}{dt}, \quad \mathcal{E} = -L\dfrac{dI}{dt}$	EMFs of mutual induction and of self-induction (the back EMF), for circuits that do not move

Faraday's law and the Faraday current

$\nabla \times \mathbf{E} = -\dfrac{\partial \mathbf{B}}{\partial t}, \quad \oint \mathbf{E} \cdot d\boldsymbol{\ell} = -\dfrac{d}{dt}\int \mathbf{B} \cdot d\mathbf{a}$	Faraday's law: a changing magnetic field produces an electric field with curl
$\mathbf{J}_{\text{Far}} = -\dfrac{\partial \mathbf{B}}{\partial t}, \quad \nabla \times \mathbf{E} = \mathbf{J}_{\text{Far}}, \quad \oint \mathbf{E} \cdot d\boldsymbol{\ell} = I_{\text{Far,enc}}$	the Faraday current density: $\mathbf{E}$ curls around $\mathbf{J}_{\text{Far}}$ as $\mathbf{B}$ curls around $\mathbf{J}$, without the μ_0; $\mathbf{J}_{\text{Far}}$ points opposite to the change in $\mathbf{B}$

Fields from potentials

$\mathbf{E} = -\nabla V - \dfrac{\partial \mathbf{A}}{\partial t}, \quad \mathbf{B} = \nabla \times \mathbf{A}$	the fields from the potentials; $-\partial\mathbf{A}/\partial t$ is the induced field
$\mathbf{A} \to \mathbf{A} + \nabla\psi, \quad V \to V - \dfrac{\partial \psi}{\partial t}$	a gauge transformation, which changes neither $\mathbf{E}$ nor $\mathbf{B}$

The Faraday-current method. When $\mathbf{J}_{\text{Far}} = -\partial\mathbf{B}/\partial t$ has the symmetry of a long wire, a long solenoid, or a slab, find $\mathbf{E}$ exactly as Ampère's law finds $\mathbf{B}$: use symmetry to find the direction of $\mathbf{E}$, draw an Amperian loop, find $I_{\text{Far,enc}} = -d\Phi_B/dt$, and solve $\oint \mathbf{E} \cdot d\boldsymbol{\ell} = I_{\text{Far,enc}}$. *Colton's Third Commandment* applies: thou shalt always draw an Amperian loop.

Numerical method. When an induced current acts back on the motion that produces it, as for a magnet falling through a coil, step the equation of motion

in time, with the EMF from the flux rule and the force from the induced current. Check the step size, and check that the energy lost by the moving object equals the energy dissipated.

Exercises

Exercises for 15.1 Motional EMF

Warm-up questions Read Section 15.1, then answer before class

W15.1 Suppose the resistor of Figure 15.2 is removed, leaving a gap between the rails, and the bar is pushed to the right at constant speed as before. Once things settle, is there a current? Is there still an EMF, and what would a voltmeter connected across the gap read? Is a push still needed to keep the bar moving? (Compare the ideal battery of Section 15.1.1.)

W15.2 In Example 15.1, the current that heats the resistor is driven by a magnetic force, yet magnetic forces do no work (Section 10.1.3). Using Figure 15.4, explain how both can be true: which part of the magnetic force on a charge in the bar does positive work, which does negative work, and where does the energy come from?

W15.3 Section 15.1.3 goes around the sliding-bar circuit counterclockwise. Redo its sign bookkeeping going clockwise instead: which way does the normal of S point, and what are Φ_B, $d\Phi_B/dt$, and $\mathscr{E}$? Which way does the current flow, and does that agree with what the section found?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P15.1 The bar of Example 15.1 ($B = 0.40\,\text{T}$, $\ell = 25\,\text{cm}$, $R = 0.50\,\Omega$) has a mass of 50 g. It is given a speed of 2.0 m/s to the right and then released, and it slides without friction.

(a) Show that its speed decreases as $v = v_0 e^{-t/\tau}$, and find τ.

(b) How far does the bar slide before it stops?

(c) Show that the total energy dissipated in the resistor is the bar's initial kinetic energy.

P15.2 A circular loop of wire of radius a rotates with constant angular velocity ω about the y axis, which is one of its diameters, in a uniform field $\mathbf{B} = B_0\hat{\mathbf{z}}$ (Figure 15.19). At $t = 0$ the loop lies in the xy plane, and it turns in the right-handed sense about $+\hat{\mathbf{y}}$, so that its normal is $\hat{\mathbf{n}} = \cos\omega t\,\hat{\mathbf{z}} + \sin\omega t\,\hat{\mathbf{x}}$.

(a) Find the EMF as a function of time, taking positive EMF to drive current in the direction related to $\hat{\mathbf{n}}$ by the right-hand rule.

(b) As seen from $+z$, which way does the current flow just before the loop completes one full turn, at that instant, and just after?

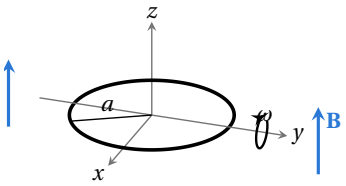


Figure 15.19 The loop of Exercise P15.2 at $t = 0$, in the xy plane, rotating about the y axis in a uniform field along $\hat{\mathbf{z}}$.

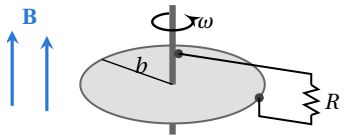


Figure 15.20 The disk of Exercise P15.3, turning in a uniform field parallel to its axle, with sliding contacts on the axle and on the rim.

(c) Find the peak EMF for $a = 10\,\text{cm}$, $B_0 = 0.20\,\text{T}$, and 50 turns per second.

P15.3 A metal disk of radius b turns with angular velocity ω about its axle in a uniform field **B** parallel to the axle (Figure 15.20). Sliding contacts touch the axle and the rim and connect them to an external circuit. This is a *homopolar generator*, or Faraday disk.

(a) First take a thin metal rod of length b turning about one end in the same field. Integrate $\mathbf{v} \times \mathbf{B}$ along the rod to show that the EMF between its ends is $\frac{1}{2}\omega B b^2$. Which end is positive, if **B** points along the angular velocity?

(b) Explain why the EMF between the axle and the rim of the disk is the same.

(c) Evaluate it for $b = 10\,\text{cm}$, $B = 0.50\,\text{T}$, and 3000 revolutions per minute.

(d) The circuit, made of the external wires and a radial line in the disk from the axle to the contact on the rim, does not move, and the flux through it does not change. Yet there is an EMF. Which assumption in the proof of the flux rule (Section 15.1.3) does this circuit fail?

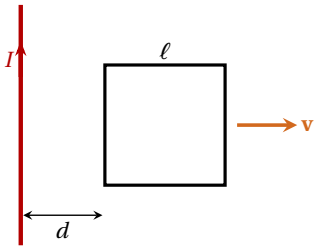


Figure 15.21 The wire and the square loop of Exercise P15.4, with the near side of the loop a distance d from the wire.

P15.4 A long straight wire carries a steady current I. A square loop of wire, of side ℓ, lies in a plane containing the wire, with two of its sides parallel to the wire, and moves directly away from it at speed v (Figure 15.21). At the moment of interest, the near side of the loop is a distance d from the wire.

(a) Find the flux through the loop as a function of d, and use the flux rule to find the EMF.

(b) Find the EMF again by integrating $\mathbf{v} \times \mathbf{B}$ around the loop, and show that the two answers agree.

(c) Which way does the current flow in the loop? Does the wire attract or repel the loop?

Exercises for 15.2 Faraday's law and the Faraday current

Warm-up questions Read Section 15.2, then answer before class

W15.4 In Faraday's ring experiment (Section 15.2.1), the galvanometer needle jumped one way when he completed the battery circuit and the other way when he broke it. Suppose instead that, while a steady current flowed, he had (a) reversed the battery's connections; (b) added a second battery in series, doubling the current. In each case, would the needle jump, and which way, compared with its jump when he completed the circuit? Explain with the flux through the second coil.

W15.5 Section 15.2.2 calls Lenz's law a statement of energy conservation. Imagine a world in which induced currents reinforced the change in flux instead of opposing it. The north pole of a magnet is nudged toward a copper ring and released. Describe what would happen next, and why it would violate the conservation of energy.

W15.6 A classmate predicts that a narrow slit cut along the whole length of the copper tube of Figure 15.8 would hardly change the magnet's slow fall, since the slit removes almost no copper. Do you agree? Explain with the paths of the eddy currents, and compare the slotted plate of Section 15.2.5.

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P15.5 A square loop of side a lies in the xy plane (Figure 15.22), with corners at the origin and at $(a, a, 0)$. A magnetic field $\mathbf{B} = B_0\,(x/a)^2 \cos\omega t\,\hat{\mathbf{z}}$ fills the region.

(a) Find the EMF around the loop, and the direction, as seen from $+z$, in which it drives current at $t = \pi/2\omega$.

(b) This field has a curl. What must be present in the region to produce it?

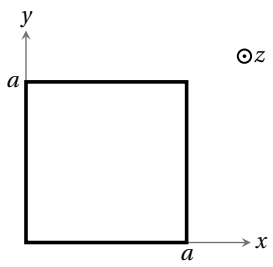


Figure 15.22 The square loop of Exercise P15.5, in the xy plane, with corners at the origin and at $(a, a, 0)$.

P15.6 Use Lenz's law to answer each question, and explain your reasoning.

(a) A bar magnet is held with its north pole down above a horizontal ring of wire and dropped, so that it falls through the ring. Seen from above, which way does the induced current circulate while the magnet approaches the ring, and after it has passed through? In each case, is the force of the ring on the magnet up or down?

(b) Two loops of wire face each other on a common axis. A current is switched on in loop 1. While it increases, do the loops attract or repel each other?

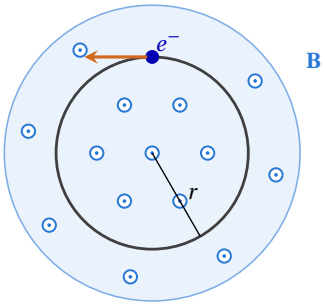


Figure 15.23 The orbit of Exercise P15.7, a circle of radius r in the midplane of an axially symmetric field that increases with time.

(c) A loop of flexible wire lies in a uniform field that points into the page, and it is squeezed so that its area decreases. Which way does the induced current flow?

(d) A loop of wire is moved parallel to a long straight wire that carries a steady current, keeping its distance from the wire fixed. Is there an induced current?

P15.7 A betatron accelerates electrons around a circle of fixed radius r in the midplane of an electromagnet whose field is symmetric about the axis of the circle and increases with time (Figure 15.23).

(a) Show that the electric field along the orbit has the magnitude

$$E = \frac{1}{2\pi r}\frac{d\Phi_B}{dt},$$

where Φ_B is the flux through the orbit, whether or not the field inside the orbit is uniform.

(b) An electron with momentum p moves on a circle of radius p/eB_{orbit} in the field B_{orbit} at the orbit (Section 10.1.4); the result holds relativistically if p is the relativistic momentum. Use $dp/dt = eE$ to show that the radius stays fixed if B_{orbit} is always half the average field inside the orbit, $\Phi_B/\pi r^2$, when both start from zero.

(c) For $r = 0.10\,\text{m}$, B_{orbit} rises from zero to $0.40\,\text{T}$ in $4.0\,\text{ms}$. Find the average energy an electron gains per turn, its final kinetic energy, and roughly how many times it goes around. How far does it travel?

Exercises for 15.3 Solving for E when given a changing B: Faraday-current method (high symmetry)

Warm-up questions Read Section 15.3, then answer before class

W15.7 For Ampère's law, every surface bounded by an Amperian loop gives the same I_{enc} only when the currents are steady. Is the same true of $I_{\text{Far,enc}}$, however **B** changes? Why? (Which of Maxwell's equations decides it?)

W15.8 Suppose a thin insulating rod with a uniform charge per unit length λ lies along the axis of the solenoid of Example 15.4. Which step of the example's argument about the direction of **E** no longer holds? Does the Faraday-current method's E_ϕ change? What is the total **E**?

W15.9 Example 15.5 finds that, while the current increases, the induced field in the inner conductor is strongest on the axis and weakest at the surface. In your own words, why does this push a rapidly alternating current toward the surface of the wire? For the same amplitude of current, would the effect be stronger at higher or at lower frequency? (How does dI/dt depend on the frequency?)

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P15.8 A thin flat sheet of metal of thickness w and conductivity σ lies between the planes $x = -w/2$ and $x = w/2$, in a uniform field $\mathbf{B} = B_0 \sin\omega t\, \hat{\mathbf{z}}$ parallel to the sheet (Figure 15.24). Far from the sheet's edges, the eddy currents flow along $\pm\hat{\mathbf{y}}$, and as much flows one way as the other. Ignore the magnetic field of the eddy currents themselves.

(a) Use a rectangular loop in the xy plane that straddles the midplane of the sheet, with its sides parallel to $\hat{\mathbf{y}}$ at x and $-x$, to show that the electric field in the sheet is $E_y = -x\,dB/dt$. (By symmetry, $E_y(-x) = -E_y(x)$.)

(b) Show that the power dissipated per unit volume, averaged over the sheet and over a cycle, is $\sigma\omega^2 B_0^2 w^2/24$.

(c) A transformer core of thickness D is built of N insulated sheets, each D/N thick. By what factor does that reduce the eddy-current loss, compared with a solid block of the same size?

(d) Evaluate the loss per kilogram for sheets of silicon steel 0.35 mm thick ($\sigma \approx 2\times10^6$ S/m, density 7650 kg/m^3),[9] with $B_0 = 1.0$ T at 60 Hz. What would the same formula give for a solid block 5 cm thick? (For so thick a block the field of the eddy currents can no longer be ignored, and the formula overestimates the loss, but the conclusion stands.)

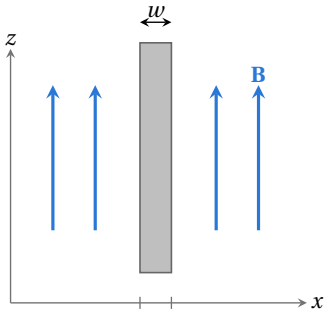


Figure 15.24 The sheet of Exercise P15.8, seen edge on, with the field **B** parallel to it, along $\hat{\mathbf{z}}$.

P15.9 A thin plastic disk of radius b is free to turn about its axis, and a charge Q is fixed to it in small beads spaced evenly around its rim (Figure 15.25). A long solenoid of radius $R < b$, with n turns per unit length and current I_0, passes through a hole at the center of the disk along its axis, without touching it. Everything is at rest. Then the current in the solenoid is switched off.

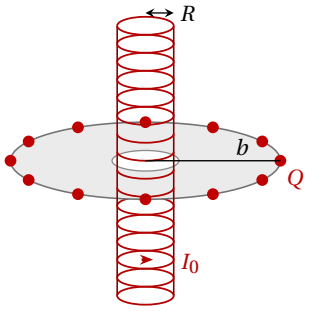


Figure 15.25 The disk of Exercise P15.9, with charged beads around its rim and a long solenoid passing through a hole at its center, without touching it.

[9] Iron with 3.0 to 3.4 percent silicon by weight (5.7 to 6.5 atomic percent) has a resistivity of about 4.5 to 5.0 $\times 10^{-7}$ Ω m at 300 K, so $\sigma \approx 2.2$ to 2.0×10^6 S/m: "Matthiessen's rule and the electrical resistivity of iron-silicon solid solutions," F. C. Schwerer, J. W. Conroy, and S. Arajs, J. Phys. Chem. Solids 30, 1513–1525 (1969). https://doi.org/10.1016/0022-3697(69)90213-3. The density is that of grain-oriented silicon steel, used in transformer cores: ATI Allegheny Ludlum, *Grain-Oriented Electrical Steel*, Technical Data Sheet, Version 1 (2012), Table 1. https://www.spacematdb.com/spacemat/manudatasheets/crgo.pdf.

(a) Find the torque on the disk while the current falls, and show that the disk ends up with the angular momentum $\frac{1}{2}Q\mu_0 nI_0R^2$. In which sense does it turn, compared with the current in the solenoid?

(b) Show that the result depends neither on b nor on how quickly the current is switched off.

(c) Before the current was switched off, nothing was turning; afterward the disk is turning, and nothing turns the other way. Angular momentum seems not to be conserved. (The resolution, that the electromagnetic field itself carried angular momentum, is in Section 20.5; Feynman discusses this puzzle in Vol. II, Section 17-4, of his lectures.) Before the switch, where are both **E** and **B** nonzero? A real solenoid has ends.

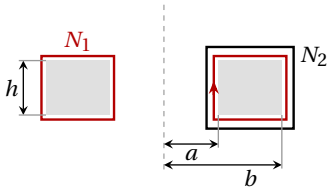


Figure 15.26 The toroid of Exercise P15.10, cut through its axis to show the rectangular cross section of its core, with the second coil wound around the core.

P15.10 A toroid has N_1 turns wound on a core of rectangular cross section, with inner radius a, outer radius b, and height h (Figure 15.26), and carries a changing current $I_1(t)$. A second coil of N_2 turns is wound around the same core.

(a) Use the field of a toroid (Example 11.4) to find the flux through the core and the EMF in the second coil. What is the mutual inductance?

(b) Evaluate the EMF for $N_1 = 1000$, $N_2 = 50$, $a = 2.0\,\text{cm}$, $b = 3.0\,\text{cm}$, $h = 1.0\,\text{cm}$, and $dI_1/dt = 100\,\text{A/s}$.

(c) A large loop of wire that goes once through the hole of the toroid, far from the windings, has an EMF of one turn's worth, even though $\mathbf{B} = 0$ everywhere along it. Explain.

(d) Explain why the Faraday-current method cannot give the induced **E** around the toroid itself.

P15.11 For the coaxial cable of Example 15.5:

(a) The source that drives the current must push it against the induced field. Show that the power per unit length it supplies to do so, $P' = -\int \mathbf{E}\cdot\mathbf{J}\,da$ over the cross section, is $L'I\,dI/dt$, with

$$L' = \frac{\mu_0}{2\pi}\left[\ln\frac{b}{a} + \frac{1}{4}\right].$$

(The thin outer conductor contributes nothing, since $\mathbf{E} = 0$ there.)

(b) Show that $P' = d(\frac{1}{2}L'I^2)/dt$, and that $\frac{1}{2}L'I^2$ is the energy per unit length stored in the magnetic field, (12.35). So L' is the inductance per unit length, (12.37).

(c) Which part of L' comes from the field inside the inner conductor?

Exercises for 15.4 Fields from potentials: time dependence

Warm-up questions Read Section 15.4, then answer before class

W15.10 In the sliding-bar circuit of Example 15.1, the uniform field **B** does not change in time, yet a current flows. Which term of (15.11) drives the current through the resistor? Is there an induced electric field anywhere in the circuit? Where, then, does the circuit's EMF come from?

W15.11 In Example 15.6, how much momentum would the particle gain, and in which direction, if instead (a) the current were switched on, from zero to I_0; (b) the current were reversed, from I_0 to $-I_0$? Check the direction in (a) with Lenz's law.

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P15.12 A static, uniform electric field $E_0\hat{\mathbf{z}}$ can be described by the potentials $V = -E_0 z$ and $\mathbf{A} = 0$.

(a) Find a gauge function ψ for which the new scalar potential is zero, and find the new vector potential.

(b) Check that the new potentials give the same **E** and **B**.

(c) A static field is now described by a vector potential that grows without limit in time. Is anything wrong?

Exercises for 15.5 Numerical method: a magnet falling through a coil

Warm-up questions Read Section 15.5, then answer before class

W15.12 Suppose the magnet of Example 15.7 were dropped with its north pole down instead of up. How would the EMF pulses change? Would the coil brake the magnet more, less, or the same? (Look at how m enters (15.13) and (15.14).)

W15.13 Suppose the magnet of Example 15.7 were dropped from 40 cm above the middle of the open coil instead of 10 cm. A classmate predicts that both EMF pulses would be taller and the areas under them larger, since the magnet would move faster. Is the classmate right about the heights? About the areas? Why?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P15.13 The coil of Example 15.7 has a self-inductance of about 4.6 mH.

(a) With the inductance included, the current obeys $L\,dI/dt + RI = \mathscr{E}$. Add this to the time-stepping scheme of Section 15.5.1, and repeat part (b) of the example. How much do the pulses and the speed of the magnet change?

(b) Explain the size of the change by comparing L/R with the duration of a pulse.

P15.14 A magnet like that of Example 15.7 ($m = 1.0\,\mathrm{A\cdot m^2}$, $M = 7.5\,\mathrm{g}$) falls down the axis of a long vertical copper tube (Figure 15.8) of mean radius $a = 10\,\mathrm{mm}$ and wall thickness $w = 1.0\,\mathrm{mm}$, with $\sigma = 5.96\times10^7\,\mathrm{S/m}$. Treat the tube as a stack of thin rings, each of height dz and resistance $2\pi a/(\sigma w\,dz)$, and ignore the inductance of the rings.

(a) Show that the drag force on the magnet, deep inside the tube, is $-kv$ with

$$k = \frac{45}{1024}\,\frac{\sigma w \mu_0^2 m^2}{a^4}.$$

You will need $\int_{-\infty}^{\infty} u^2\,du/(a^2+u^2)^5 = 5\pi/128a^7$.

(b) Find the terminal speed, and the time the magnet takes to fall 1.0 m at that speed.

(c) The magnet is dropped from rest 10 cm above the top of the tube. Use the method of Section 15.5.1, with rings 0.5 mm high, to find how far into the tube it goes before its speed is within 10% of the terminal speed.

Exercises for 15.6 Summary

P15.15 True or false? Explain your reasoning.

(a) The rule $\mathscr{E} = -d\Phi_B/dt$ is a consequence of changing electric fields creating changing magnetic fields.

(b) A magnetic field points along $+\hat{\mathbf{z}}$ and is increasing. The electric field it induces circles the z axis in the $+\hat{\boldsymbol{\phi}}$ direction.

(c) The current in a long solenoid is increasing. The curl of the induced electric field is nonzero everywhere.

(d) Outside a long solenoid whose current is changing, there is no electric field, because there is no magnetic field there.

(e) When the fields change in time, $\mathbf{E} = -\nabla V$ still holds if V is defined correctly.

P15.16 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as a magnet falling through a coil or a copper tube, the eddy currents in a spinning metal disk, the induced electric field of a finite solenoid whose current changes, the EMF of a loop that rotates in the nonuniform field of a bar magnet, the eddy-current losses in a laminated transformer core, the potentials of the same fields in two different gauges, or the orbit of an electron in a betatron.

Chapter 16

The Ampère–Maxwell Law

Chapter 15 gave the curl of the electric field the term it needs when the magnetic field changes. This chapter does the same for the curl of the magnetic field. A changing electric field gives the magnetic field a curl, which is Maxwell's displacement current, and Ampère's law becomes the Ampère–Maxwell law. With both effects the four equations of Chapter 2 are complete, the two fields stop being independent of each other, and the equations predict waves that travel at the speed of light.

The chapter begins with the contradiction that Ampère's law runs into when a capacitor is charged, and with Maxwell's term, which resolves it. It then develops the displacement-current method for finding the magnetic field that curls around a changing electric field, the counterpart of the Faraday-current method of Chapter 15, and a section on how the displacement current produces magnetic fields. The last sections ask when the new effects can safely be ignored and show how the two of them together predict electromagnetic waves.

16.1 Displacement current and the Ampère–Maxwell law

Faraday's law tells us that the curl of **E** is not zero when **B** changes in time. Ampère's law, the equation for the curl of **B**, must also be modified when the fields change, but for a different reason. In the form (11.4), it was derived from the Biot–Savart law for steady currents, and when the currents are not steady it is inconsistent with charge conservation.

16.1.1 A contradiction: the charging capacitor

Consider a capacitor that is being charged through a wire that carries a current I, and an Amperian loop C that circles the wire a short distance from the capacitor (Figure 16.1). The integral form of Ampère's law, (11.5), says that $\oint_C \mathbf{B}\cdot d\boldsymbol{\ell} = \mu_0 I_{\text{enc}}$, where I_{enc} is the current through any surface bounded by C. For the flat surface $\mathscr{S}_1$, which the wire pierces, $I_{\text{enc}} = I$. But the surface $\mathscr{S}_2$, bounded by the same loop, bulges out to pass between the plates, and no charge crosses it: the charges

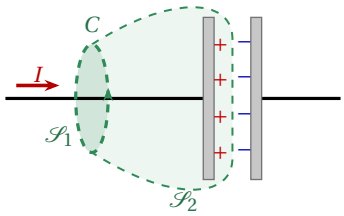


Figure 16.1 A capacitor being charged by a current I. The Amperian loop C circles the wire. The flat surface $\mathscr{S}_1$ bounded by C is pierced by the wire; the bulging surface $\mathscr{S}_2$, bounded by the same loop, passes between the plates, where no charge flows. Animation of the capacitor charging, with the displacement current and the magnetic field: outside the edge of the plates, **B** is the same around the gap as around the wires, and it vanishes when the current stops, while **E** remains.

that the wire carries stop at the plate. For $\mathscr{S}_2$, $I_{\text{enc}} = 0$. Ampère's law gives two different answers for the circulation around the same loop: $\mu_0 I$ and zero.

For steady currents, constant in time and not stopping or starting in space, the contradiction never arises. The surfaces $\mathscr{S}_1$ and $\mathscr{S}_2$ together form a closed surface, and when the currents are steady, as much current flows out of any closed surface as flows in (Section 10.4.1), so the currents through any two surfaces bounded by the same loop are equal. Here current flows into the closed surface through $\mathscr{S}_1$ and none flows out through $\mathscr{S}_2$, and the charge inside, on the plate, grows.

The differential form shows the same problem. The divergence of a curl is always zero, (1.57), so taking the divergence of both sides of $\nabla \times \mathbf{B} = \mu_0 \mathbf{J}$ gives

$$0 = \mu_0 \nabla \cdot \mathbf{J}.$$

That is the condition (10.13) for steady currents. But the continuity equation, (10.12), says that $\nabla \cdot \mathbf{J} = -\partial\rho/\partial t$, and wherever charge piles up, as it does on the plate of the capacitor, $\partial\rho/\partial t$ is *not* zero. As it stands, Ampère's law is incompatible with the conservation of charge.

16.1.2 Maxwell's term

What is needed is a quantity whose divergence is zero even where charge is piling up, and which reduces to $\mathbf{J}$ when the currents are steady. Gauss's law, $\nabla \cdot \mathbf{E} = \rho/\varepsilon_0$, (4.2), holds whether or not the fields change, so the charge density in the continuity equation can be written in terms of $\mathbf{E}$:

$$\begin{aligned}
\nabla \cdot \mathbf{J} &= -\frac{\partial \rho}{\partial t} \\
\nabla \cdot \mathbf{J} &= -\frac{\partial (\varepsilon_0 \nabla \cdot \mathbf{E})}{\partial t} \\
\nabla \cdot \mathbf{J} &= -\nabla \cdot \left(\varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right) \\
\nabla \cdot \left(\mathbf{J} + \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right) &= 0.
\end{aligned}$$

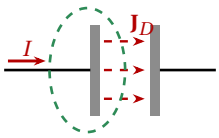


Figure 16.2 A closed surface (dashed) around one plate of a charging capacitor. Current flows in through the wire and none flows out, so charge piles up on the plate; but the displacement current (dashed arrows) leaves mostly through the gap, and $\mathbf{J} + \varepsilon_0 \partial\mathbf{E}/\partial t$ has as much flux out of the surface as in.

So the combination $\mathbf{J} + \varepsilon_0 \partial\mathbf{E}/\partial t$ always has zero divergence, and when nothing changes in time it is simply $\mathbf{J}$. Replacing $\mathbf{J}$ by it in Ampère's law gives

$$\boxed{\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \mu_0 \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t}}, \tag{16.1}$$

the fourth of Maxwell's equations, (2.4). This is the **Ampère–Maxwell law**, and $\mu_0 \varepsilon_0 \partial\mathbf{E}/\partial t$ is **Maxwell's term**, or Maxwell's correction to Ampère's law. With it, the divergence of the law gives the continuity equation (Exercise P2.1; Figure 16.2) instead of a contradiction of it. Section 20.1 takes up the continuity equation in full. Many people call the corrected equation Ampère's law as well. In this book Ampère's law means the static version, (11.4), and the corrected equation is the Ampère–Maxwell law.

The argument shows that *something* has to be added to Ampère's law and that this term works. It does not show that nothing else would work, since any added term with zero divergence would leave the continuity equation as it is. Maxwell's term is established by experiment: with it, Maxwell's equations predict electromagnetic waves that travel at the speed of light (Section 16.5), and Heinrich Hertz produced and detected such waves in 1886–88. Nor is this argument, now ubiquitously used, actually how Maxwell found the term. Maxwell did know that Ampère's law in this form holds only for closed currents — he pointed it out in 1855–56, and again in his *Treatise* of 1873[1] — but as Section 2.4 describes, the term itself came out of his mechanical model of the field, in 1861–62.

16.1.3 The displacement current

Compare the two curl equations:

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \qquad \nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \mu_0 \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t}.$$

A changing **B** creates an electric field with curl, and a changing **E** creates a magnetic field with curl. More than that, mathematically speaking, the changing **B** and **E** fields create **E** and **B** fields in exactly the same way that a current density **J** creates a magnetic field. In Section 15.2.4 we called $-\partial \mathbf{B}/\partial t$ the Faraday current density, even though it is not a current density at all. On the other hand, this new term $\varepsilon_0 \partial \mathbf{E}/\partial t$ *does* have units of current density. It is the empty-space form of the **displacement current density** $\mathbf{J}_D$, or more colloquially just the **displacement current**:

$$\mathbf{J}_D = \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t}. \qquad \text{(empty space)} \quad (16.2)$$

The general definition uses the **D** of Section 9.5:[2]

$$\mathbf{J}_D = \frac{\partial \mathbf{D}}{\partial t}. \qquad \text{(general definition)} \quad (16.3)$$

In empty space, where $\mathbf{D} = \varepsilon_0 \mathbf{E}$, this is $\varepsilon_0 \partial \mathbf{E}/\partial t$, which does not involve any real currents, and the Ampère–Maxwell law reads $\nabla \times \mathbf{B} = \mu_0 (\mathbf{J} + \mathbf{J}_D)$. Inside matter **D** also contains the polarization, and Section 16.1.4 talks more about that case. The integral form follows from the curl theorem, as Ampère's law did:

$$\oint \mathbf{B} \cdot d\boldsymbol{\ell} = \mu_0 \left(I_{\text{enc}} + I_{D,\text{enc}} \right), \qquad I_{D,\text{enc}} = \int_S \mathbf{J}_D \cdot d\mathbf{a} = \varepsilon_0 \frac{d\Phi_E}{dt}, \qquad (16.4)$$

where $\Phi_E = \int_S \mathbf{E} \cdot d\mathbf{a}$ is the flux of **E** through the surface S bounded by the loop.

[1] J. C. Maxwell, *A Treatise on Electricity and Magnetism* (Clarendon Press, Oxford, 1873), Vol. 2, Art. 607.

[2] Some books instead define the displacement current density as $\varepsilon_0 \partial \mathbf{E}/\partial t$, even inside matter. The two definitions agree in empty space; inside matter they differ by the polarization current of Section 16.1.4.

The displacement current settles the capacitor's contradiction. Between the plates, if the plates have area A and charges $\pm Q$, and if we ignore the fringing at the edges, the field is $E = Q/\varepsilon_0 A$, pointing from the positive plate toward the negative one (Example 4.6). The displacement current through $\mathscr{S}_2$ is then

$$\begin{aligned} I_{D,\text{enc}} &= \varepsilon_0 \frac{d}{dt}(EA) \\ &= \frac{dQ}{dt} = I, \end{aligned}$$

exactly the current in the wire. Through $\mathscr{S}_1$ there is a current I and no displacement current, and through $\mathscr{S}_2$ there is a displacement current I and no current, so with the fringing ignored both surfaces give $\oint \mathbf{B}\cdot d\boldsymbol{\ell} = \mu_0 I$. (The fringing field crosses $\mathscr{S}_1$ as well; Section 16.3.2 takes up what the fringing changes.) The current that stops at the plate continues across the gap as displacement current, and the sum $\mathbf{J} + \mathbf{J}_D$ flows in closed loops, as a steady current does. The magnetic field between the plates of a charging capacitor is weak, but it has been measured: it grows in proportion to the distance from the axis, as it must if the displacement current is spread uniformly between the plates (as in Example 16.1), with a slope that agrees with the prediction to within 5%.[3]

The two curl equations differ in one important way: the Ampère–Maxwell law has no minus sign. The magnetic field curls around an increasing $\mathbf{E}$ in the same sense that it curls around a current pointing along $\mathbf{E}$ (Figure 16.3), while the electric field curls around an increasing $\mathbf{B}$ in the opposite sense. So there is nothing like Lenz's law for the induced magnetic field. The induced $\mathbf{B}$ does not oppose the change in $\mathbf{E}$, and arguments that borrow the minus sign from Faraday's law give the wrong direction. The difference in sign is essential: it is what allows the two equations together to describe waves (Section 16.5).

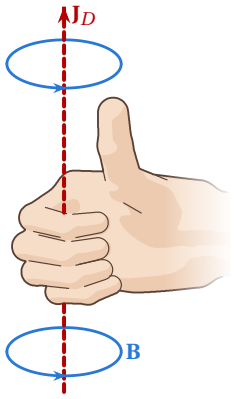


Figure 16.3 The right-hand rule for the displacement current: with the thumb of your right hand along $\mathbf{J}_D = \varepsilon_0 \partial\mathbf{E}/\partial t$ (dashed), your fingers curl around in the direction of the magnetic field $\mathbf{B}$.

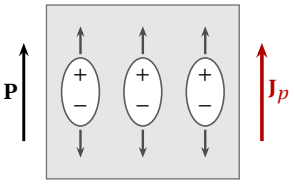


Figure 16.4 When the polarization $\mathbf{P}$ of a dielectric grows, the positive end of each molecule moves along $\mathbf{P}$ and the negative end moves the other way (small arrows). Both motions carry current along the change in $\mathbf{P}$: the polarization current density $\mathbf{J}_p = \partial\mathbf{P}/\partial t$.

16.1.4 In materials

Inside matter, the current density $\mathbf{J}$ in the Ampère–Maxwell law is the total current density. When the fields change, it has a third part besides the free current $\mathbf{J}_f$ and the bound current $\nabla\times\mathbf{M}$ of (13.13): when the polarization of a dielectric changes, its bound charges move, and their motion is the **polarization current density** $\mathbf{J}_p = \partial\mathbf{P}/\partial t$ (Figure 16.4). There are thus three real currents that must be accounted for in the Ampère–Maxwell law — free, bound, and polarization — in addition to the new $\varepsilon_0\partial\mathbf{E}/\partial t$ term. It can be written several equivalent ways:

$$\nabla\times\mathbf{B} = \mu_0\left(\mathbf{J}_f + \nabla\times\mathbf{M} + \frac{\partial\mathbf{P}}{\partial t}\right) + \mu_0\varepsilon_0\frac{\partial\mathbf{E}}{\partial t},$$

or, using $\mathbf{D} = \varepsilon_0\mathbf{E} + \mathbf{P}$,

$$\nabla\times\mathbf{B} = \mu_0\left(\mathbf{J}_f + \nabla\times\mathbf{M}\right) + \mu_0\frac{\partial\mathbf{D}}{\partial t},$$

[3] "Measuring Maxwell's displacement current inside a capacitor," D. F. Bartlett and T. R. Corle, Phys. Rev. Lett. 55, 59–62 (1985). https://doi.org/10.1103/PhysRevLett.55.59.

or, using $\mathbf{H} = \mathbf{B}/\mu_0 - \mathbf{M}$,

$$\nabla \times \mathbf{H} = \mathbf{J}_f + \frac{\partial \mathbf{D}}{\partial t}. \tag{16.5}$$

The latter form is typically what is used for materials. By contrast, Faraday's law is the same in matter as in vacuum, because it contains no charges or currents. Section 19.1 derives the polarization current equation, and discusses the new boundary conditions that go with the Ampère–Maxwell law in matter.

16.2 Solving for B when given a changing E: displacement-current method (high symmetry)

The integral form of the Ampère–Maxwell law, (16.4), is Ampère's law with the displacement current added to the current. So this is a new type of problem, but not a new method. Given a changing electric field, find the displacement current density, treat it like a regular current density, and use it to figure out **B**. Whenever the current and the displacement current together have one of the symmetries of Section 11.2.2, the magnetic field follows as in Section 11.2:

1. Find **E** and the displacement current density, $\mathbf{J}_D = \varepsilon_0 \partial \mathbf{E}/\partial t$ in empty space, and sketch it. It points along the change in **E**, with no minus sign.
2. If there are real currents as well, add them: **B** curls around $\mathbf{J} + \mathbf{J}_D$. Use the symmetry of the total to determine the direction of **B** and which coordinates its magnitude can depend on.
3. Choose an Amperian loop that matches the symmetry and draw it (Colton's Third Commandment), with the direction in which you will go around it.
4. Calculate $I_{\text{enc}} + I_{D,\text{enc}}$, with $I_{D,\text{enc}} = \int \mathbf{J}_D \cdot d\mathbf{a}$ (which also equals $\varepsilon_0 \, d\Phi_E/dt$).
5. Solve $\oint \mathbf{B} \cdot d\boldsymbol{\ell} = \mu_0 (I_{\text{enc}} + I_{D,\text{enc}})$ for B, and give **B** its direction.

In most problems only one of the two currents passes through the loop: between the plates of a capacitor there is only displacement current, and inside a metal wire the displacement current is negligible compared with the current (Section 16.4). Although it has the units of a current density, the displacement current density is (in a vacuum) *not* a current at all: no charge moves. But **B** curls around it in exactly the same way as around a regular current density **J**, and the right-hand rule gives the direction: point your right thumb along $\mathbf{J}_D$, and your fingers curl around in the direction of **B** (Figure 16.3).

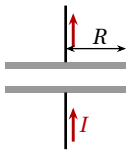


Figure 16.5 The capacitor of Example 16.1, edge on.

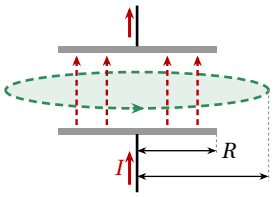
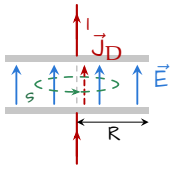


Figure 16.6 The capacitor of Example 16.1 from the side, with an Amperian loop of radius $s > R$ in the gap and the flat disk it bounds (shaded). With the fringing ignored, the displacement current (dashed arrows) crosses only the part of the disk inside the edge of the plates, and the loop encloses all of it.

Example 16.1

A capacitor with circular plates of radius R, a small distance apart, is charged by a constant current I in wires that run along its axis to the centers of the plates. Ignoring the fringing of the electric field at the edges, find the magnetic field between the plates at a distance s from the axis, both inside the edge of the plates ($s < R$) and outside it ($s > R$). How strong is the field at the edge of plates 5.0 cm in radius when $I = 0.50$ A?

Solution:



We begin with a sketch of what is given and what is asked: the capacitor from the side, with the current flowing up the wire into the bottom plate and out of the top one. The bottom plate gains positive charge, so $\mathbf{E}$ between the plates points up and grows, and $\mathbf{J}_D$ points up too. The loops are circles of radius s between the plates, centered on the axis, which we traverse counterclockwise as seen from above.

The charge Q on the bottom plate increases at the rate $dQ/dt = I$. Between the plates the field is $\mathbf{E} = (Q/\varepsilon_0 \pi R^2)\,\hat{\mathbf{z}}$, and with no fringing it is zero outside the edge, so

$$\mathbf{J}_D = \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} = \begin{cases} \dfrac{I}{\pi R^2}\,\hat{\mathbf{z}}, & s < R, \\ 0, & s > R. \end{cases}$$

The displacement current is spread uniformly over the gap, points the same way as the current in the wire, and carries the same total, I. The whole pattern, current and displacement current together, is symmetric about the axis, and every part of it lies in a plane through the axis: up the wire, outward across the bottom plate, up across the gap, and inward across the top plate. As for a straight wire, the magnetic field circles the axis. (The circulation around a circle centered on the axis would give the ϕ component even without this argument, since by symmetry it is the same all the way around.)

Going counterclockwise around a circle of radius s in the gap, $\oint \mathbf{B}\cdot d\boldsymbol{\ell} = 2\pi s B_\phi$, and the flat disk bounded by the circle has no current through it, only displacement current. Inside the edge it catches the fraction s^2/R^2 of it, and outside the edge all of it:

$$2\pi s B_\phi = \mu_0 I \frac{s^2}{R^2} \quad (s < R), \qquad 2\pi s B_\phi = \mu_0 I \quad (s > R).$$

For $s > R$ the changing field crosses only the part of the disk with $s < R$, not the whole disk, but the law counts the displacement current wherever it crosses the surface, and the loop encloses all of it (Figure 16.6). So

$$\mathbf{B} = \begin{cases} \dfrac{\mu_0 I s}{2\pi R^2}\,\hat{\boldsymbol{\phi}}, & s < R, \\ \dfrac{\mu_0 I}{2\pi s}\,\hat{\boldsymbol{\phi}}, & s > R, \end{cases} \tag{16.6}$$

at every height in the gap. (Why is the field the same at every height in the gap? Answer in footnote.[4]) This is the field of the thick wire of Example 11.1, with the displacement current in the place of the current: it grows in proportion to s inside the edge (Figure 16.7), and outside, it is the same as the field around the wire leading to the capacitor. With the fringing ignored, the magnetic field outside the edge cannot tell that the wire has a gap in it. (Section 16.3 looks at what the fringing changes, and at where this field comes from.)

At the edge of the given plates,

$$\begin{aligned} B &= \frac{\mu_0 I}{2\pi R} \\ &= \frac{(4\pi \times 10^{-7}\,\mathrm{T\cdot m/A})(0.50\,\mathrm{A})}{2\pi(0.050\,\mathrm{m})} \\ &= 2.0 \times 10^{-6}\,\mathrm{T}, \end{aligned}$$

less than a tenth of Earth's field, which is one reason this field is hard to measure.

Reality check: The units are those of $\mu_0 I$/length, teslas; the field is continuous at $s = R$; and the direction is the one the right-hand rule gives for a current up the axis, as it should be, since the Ampère-Maxwell law has no minus sign.

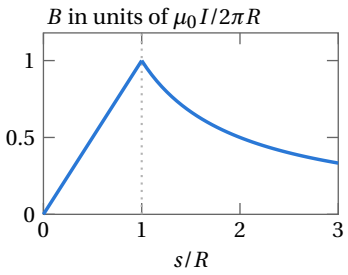


Figure 16.7 The magnetic field between the plates of Example 16.1, against s/R: it grows as s inside the edge and falls as $1/s$ outside, like the field of a thick wire.

Tutorial 16.1 practices both methods, with fields that change in a way that is not uniform in time.

Tutorial 16.1 Time-dependent fields

Printable version

Background

This chapter and the previous one complete Maxwell's equations by fixing the curl equations to account for time-dependent fields:

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \qquad \nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \mu_0 \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t}.$$

This says that a changing **B** field will create an electric field with curl, and a changing **E** field will create a magnetic field with curl. More than that, mathematically speaking, the changing **B** and **E** fields will create **E** and **B** fields in exactly the same way that a current density **J** creates a magnetic field. In particular, the Amperian-loop technique we have used to find **B** from $\mu_0 \mathbf{J}$ can be used to find **E** from the Faraday current density $\mathbf{J}_{\mathrm{Far}} = -\partial \mathbf{B}/\partial t$, and to find **B** from μ_0 times the displacement current density $\mathbf{J}_D = \varepsilon_0 \partial \mathbf{E}/\partial t$.

Focusing on the second equation for a moment: just as in high symmetry situations we can draw an Amperian loop and use Ampère's law in integral form to find **B** from **J**,

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} \quad \leftrightarrow \quad \oint \mathbf{B} \cdot d\boldsymbol{\ell} = \mu_0 I_{\mathrm{enc}},$$

[4] With the fringing ignored, $\mathbf{J}_D$ is the same at every height between the plates, so a circle of radius s encloses the same displacement current wherever it is in the gap.

we can pretend that $\varepsilon_0 \partial \mathbf{E}/\partial t$ is a current density, and find out how much of that type of "current" is enclosed by our Amperian loop:

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J}_D \quad \leftrightarrow \quad \oint \mathbf{B} \cdot d\boldsymbol{\ell} = \mu_0 I_{D,\text{enc}}.$$

If you have both regular $\mathbf{J}$ and also $\mathbf{J}_D$, the loop encloses both, and $\oint \mathbf{B} \cdot d\boldsymbol{\ell} = \mu_0 (I_{\text{enc}} + I_{D,\text{enc}})$. Similarly, Faraday's law becomes

$$\nabla \times \mathbf{E} = \mathbf{J}_{\text{Far}} \quad \leftrightarrow \quad \oint \mathbf{E} \cdot d\boldsymbol{\ell} = I_{\text{Far,enc}}.$$

The goal of this tutorial is to use the last two equations to calculate induced fields for two situations.

Part 1 — A changing E field creates a B field

The charge of a circular parallel-plate capacitor is increasing according to $q = q_0\left(1 - e^{-t/\tau}\right)$, where q_0 and τ are positive constants. The top plate has the given charge q; the bottom plate has charge $-q$. The plates have radius R and are separated in the z direction by a distance d, where $R \gg d$, so you can neglect edge effects. What magnetic field is produced in between the capacitor plates?

(a) Draw a picture or diagram to help you visualize what is going on.

(b) Why is there a magnetic field? Describe in words.

(c) What will the displacement current be (magnitude, direction)? Make a rough sketch of the magnetic field that curls around it on your picture above.

(d) Now, getting ready to solve the problem, describe what you have to work with in terms of givens, knowns, unknowns, target variables, and so forth.

(e) Come up with a plan to use the integral form of the Ampère–Maxwell law, with the displacement current, to solve this problem, similar to how you have used Ampère's law in the past, and carry out that plan.

Part 2 — A changing B field creates an E field

The current in a solenoid (n turns per length) is increasing according to $I = I_0\left(1 - e^{-t/\tau}\right)$, where I_0 and τ are positive constants. The solenoid's length is much longer than its radius, so you can neglect edge effects. What electric field is produced inside the solenoid?

(a) Do this by drawing a picture, finding the Faraday current, drawing an appropriate Amperian loop, and using Faraday's law in the form $\oint \mathbf{E} \cdot d\boldsymbol{\ell} = I_{\text{Far,enc}}$ to solve for $\mathbf{E}$.

Part 3 — Wrap-up

(a) Are there any ways you can think of to consider whether the answers you obtained in 1(e) and 2(a) are reasonable? (Limiting cases, units, comparison to other problems, and so forth.)

16.3 How the displacement current produces magnetic fields

Section 16.2 treated the **displacement current** as if it were a current, and Example 16.1 used it to find the magnetic field between the plates of a charging capacitor. Inside the edge of the plates that field grows in proportion to s, just as the field inside a thick wire does, and it is natural to conclude that the displacement current between the plates produces the field there, as the current in a wire produces the field around the wire. That conclusion is mistaken, and the mistake is a common one.[5] The statement that a changing electric field produces a magnetic field in the same way that a current does can be read in two ways. Both readings are correct, but they say different things.

In the first reading, the displacement current is part of the *curl* of $\mathbf{B}$: $\nabla \times \mathbf{B} = \mu_0(\mathbf{J} + \mathbf{J}_D)$ at every point, and $\oint \mathbf{B} \cdot d\boldsymbol{\ell} = \mu_0(I_{\text{enc}} + I_{D,\text{enc}})$ around every loop. In this sense the displacement current is indispensable. Between the plates of a charging capacitor, where there is no current, it is the whole curl of $\mathbf{B}$, and an Amperian loop there encloses nothing else. Every Amperian-loop calculation in this chapter uses the Ampère–Maxwell law in this sense.

In the second reading, the displacement current is a *source* of $\mathbf{B}$, in the way that a current is the source of the field in the Biot–Savart law. In this sense, too, the displacement current enters exactly as a current does. By the Helmholtz theorem (Section 2.3), a field that goes to zero far away is fixed by its divergence and its curl, so $\nabla \cdot \mathbf{B} = 0$ and $\nabla \times \mathbf{B} = \mu_0(\mathbf{J} + \mathbf{J}_D)$ give, at every instant,

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int \frac{(\mathbf{J} + \mathbf{J}_D) \times \hat{\boldsymbol{\imath}}}{\imath^2} \, d\tau', \tag{16.7}$$

the Biot–Savart law with $\mathbf{J} + \mathbf{J}_D$ in the place of $\mathbf{J}$, where the integral runs over all of space. But the displacement current's share of this integral can be zero, and for the charging capacitor it is.

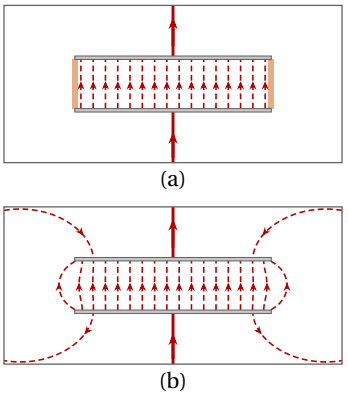


Figure 16.8 The displacement current (dashed) of a capacitor charged by a constant current, in a cross section through its axis. (a) With the fringing ignored, as in Example 16.1, it is uniform between the plates and stops at their edge, where it has a curl (orange). (b) Computed for conducting plates whose separation is half their radius: some of it leaves the edges and outer faces of the plates and loops around outside, and it has no curl anywhere. The lines are spaced evenly where the field is uniform; elsewhere the spacing is rough.

16.3.1 The displacement current of a charging capacitor

In Example 16.1 the current is constant, so the magnetic field does not change in time, and by Faraday's law the electric field has no curl. The electric field can therefore be written as $\mathbf{E} = -\nabla V$, as in electrostatics. Then we have

$$\mathbf{J}_D = \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} = -\nabla \left(\varepsilon_0 \frac{\partial V}{\partial t} \right),$$

and as you can see, the displacement current is a gradient field (Figure 16.8b). The Biot–Savart integral of a gradient field over all space is zero. To see why, consider an arbitrary vector field $\mathbf{F}$ which falls off faster than $1/r$ far away. Exercise P16.6

[5] See "Displacement currents and magnetic fields," A. P. French and J. R. Tessman, Am. J. Phys. 31, 201–204 (1963), https://doi.org/10.1119/1.1969359; "Calculation of magnetic and electric fields from displacement currents," R. M. Whitmer, Am. J. Phys. 33, 481–484 (1965), https://doi.org/10.1119/1.1971712; "Maxwell's displacement current revisited," J. D. Jackson, Eur. J. Phys. 20, 495–499

shows that $\int \mathbf{F} \times (\hat{\boldsymbol{\imath}}/\imath^2)\, d\tau' = \int (\nabla' \times \mathbf{F})/\imath\, d\tau'$, and hence, for the Biot–Savart integral,

$$\frac{\mu_0}{4\pi} \int \frac{\mathbf{F} \times \hat{\boldsymbol{\imath}}}{\imath^2}\, d\tau' = \frac{\mu_0}{4\pi} \int \frac{\nabla' \times \mathbf{F}}{\imath}\, d\tau', \tag{16.8}$$

where ∇' takes derivatives with respect to the source point $\mathbf{r}'$. With words: the Biot–Savart integral sees only the curl of what is put into it. A gradient field has no curl anywhere, so it contributes nothing. The displacement current of the charging capacitor therefore adds nothing to the magnetic field, which, between the plates as well as outside them, is the Biot–Savart field of the real currents alone: the current in the wires, and the currents that spread out over the plates to carry the charge onto them.

How, then, did Example 16.1 find the field between the plates from the displacement current alone? It used the Ampère–Maxwell law in the first sense, counting the displacement current through a disk bounded by the loop, and that is legitimate (Section 16.3.2 returns to it). The mistaken conclusion comes from applying the second reading to the example's displacement current. With the fringing ignored, $\mathbf{J}_D$ is uniform between the plates and zero outside their edge (Figure 16.8a). That is not a gradient field: it stops abruptly at the edge of the plates, so it has a curl there, on the cylinder $s = R$ between the plates (Exercise P1.21), and by (16.8) that curl gives it a Biot–Savart integral that is not zero. The real displacement current does not stop at the edge. It follows the fringing field (Figure 16.8b): some of it leaves the edges and the outer faces of the plates and loops around outside before it reaches the other plate, and it has no curl anywhere. Its Biot–Savart integral is zero, the part outside the gap exactly canceling the part between the plates.

The cancellation can also be seen without any mathematics. We can model the increasing charge on the plates as a collection of point charges which are growing in magnitude. Each increasing charge contributes to the overall displacement current, with amount $\varepsilon_0 \partial \mathbf{E}/\partial t = (dq/dt)\,\hat{\mathbf{r}}/4\pi r^2$. The displacement current for a single increasing charge (positive dq/dt) thus "streams" away from the charge, going straight out from it in every direction (Figure 16.9). Consider an arbitrary point P. The displacement current contributions to the field at that point can be summed up in pairs of streams placed symmetrically above and below the line connecting the charge to P. The two streaming elements contribute equal and opposite amounts to the Biot–Savart integral at P: the one above the symmetry line creating a field into the page and the one below the line creating a field out of the page, as per the figure. So the field created by the displacement current from each charge adds up to zero, as does the field created by all the charges together.

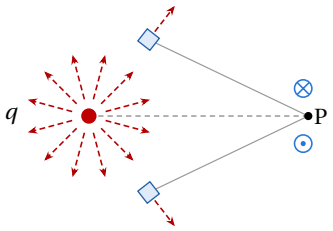


Figure 16.9 The displacement current (dashed) of a growing point charge q points straight out from it. Two elements placed symmetrically about the line from q to a field point P contribute equal and opposite amounts to the Biot–Savart integral at P, one into the page and one out of it.

(1999), https://doi.org/10.1088/0143-0807/20/6/315; and "Untold secrets of the slowly charging capacitor," J. A. Milsom, Am. J. Phys. 88, 194–199 (2020), https://doi.org/10.1119/10.0000635, which computes the field of the currents in the wire and on the plates point by point.

16.3.2 The real currents make the field

For the charging capacitor situation, then, the correct analysis says that the Biot–Savart law with only the *real* currents gives the field everywhere, including between the plates. This can be calculated in just a few steps. First imagine the wire unbroken, with the current I running straight through the gap; by itself it gives the field $\mu_0 I/2\pi s$ everywhere around it, between the plates as well. For plates which are close together, that is a good approximation to the field of the actual wire, which is broken at the gap. However, there are also currents on the plates. The current that arrives at the center of the bottom plate spreads outward over it, depositing charge as it goes, and the same current flows inward over the top plate to leave it. If the charge piles up uniformly over the plates, as Example 16.1 assumed, then at a distance s from the axis the current still spreading outward is the part of I not yet deposited, $I(1-s^2/R^2)$, and the surface current density on each plate is

$$K(s) = \frac{I}{2\pi s}\left(1-\frac{s^2}{R^2}\right).$$

The two plates carry equal and opposite surface currents a small distance apart, so close to them they act like two large sheets with opposite currents (Section 11.2.4): the field between them is $\mu_0 K$, and outside them it nearly cancels. Between the plates the field of the two sheets points opposite to that of the wire, so

$$B = \frac{\mu_0 I}{2\pi s} - \mu_0 K(s) = \frac{\mu_0 I s}{2\pi R^2},$$

the answer of Example 16.1 inside the edge, with no displacement current anywhere in the calculation, only the real currents of Figure 16.10. Outside the edge there are no plate currents, and the field is that of the wire.

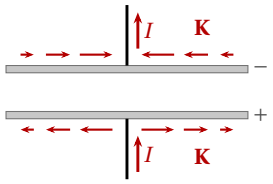


Figure 16.10 The currents that charge the capacitor of Example 16.1: up the wires, outward over the positive plate and inward over the negative one, depositing charge as they go. The plate currents are weaker farther out, where less charge remains to be deposited. Animation of their Biot–Savart field added up, the wires first and then the plates: the sum is the whole magnetic field, between the plates as well.

Real plates are conductors, and that changes the numbers. The fringing field holds part of the charge on the outer faces of the plates and near their edges, so less of the current is deposited between the plates, and the field there is smaller. The Ampère–Maxwell law gives the correction exactly for a circle well inside the edge, where the field between conducting plates is uniform. If ΔV is the voltage between the plates, then $E = \Delta V/d$ there, and with $Q = C\,\Delta V$ and $dQ/dt = I$,

$$I_{D,\text{enc}} = \varepsilon_0 \pi s^2 \frac{dE}{dt} = \frac{\varepsilon_0 \pi s^2}{d}\frac{I}{C} = \frac{C_0}{C} I \frac{s^2}{R^2},$$

where $C_0 = \varepsilon_0 \pi R^2/d$ is the capacitance with the fringing ignored and C is the real one. So

$$B = \frac{C_0}{C}\frac{\mu_0 I s}{2\pi R^2}, \tag{16.9}$$

smaller than (16.6) by the factor C_0/C at every s well inside the edge. The real capacitance is larger than C_0 because of the charge that the fringing field holds outside the gap. For thin conducting plates, $C/C_0 = 1.044$ when $R = 50d$ and

1.145 when $R = 12.5d$,[6] so the field between the plates is smaller than that of Example 16.1 by about 4% and 13%. For the plates of that example, 5.0 cm in radius, with a gap of 1.0 mm, the field at the edge is about 1.9 μT rather than 2.0 μT.

Both calculations give the same field, as they must. The Ampère–Maxwell law is usually the easier one, because a loop between the plates encloses only displacement current and needs only its flux. But that displacement current is what the loop encloses, not where the field comes from: the field is the Biot–Savart field of the real currents, in the wires and on the plates. Ignoring the fringing is a good approximation for plates close together. What it leaves out is the displacement current outside the gap, and with it the reason that the displacement current adds nothing to the Biot–Savart integral.

16.3.3 When the displacement current does contribute

The argument of Section 16.3.1 used a single fact: at each instant the changing electric field was the field that the charges would produce by Coulomb's law, so its displacement current was a gradient field. Two kinds of displacement current are not gradient fields, and they do contribute to the Biot–Savart integral (16.7).

The first is in matter. There the displacement current (16.3) contains the polarization current, $\mathbf{J}_D = \varepsilon_0 \partial \mathbf{E}/\partial t + \mathbf{J}_p$ (Section 16.1.4). The polarization current is the motion of real charges, the bound charges of the dielectric, and it contributes to the Biot–Savart integral as any current does (Example 16.2).

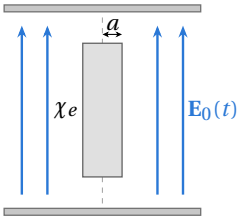
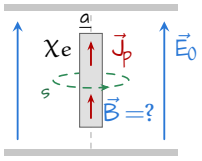


Figure 16.11 The capacitor and dielectric rod of Example 16.2, edge on.

Example 16.2

A long rod of radius a, made of a linear dielectric with susceptibility χ_e, lies along the axis of a large parallel-plate capacitor, far from both plates. The capacitor is being charged so that its field, $E_0(t)\,\hat{\mathbf{z}}$, increases at a constant rate. Near the middle of the rod, find the polarization current and the magnetic field that it produces, inside and outside the rod. Compare that field, at the surface of the rod, with the field that the capacitor itself produces there.

Solution:



First, a sketch of what is given and what is asked: the rod on the axis of the capacitor, edge on, with the field $\mathbf{E}_0$ growing upward. Inside the rod the polarization grows as well, so its bound charges move: the polarization current $\mathbf{J}_p$, up the rod. The loop is a circle of radius s around the middle of the rod.

Near the middle of a long rod, far from its ends, the field inside is the capacitor's field, $E_0\,\hat{\mathbf{z}}$: the field lines run along the rod, $E_\parallel$ is continuous across its side (Section 5.2), and

[6] Computed for two thin conducting disks by dividing each into narrow rings and solving for the charges on the rings that make each disk an equipotential.

the bound charges at the ends of the rod are too far away to matter. So inside the rod

$$\mathbf{P} = \varepsilon_0 \chi_e E_0 \hat{\mathbf{z}}, \qquad \mathbf{J}_p = \frac{\partial \mathbf{P}}{\partial t} = \varepsilon_0 \chi_e \frac{dE_0}{dt} \hat{\mathbf{z}},$$

uniform over the cross section, a total current

$$I_p = \pi a^2 \varepsilon_0 \chi_e \frac{dE_0}{dt}$$

up the rod. This is a current of real charges, and near the middle of the rod it is like the current in a long straight wire of radius a. Its field is that of Example 11.1:

$$\mathbf{B}_{\text{rod}} = \begin{cases} \dfrac{\mu_0 I_p s}{2\pi a^2} \hat{\boldsymbol{\phi}}, & s < a, \\ \dfrac{\mu_0 I_p}{2\pi s} \hat{\boldsymbol{\phi}}, & s > a. \end{cases}$$

The rest of the displacement current, $\varepsilon_0 \partial \mathbf{E}/\partial t$, belongs to the field that the charges on the plates and at the ends of the rod would produce by Coulomb's law, so it adds nothing to the Biot-Savart integral (Section 16.3.1). The polarization current alone makes the rod's field.

The capacitor's own field near its axis, made by the currents in its wires and plates, follows from the Ampère-Maxwell law as in Example 16.1: a circle of radius s encloses the displacement current $\varepsilon_0 \pi s^2 \, dE_0/dt$, so $B_{\text{cap}} = \mu_0 \varepsilon_0 s (dE_0/dt)/2$. At the surface of the rod,

$$\frac{B_{\text{rod}}}{B_{\text{cap}}} = \frac{\mu_0 I_p / 2\pi a}{\mu_0 \varepsilon_0 a (dE_0/dt)/2} = \chi_e.$$

A rod with a large susceptibility adds much more to the field beside it than the capacitor provides.

Reality check: The two fields circle the axis in the same sense, counterclockwise as seen from above, because the polarization current points along the change in $\mathbf{E}$, as the displacement current between the plates does. The Ampère-Maxwell law agrees: a circle of radius $s > a$ near the middle of the rod encloses the displacement current $\varepsilon_0 \pi s^2 \, dE_0/dt + I_p$, the capacitor's part and the rod's (Section 16.1.4), so $2\pi s B = \mu_0(\varepsilon_0 \pi s^2 \, dE_0/dt + I_p)$, which is $B_{\text{cap}} + B_{\text{rod}}$. As $\chi_e \to 0$, the rod's field vanishes.

The second is the displacement current of an *induced* electric field. Any electric field is the sum of the field that the charges would produce by Coulomb's law and the field induced by the changing magnetic field, which curls around the Faraday current (Section 15.2.4). In the Coulomb gauge these are the two terms of $\mathbf{E} = -\nabla V - \partial \mathbf{A}/\partial t$, (15.11). Only the first is a gradient field. In the solenoid of Example 15.4, if the current does not change at a steady rate, the induced field changes in time, and its displacement current circles the axis like the current in the windings. It is not a gradient field, and it adds to the magnetic field. Such additions are the corrections that the next section estimates, and when the fields

change slowly they are very small. Far from all charges and currents, though, the electric field is entirely induced, and the displacement current is all that is left of the curl of **B**. For a pulse of light whose source has stopped, with no currents anywhere, (16.7) makes **B** at each instant the Biot–Savart integral of $\varepsilon_0 \partial \mathbf{E}/\partial t$ alone. This is how the two fields of an electromagnetic wave keep each other going (Section 16.5).

The rule, then, is this: the displacement current of the field that the charges would produce by Coulomb's law, where they are at that instant, never adds anything to the Biot–Savart integral, however quickly the charges change, because that field is a gradient field at every instant. The polarization current and the displacement current of an induced electric field do add to it. In empty space, then, the magnetic field of slowly changing currents is the Biot–Savart field of the currents at that instant, whether or not the currents are steady; the displacement current of the induced field gives corrections of the size that the next section estimates.

16.4 When is fast, when is slow

The examples of Sections 15.3 and 16.1–16.3 each used one of the two new effects and ignored the other. Example 15.4 found the electric field induced by a changing magnetic field. But that electric field changes in time too, unless the current changes at a steady rate, so it has a displacement current of its own; the displacement current produces a magnetic field, which also changes, and induces another electric field, and so on. Example 16.1 found the magnetic field of a charging capacitor from its displacement current, and if the current charging the capacitor had not been constant, that magnetic field would have changed and induced an electric field of its own. In each case we stopped after one step (Figure 16.12). When is that safe?

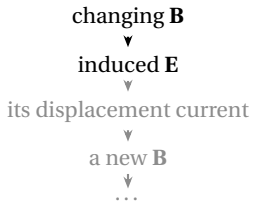


Figure 16.12 The chain that Example 15.4 cut after its first step: the induced **E** also changes, so it has a displacement current, which makes a new **B**, and so on (gray).

16.4.1 The size of the corrections

Suppose the fields change appreciably in a time $1/\omega$ (for fields that oscillate, ω is the angular frequency), and that the region we care about has a size L. In Faraday's law, the curl of **E** is of order E/L and the time derivative of **B** is of order ωB, so a magnetic field B induces an electric field of order

$$E \sim \omega L B.$$

In the Ampère–Maxwell law, an electric field E produces, through its displacement current, a magnetic field of order

$$B \sim \mu_0 \varepsilon_0 \, \omega L E = \frac{\omega L}{c^2} E,$$

where $c = 1/\sqrt{\mu_0 \varepsilon_0} = 3.00 \times 10^8 \,\mathrm{m/s}$, the combination of constants we met in Section 10.5. Start with a magnetic field B_0 produced by currents. It induces an

electric field $E_1 \sim \omega L B_0$, whose displacement current produces a correction to the magnetic field of order

$$B_2 \sim \frac{\omega L}{c^2} E_1 \sim \left(\frac{\omega L}{c}\right)^2 B_0.$$

Each round trip from **B** to **E** and back multiplies the correction by $(\omega L/c)^2$, and the same is true starting from an electric field. So the first-step fields are accurate when

$$\frac{\omega L}{c} \ll 1, \qquad \text{or} \qquad \frac{L}{c} \ll \frac{1}{\omega}. \tag{16.10}$$

With words: the fields change slowly if light can cross the region in a time much shorter than the time it takes the fields to change. Such fields are called **quasistatic**. In the **quasistatic approximation**, the magnetic field at each instant is the one that the currents at that instant would produce if they were steady, and the induced electric field is the one that the rate of change of that magnetic field produces, through the Faraday current. All of the examples of Chapter 15 and of this chapter so far have used it.

For the alternating current of the power grid, at 60 Hz, $c/\omega = 800$ km. Everything in a house or a laboratory is quasistatic, and only power lines hundreds of kilometers long are not. At 1 GHz, in the range of cell phones and microwave ovens, c/ω is only 5 cm, and circuits a few centimeters across are no longer quasistatic; such circuits radiate (Chapter 28). The same estimate applies to the electric field inside a conductor. There the current density is $\mathbf{J} = \sigma\mathbf{E}$, where σ is the conductivity, so the displacement current is smaller than the current by the factor $\varepsilon_0\omega/\sigma$. For copper, with $\sigma = 5.96 \times 10^7$ S/m at room temperature,[7] that factor is about 6×10^{-17} at 60 Hz, and even at 1 GHz it is about 10^{-9} (Exercise P16.9).

The steady ramps of the earlier examples are special (Figure 16.13). In the solenoid with dI/dt constant, the induced field (15.9) does not change in time, so it has no displacement current, and there is nothing more to add: the first step is exact. In the capacitor charged by a constant current, the magnetic field does not change in time, so it induces no electric field, and again the first step is exact, although (16.6) gives that field only with the fringing ignored (Section 16.3.2). (Both statements hold once the ramp has been going long enough. When it starts, the change spreads outward at the speed of light, Section 16.5.)

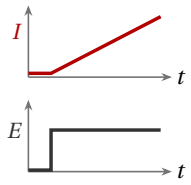


Figure 16.13 A steady ramp: the current rises at a constant rate (top), so the field it induces is constant in time (bottom), has no displacement current, and adds nothing more.

16.4.2 Successive approximations

When the fields are not quasistatic, the corrections can still be added one at a time, and in one case the process can be carried to the end.

[7] This is the conductivity of pure, annealed copper at 20°C, from the recommended resistivity $1.678 \times 10^{-8}\,\Omega$m (uncertain by 0.5%) in "Electrical resistivity of copper, gold, palladium, and silver," R. A. Matula, J. Phys. Chem. Ref. Data 8, 1147–1298 (1979). https://doi.org/10.1063/1.555614

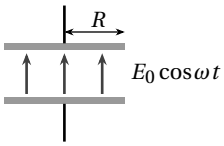
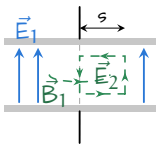


Figure 16.14 The capacitor of Example 16.3, driven by an alternating current.

Example 16.3

The capacitor of Example 16.1, with circular plates of radius R, is driven by an alternating current, so that the electric field near its axis is $E_0 \cos\omega t\,\hat{\mathbf{z}}$. Including the magnetic field that curls around the changing $\mathbf{E}$ and the electric field that this changing $\mathbf{B}$ induces in turn, find the first correction to the electric field between the plates. At what frequency does it change the field at the edge of plates 5.0 cm in radius by 1%?

Solution:



The first step is a sketch of what is given and what is asked: the capacitor from the side, with the field $\mathbf{E}_1$ between the plates, and two loops, a circle for the magnetic field $\mathbf{B}_1$ that curls around the displacement current of $\mathbf{E}_1$, and a rectangle with one side on the axis for the correction $\mathbf{E}_2$ that the changing $\mathbf{B}_1$ induces.

Start with the uniform field $\mathbf{E}_1 = E_0 \cos\omega t\,\hat{\mathbf{z}}$. Its displacement current is $\mathbf{J}_D = \varepsilon_0 \partial \mathbf{E}_1/\partial t = -\varepsilon_0 \omega E_0 \sin\omega t\,\hat{\mathbf{z}}$, uniform across the gap, and the circle of radius s gives, as in Example 16.1, $2\pi s B_{1\phi} = \mu_0 J_D \pi s^2$, or

$$\mathbf{B}_1 = -\frac{\omega s}{2c^2} E_0 \sin\omega t\,\hat{\boldsymbol{\phi}}.$$

This field changes in time, so it has a Faraday current,

$$\mathbf{J}_{\text{Far}} = -\frac{\partial \mathbf{B}_1}{\partial t} = \frac{\omega^2 s}{2c^2} E_0 \cos\omega t\,\hat{\boldsymbol{\phi}},$$

which circles the axis, like the Faraday current in the coaxial cable of Example 15.5. So the correction $\mathbf{E}_2$ points along the axis and depends on s. Take a rectangle of height ℓ to the right of the axis, with one side on the axis, where $E_2 = 0$, and the other side at s. (Why is E_2 zero on the axis? Answer in footnote.[8]) Going up the side at s and down the axis is counterclockwise as drawn, so the normal points out of the page, along $-\hat{\boldsymbol{\phi}}$, and

$$\begin{aligned} E_2(s)\,\ell &= I_{\text{Far,enc}} \\ &= -\int_0^s \frac{\omega^2 \bar{s}}{2c^2} E_0 \cos\omega t\,\ell\, d\bar{s} \\ &= -\frac{\omega^2 s^2}{4c^2} E_0 \cos\omega t\,\ell. \end{aligned}$$

The field to this order is

$$\mathbf{E} = E_0 \cos\omega t \left[1 - \frac{1}{4}\left(\frac{\omega s}{c}\right)^2\right] \hat{\mathbf{z}}.$$

[8] The problem gives the field near the axis as $E_0 \cos\omega t$, so E_0 is the whole field there. The rectangle gives only the difference between E_2 at s and on the axis, so this one value has to be fixed.

The field is weaker at the edge than at the center by the fraction $(\omega R/2c)^2$. It reaches 1% when $\omega R/2c = 0.10$, or

$$\begin{aligned}\nu &= \frac{\omega}{2\pi} \\ &= \frac{0.20\,c}{2\pi R} \\ &= \frac{0.20\,(3.00\times10^8\,\mathrm{m/s})}{2\pi(0.050\,\mathrm{m})} \\ &= 190\,\mathrm{MHz}.\end{aligned}$$

At 60 Hz the fraction is 10^{-15}.

Reality check: The quantity $\omega R/c$ has no units, as it must in a correction factor, and the fraction $(\omega R/2c)^2$ is of the order $(\omega L/c)^2$ that the estimate of Section 16.4.1 gave, with $L = R$. As $\omega \to 0$ the correction vanishes, leaving the uniform field of Example 16.1.

The correction $\mathbf{E}_2$ changes in time too, so its displacement current adds to the curl of $\mathbf{B}$ and corrects $\mathbf{B}_1$, whose change corrects $\mathbf{E}$ again, and so on. Each step brings another factor of $-(\omega s/2c)^2$, divided by the square of the step number, and the field becomes (Exercise P16.10)

$$\mathbf{E} = E_0\cos\omega t\left[1-\left(\frac{x}{2}\right)^2+\frac{1}{(2!)^2}\left(\frac{x}{2}\right)^4-\frac{1}{(3!)^2}\left(\frac{x}{2}\right)^6+\cdots\right]\hat{\mathbf{z}}, \qquad x=\frac{\omega s}{c}.$$

The series in brackets is the Bessel function $J_0(x)$ of Section 6.3.[9] At the first zero of J_0, $x = 2.405$, the field at the edge of the plates is zero. For plates 5.0 cm in radius that happens at 2.3 GHz, close to the frequency of a microwave oven. Then the capacitor is no longer a capacitor at all: closed off by a conducting cylinder at its edge, it becomes a resonant cavity, a kind of device taken up in Section 23.7.

16.5 The prediction of waves

Far from any charges and currents, the two curl equations feed each other. A changing magnetic field produces an electric field; if that field changes, it produces a magnetic field; if that one changes, it produces another electric field; and nothing else is needed (Figure 16.15). The fields can keep each other going in empty space, with no charges or currents anywhere near. In empty space, where $\rho = 0$ and $\mathbf{J} = 0$, Maxwell's equations are

$$\nabla\cdot\mathbf{E}=0, \qquad \nabla\cdot\mathbf{B}=0, \qquad \nabla\times\mathbf{E}=-\frac{\partial\mathbf{B}}{\partial t}, \qquad \nabla\times\mathbf{B}=\mu_0\varepsilon_0\frac{\partial\mathbf{E}}{\partial t}. \tag{16.11}$$

Take the curl of Faraday's law, and use the Ampère–Maxwell law to replace the curl of $\mathbf{B}$:

$$\nabla\times(\nabla\times\mathbf{E}) = -\frac{\partial}{\partial t}(\nabla\times\mathbf{B}) = -\mu_0\varepsilon_0\frac{\partial^2\mathbf{E}}{\partial t^2}.$$

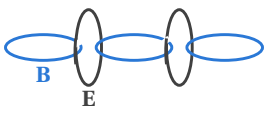


Figure 16.15 A schematic of fields that keep each other going: a changing magnetic field (blue loops) is circled by an electric field (black), whose change is circled by a magnetic field, and so on.

[9] The calculation is from The Feynman Lectures on Physics, Vol. II, Section 23-2. https://feynmanlectures.caltech.edu/II_23.html.

By the identity (1.58), the left side is $\nabla(\nabla\cdot\mathbf{E}) - \nabla^2\mathbf{E}$, and the first term is zero because $\nabla\cdot\mathbf{E} = 0$. So

$$\nabla^2\mathbf{E} = \mu_0\varepsilon_0\frac{\partial^2\mathbf{E}}{\partial t^2}, \qquad \text{and in the same way} \qquad \nabla^2\mathbf{B} = \mu_0\varepsilon_0\frac{\partial^2\mathbf{B}}{\partial t^2} \tag{16.12}$$

(Exercise P16.12). In Cartesian coordinates the Laplacian of a vector is the Laplacian of each component, so every Cartesian component of **E** and of **B** obeys the same equation,

$$\nabla^2 f = \frac{1}{v^2}\frac{\partial^2 f}{\partial t^2}, \qquad v = \frac{1}{\sqrt{\mu_0\varepsilon_0}}. \tag{16.13}$$

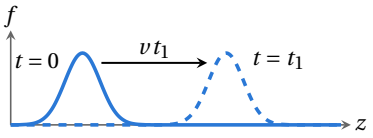


Figure 16.16 A solution $g(z-vt)$ of the one-dimensional wave equation, at $t = 0$ and at a later time t_1. The pulse keeps its shape and moves a distance vt_1 along $+z$.

This is the **wave equation**. To see why it has that name, take a function that depends only on z and t, so that the equation is $\partial^2 f/\partial z^2 = (1/v^2)\partial^2 f/\partial t^2$. Any function of the combination $z - vt$, $f = g(z-vt)$, solves it, whatever the shape of g: with $u = z - vt$, the chain rule gives $\partial^2 f/\partial z^2 = g''(u)$ and $\partial^2 f/\partial t^2 = v^2 g''(u)$. And $g(z-vt)$ is the shape $g(z)$ moved a distance vt along $+z$ (Figure 16.16): it describes a disturbance that travels in the $+z$ direction at the speed v without changing shape. A function $h(z+vt)$ travels in the $-z$ direction, and the general solution in one dimension is a sum of the two. Electric and magnetic fields in empty space travel as waves, at the speed

$$\begin{aligned} v = \frac{1}{\sqrt{\mu_0\varepsilon_0}} &= \frac{1}{\sqrt{(1.25664\times10^{-6}\,\mathrm{N/A^2})(8.85419\times10^{-12}\,\mathrm{C^2/(N\cdot m^2)})}} \\ &= 2.9979\times10^8\,\mathrm{m/s}. \end{aligned}$$

That is the speed of light. Light is an electromagnetic wave.

Maxwell made this calculation, with the measurements of his day, in 1862 and again in 1865 (Section 2.4). In 1865 the speed that his equations gave, 3.107×10^8 m/s, came from purely electrical measurements by Wilhelm Weber and Rudolf Kohlrausch, and he compared it with the speed of light measured by Hippolyte Fizeau, 3.149×10^8 m/s, and by Léon Foucault, 2.98×10^8 m/s.[10] Nothing about light had gone into the electrical measurements. Today the agreement is exact by construction: the speed of light has had a defined value since 1983, and since 2019 μ_0 and ε_0 have been measured quantities whose product is exactly $1/c^2$ (Section 17.4). From here on we write c for $1/\sqrt{\mu_0\varepsilon_0}$. Two decades after Maxwell's paper, Hertz made electromagnetic waves with electric circuits, detected them across his laboratory, and showed that they travel at about the speed of light.

Both new terms are needed for this. Without Maxwell's term, the curl of **B** would be zero in empty space, and the same steps would give $\nabla^2\mathbf{E} = 0$: the electric field would obey Laplace's equation at every instant, and there would be no waves at all. And the sign matters. If the Ampère–Maxwell law had a minus sign like Faraday's law, the steps would give $\nabla^2\mathbf{E} = -\mu_0\varepsilon_0\,\partial^2\mathbf{E}/\partial t^2$, which is Laplace's equation in the variables x, y, z, and vt. Its solutions do not travel: a field pattern $\sin kz$, for example, grows or dies away in place, as $e^{\pm kvt}$ (Exercise P16.14).



Heinrich Hertz (1857–1894, German) Hertz was born in Hamburg. After a start in engineering he studied physics in Berlin with Hermann von Helmholtz and Gustav Kirchhoff, took his doctorate under Helmholtz in 1880, and stayed on as Helmholtz's assistant. In 1885 he became professor of physics at the technical college in Karlsruhe, where in 1886–88 he made electromagnetic waves with a spark gap and detected them with a loop of wire that had a tiny gap of its own. With a zinc sheet as a reflector he set up standing waves and measured their wavelength, and he showed that the waves are reflected, refracted and polarized as light is. In 1887 he also noticed that ultraviolet light helps a spark to jump, the first observation of the photoelectric effect. He moved to Bonn in 1889 and died there at thirty-six. The SI unit of frequency is named after him. (Wikipedia)

[10] "A Dynamical Theory of the Electromagnetic Field," J. C. Maxwell, Phil. Trans. R. Soc. Lond. 155, 459–512 (1865). https://doi.org/10.1098/rstl.1865.0008.

The wave equations alone do not say everything. The fields must also satisfy the four equations (16.11) themselves, and those tie **E** and **B** together. A wave that travels along z with **E** along x, for example, must have a magnetic field along y, with $B = E/c$ (Exercise P16.13). Chapter 21 takes up these waves, the energy and momentum they carry, and what happens to them in matter.

16.6 Summary

When the electric field changes in time, the curl equation for the magnetic field gains a term: the magnetic field curls around $\varepsilon_0 \partial \mathbf{E}/\partial t$ in exactly the way that it curls around a current. With this term and Faraday's law, Maxwell's equations are complete, the electric and magnetic fields are no longer independent, and together they predict electromagnetic waves that travel at the speed of light.

The two curl equations

$\nabla \times \mathbf{E} = -\dfrac{\partial \mathbf{B}}{\partial t}, \quad \oint \mathbf{E} \cdot d\boldsymbol{\ell} = -\dfrac{d}{dt}\int \mathbf{B} \cdot d\mathbf{a}$	Faraday's law: a changing magnetic field produces an electric field with curl
$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \mu_0 \varepsilon_0 \dfrac{\partial \mathbf{E}}{\partial t}$	the Ampère–Maxwell law; the conservation of charge requires Maxwell's term
$\mathbf{J}_D = \dfrac{\partial \mathbf{D}}{\partial t}$, $\oint \mathbf{B} \cdot d\boldsymbol{\ell} = \mu_0 \left(I_{\text{enc}} + I_{D,\text{enc}}\right)$	the displacement current density, $\varepsilon_0 \partial \mathbf{E}/\partial t$ in empty space; **B** curls around it as around a current, with no minus sign; $I_{D,\text{enc}} = \varepsilon_0 \, d\Phi_E/dt$
$\mathbf{J}_p = \dfrac{\partial \mathbf{P}}{\partial t}, \quad \nabla \times \mathbf{H} = \mathbf{J}_f + \mathbf{J}_D$	the polarization current, part of $\mathbf{J}_D$ in matter, and the Ampère–Maxwell law in terms of free currents; Faraday's law is the same in matter

Slow fields and waves

$\dfrac{\omega L}{c} \ll 1$	the fields are quasistatic: light crosses the region of size L in much less than the time $1/\omega$ in which the fields change, and each round trip between **E** and **B** is smaller by $(\omega L/c)^2$
$\nabla^2 \mathbf{E} = \mu_0 \varepsilon_0 \dfrac{\partial^2 \mathbf{E}}{\partial t^2}, \quad c = \dfrac{1}{\sqrt{\mu_0 \varepsilon_0}}$	the wave equation, for **E** and for **B** in empty space; its solutions travel at the speed of light

The displacement-current method. When the current and the displacement current ($\mathbf{J}_D = \varepsilon_0 \partial \mathbf{E}/\partial t$ in empty space) together have the symmetry of a long

wire, a long solenoid, or a slab, find $\mathbf{B}$ from $\oint \mathbf{B}\cdot d\boldsymbol{\ell} = \mu_0(I_{\text{enc}} + I_{D,\text{enc}})$, treating $\mathbf{J}_D$ as a current. *Colton's Third Commandment* applies: thou shalt always draw an Amperian loop. Lenz's law has no counterpart here.

Where the magnetic field comes from. At every instant $\mathbf{B}$ is the Biot–Savart integral of $\mathbf{J}+\mathbf{J}_D$ over all of space, (16.7), but the displacement current of the field that the charges would produce by Coulomb's law, where they are at that instant, is a gradient field and adds nothing to it; only the polarization current and the displacement current of an induced electric field do. So the magnetic field of slowly changing currents is the Biot–Savart field of the real currents, such as those in the wires and on the plates of a charging capacitor.

Which method? Figure 16.17 gathers the methods for fields that change in time into one chart, as Figures 14.9 and 14.10 do for static fields.

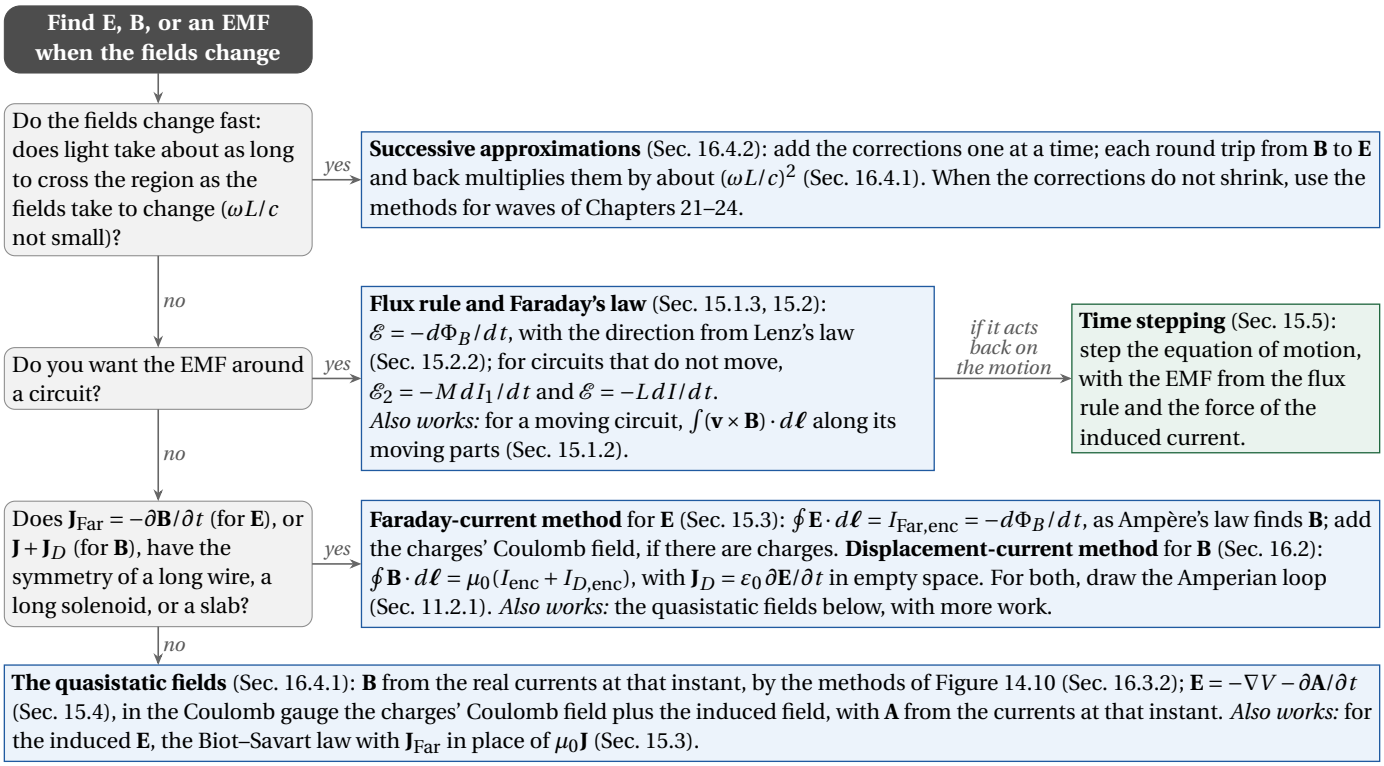


Figure 16.17 Which method? The methods of Chapters 15 and 16 for finding **E**, **B**, or an EMF when the fields change in time. Gray: questions; blue: methods worked by hand; green: numerical methods. "Also works" names a second method that usually takes longer.

Exercises

Exercises for 16.1 Displacement current and the Ampère–Maxwell law

Warm-up questions Read Section 16.1, then answer before class

W16.1 Section 16.1.3 compares the displacement current density $\varepsilon_0 \partial \mathbf{E}/\partial t$ with the Faraday current density $-\partial \mathbf{B}/\partial t$ of Section 15.2.4. Which of the two has the units of a current density? In empty space, is either one a flow of charge? Explain.

W16.2 A footnote in Section 16.1.3 notes that some books define the displacement current density as $\varepsilon_0 \partial \mathbf{E}/\partial t$ even inside matter. With that definition, where does the polarization current $\partial \mathbf{P}/\partial t$ appear in the Ampère–Maxwell law, and where does it appear with this book's $\mathbf{J}_D = \partial \mathbf{D}/\partial t$? What mistake could you make by mixing the two?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P16.1 Show that $I_{\text{enc}} + I_{D,\text{enc}}$ through a surface bounded by a given loop is the same for every such surface. (Two such surfaces together form a closed surface; use the divergence theorem, the continuity equation, and Gauss's law.)

P16.2 A metal ball of radius a with charge Q_0 is embedded in a large block of material that conducts slightly, with conductivity σ (Figure 16.18). Take the permittivity of the material to be ε_0.

(a) Show that the charge on the ball decreases as $Q_0 e^{-t/\tau}$, with $\tau = \varepsilon_0/\sigma$.

(b) Find the current density $\mathbf{J}$ and the displacement current density $\mathbf{J}_D$ in the material, and show that $\mathbf{J} + \mathbf{J}_D = 0$.

(c) What is the magnetic field? Explain with a symmetry argument, and explain why Ampère's law without Maxwell's term would contradict that argument.

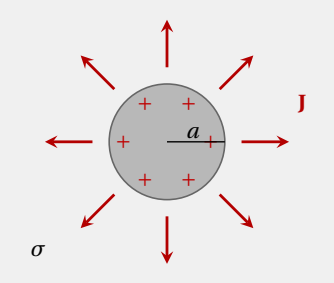


Figure 16.18 The charged ball of Exercise P16.2, embedded in a slightly conducting material, with its charge flowing outward in all directions.

P16.3 A parallel-plate capacitor with circular plates of radius R and area $A = \pi R^2$ is filled with a linear dielectric of relative permittivity ε_r, and a current I flows onto its positive plate (Figure 16.19). Ignore fringing.

(a) Find $\mathbf{D}$, $\mathbf{E}$, and $\mathbf{P}$ between the plates when the free charge on the positive plate is Q.

(b) Find $\varepsilon_0 \partial \mathbf{E}/\partial t$ and the polarization current density $\mathbf{J}_p = \partial \mathbf{P}/\partial t$, and show that their sum, the displacement current density $\mathbf{J}_D = \partial \mathbf{D}/\partial t$, is $(I/A)\,\hat{\mathbf{n}}$, where $\hat{\mathbf{n}}$ points from the positive plate toward the negative one.

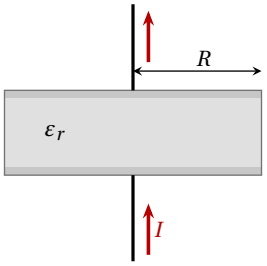


Figure 16.19 The capacitor of Exercise P16.3, filled with a dielectric and charged by a current I.

(c) What fraction of the displacement current between the plates is polarization current? Pure water at 20°C has the dielectric constant 80.1 (Table 9.2), and it keeps this value at frequencies up to far above 1 MHz.[11] At 1 MHz it also conducts so little that its conduction current can be ignored. Show that in a capacitor filled with pure water and driven at 1 MHz, 98.8% of the displacement current between the plates is polarization current.

(d) Show that the magnetic field between the plates is the same as without the dielectric, (16.6).

(e) Find the same result from (16.5), using $\mathbf{H}$ and $\partial\mathbf{D}/\partial t$.

P16.4 Section 16.1.2 points out that its argument does not show that nothing else would work. Find a term, other than Maxwell's, that could be added to the right side of (16.1) without spoiling its agreement with the continuity equation, and that would vanish whenever nothing changes in time. (Look at the other curl equation.) What, then, establishes Maxwell's term?

Exercises for 16.2 Solving for B when given a changing E: displacement-current method (high symmetry)

Warm-up questions Read Section 16.2, then answer before class

W16.3 A classmate predicts that if the plates of Example 16.1 were moved twice as far apart (still close enough that the fringing can be ignored), the magnetic field between them would be half as strong, since $E = \Delta V/d$. Do you agree? Explain with the displacement current.

W16.4 Suppose the gap of Example 16.1 were filled with a material that conducts slightly (take its permittivity to be ε_0), so that some charge flows across from plate to plate while the capacitor charges. Would the magnetic field between the plates be different? (What becomes of the current that reaches the bottom plate?)

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

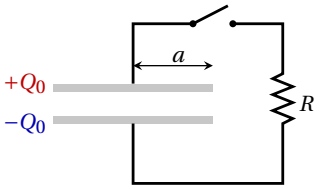


Figure 16.20 The capacitor of Exercise P16.5, discharging through a resistor R.

P16.5 A capacitor has circular plates of radius a separated by a distance $d \ll a$, with air between them. It is charged to a voltage V_0 and, at $t = 0$, connected across a resistor R (Figure 16.20). Ignore the fringing of the electric field at the edges of the plates.

[11] The main dielectric relaxation of water, near which its molecules fall behind the field, is at about 17 GHz at 20°C: "A model for the complex permittivity of water at frequencies below 1 THz," H. J. Liebe, G. A. Hufford, and T. Manabe, Int. J. Infrared Millim. Waves 12, 659–675 (1991). https://doi.org/10.1007/BF01008897.

(a) Find the charge on the capacitor as a function of time.

(b) Find the magnetic field between the plates, inside and outside the edge of the plates. How does its direction compare with that of the field around the wire that leads to the positive plate?

(c) Evaluate the field at the edge of the plates at $t = 0$, for $a = 5.0\,\text{cm}$, $d = 1.0\,\text{mm}$, $V_0 = 100\,\text{V}$, and $R = 1.0\,\text{k}\Omega$. How long does the discharge take?

Exercises for 16.3 How the displacement current produces magnetic fields

Warm-up questions Read Section 16.3, then answer before class

W16.5 Example 16.1 found the magnetic field between the plates by counting only displacement current, and Section 16.3.2 found the same field from the real currents alone, with no displacement current in the calculation. How can both be right? Which of the two readings described at the start of this section does each calculation use?

W16.6 Section 16.3.2 finds the field between the plates as the field of an unbroken wire, $\mu_0 I/2\pi s$, minus that of the currents on the plates, $\mu_0 K(s)$. Close to the axis the wire's field is very large, yet the field between the plates is small there. Why does the plates' field nearly cancel the wire's there? Where on the plates is $K(s)$ zero, and why?

W16.7 For each of the following, does it add anything to the Biot–Savart integral (16.7) for **B**? Give a reason for each. (a) The displacement current of a growing point charge (Figure 16.9). (b) The polarization current in the rod of Example 16.2. (c) The rest of the displacement current in that rod, $\varepsilon_0 \partial \mathbf{E}/\partial t$.

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P16.6 This exercise derives (16.8) and applies it to the charging capacitor.

(a) Let **F** be a vector field that falls off faster than $1/r'$ far away. Show that $\hat{\boldsymbol{\imath}}/\imath^2 = \nabla'(1/\imath)$, where ∇' takes derivatives with respect to $\mathbf{r}'$, and use the product rule for the curl of a function times a vector (Section 1.7) to show that

$$\mathbf{F} \times \frac{\hat{\boldsymbol{\imath}}}{\imath^2} = \frac{\nabla' \times \mathbf{F}}{\imath} - \nabla' \times \left(\frac{\mathbf{F}}{\imath}\right).$$

(b) Show that the integral of the last term over all of space is zero, and so derive (16.8). (The volume integral of a curl can be turned into a surface integral, $\int \nabla \times \mathbf{G}\, d\tau = -\oint \mathbf{G} \times d\mathbf{a}$, by applying the divergence theorem to $\mathbf{G} \times \mathbf{c}$, where **c** is a constant vector. Leave out a small

sphere around the field point, and show that its surface contributes nothing either.)

(c) A capacitor is charged by a constant current, as in Example 16.1, but without ignoring the fringing. Explain why its displacement current is a gradient field everywhere and falls off fast enough far away, and conclude that its Biot–Savart integral is zero.

(d) The displacement current of Example 16.1 is uniform between the plates and zero outside them. Find its curl, as in Exercise P1.21. (It is zero everywhere except on the cylinder $s = R$ between the plates, where it is a delta function in s.)

P16.7 Section 11.1.2 derived $\nabla \times \mathbf{B} = \mu_0 \mathbf{J}$ from the Biot–Savart law and dropped a term that contains $\nabla' \cdot \mathbf{J}$, which is zero for steady currents. Suppose instead that the currents are not steady, as when a capacitor is charging, and take $\mathbf{B}$ at each instant to be the Biot–Savart field of the currents at that instant.

(a) Keep the dropped term, use the continuity equation (10.12), and show that

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \mu_0 \varepsilon_0 \frac{\partial \mathbf{E}_C}{\partial t},$$

where $\mathbf{E}_C$ is the field that the charges would produce by Coulomb's law at that instant.

(b) Explain why this means that the Biot–Savart field of the real currents satisfies the Ampère–Maxwell law exactly whenever the electric field is $\mathbf{E}_C$, and why it does not when part of the electric field is induced.

P16.8 Section 16.3.3 showed that the displacement current of the field that the charges would produce by Coulomb's law, where they are at that instant, adds nothing to the Biot–Savart integral, because that field is a gradient field at every instant. This exercise shows why the same argument never applies to the Faraday current.

(a) Show that $\nabla \cdot \mathbf{J}_{\text{Far}} = 0$ everywhere and at every instant.

(b) Suppose that $\mathbf{J}_{\text{Far}} = \nabla f + \mathbf{G}$, where $\nabla \cdot \mathbf{G} = 0$ and f goes to zero far away. Show that f satisfies Laplace's equation everywhere, and use the first uniqueness theorem (Section 7.1.3) to show that $f = 0$. A Faraday current has no gradient part.

(c) In empty space the displacement current has a divergence: show that $\nabla \cdot \mathbf{J}_D = \partial \rho / \partial t$. Show also that $\nabla \cdot (\mathbf{J} + \mathbf{J}_D) = 0$, and explain in words why $\mathbf{J} + \mathbf{J}_D$, not $\mathbf{J}_D$ alone, is the quantity that plays the role of $\mathbf{J}_{\text{Far}}$.

(d) Suppose that magnetic charges existed, with density ρ_m and current density $\mathbf{J}_m$, so that $\nabla \cdot \mathbf{B} = \mu_0 \rho_m$ and Faraday's law became $\nabla \times \mathbf{E} = -\mu_0 \mathbf{J}_m - \partial \mathbf{B}/\partial t$. Would the Faraday current then have a gradient part? What would play the role of $\mathbf{J} + \mathbf{J}_D$?

Exercises for 16.4 When is fast, when is slow

Warm-up questions Read Section 16.4, then answer before class

W16.8 A classmate says that the quasistatic approximation must work better for a capacitor driven by a small alternating current than for one driven by a large current of the same frequency, since a small current makes small fields and so small corrections. Do you agree? Explain with the estimate of Section 16.4.1.

W16.9 In Example 16.3, the field at the edge of plates 5.0 cm in radius is 1% weaker than at the center when the frequency is 190 MHz. At that frequency, about how much weaker would the field be at the edge of plates 10 cm in radius? At what frequency would those larger plates show a 1% difference?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P16.9 A straight copper wire of radius a, length ℓ, and conductivity σ has a voltage $V_0 \cos \omega t$ between its ends (Figure 16.21).

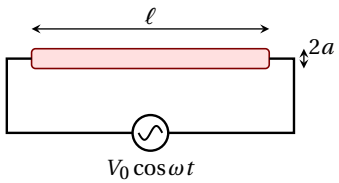


Figure 16.21 The wire of Exercise P16.9, with an alternating voltage between its ends.

(a) Assume that the electric field and the current density are uniform throughout the wire, and find $\mathbf{E}$, $\mathbf{J}$, and the displacement current density $\mathbf{J}_D$ in it, taking $\mathbf{D} = \varepsilon_0 \mathbf{E}$ in the metal.

(b) Find the amplitudes of the current and of the displacement current through a cross section of the wire, and show that their ratio is $\varepsilon_0 \omega / \sigma$.

(c) For copper, $\sigma = 6.0 \times 10^7$ S/m. At what frequency would the two be equal? Compare it with the frequency of visible light, about 5×10^{14} Hz.

(d) The answer to part (c) is misleading, because the assumption of part (a) fails at much lower frequencies. Use Example 15.5 to explain why an alternating current does not stay spread uniformly over the cross section of a wire.

P16.10 This exercise carries Example 16.3 to the end.

(a) Find the correction $\mathbf{B}_3$ to the magnetic field that curls around the displacement current of $\mathbf{E}_2$, and the correction $\mathbf{E}_4$ to the electric field that its change induces. Show that $\mathbf{E}_4 = \frac{1}{64}(\omega s/c)^4 E_0 \cos \omega t \, \hat{\mathbf{z}}$.

(b) Show that the fields

$$\mathbf{E} = E_0 J_0\left(\frac{\omega s}{c}\right)\cos\omega t\,\hat{\mathbf{z}}, \qquad \mathbf{B} = -\frac{E_0}{c} J_1\left(\frac{\omega s}{c}\right)\sin\omega t\,\hat{\boldsymbol{\phi}}$$

satisfy Faraday's law and the Ampère–Maxwell law exactly between the plates. You will need $J_0'(x) = -J_1(x)$ and $(1/x)\,d(xJ_1)/dx = J_0$.

(c) Plot E at $t = 0$ against s/R for $\omega R/c = 0.5$, 1, and 2.405.

P16.11 For each of the following, find $\omega L/c$ and decide whether the fields are quasistatic.

(a) A 60 Hz transformer 30 cm across.

(b) A microwave oven, at 2.45 GHz, 30 cm across.

(c) A computer processor chip 2 cm across, with a clock frequency of 3 GHz.

(d) The electrical activity of the brain, at frequencies up to about 100 Hz, over a head 15 cm across.

(e) The antenna of an AM radio station, at 1 MHz, 75 m tall.

Exercises for 16.5 The prediction of waves

Warm-up questions Read Section 16.5, then answer before class

W16.10 Which of these functions solve the one-dimensional wave equation, $\partial^2 f/\partial z^2 = (1/v^2)\,\partial^2 f/\partial t^2$? For each one that does, describe how it changes in time. (a) $e^{-(z+vt)^2/w^2}$; (b) $\sin kz\cos kvt$; (c) $e^{-(z^2+v^2t^2)/w^2}$.

W16.11 Suppose that Maxwell's term were missing, so that, as this section shows, the electric field in empty space would obey Laplace's equation at every instant. If the current in a loop of wire far away were suddenly switched on, when would the fields here first change? (What fixes the solution of Laplace's equation at each instant?) How does Maxwell's term change the answer?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

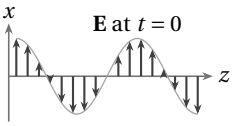


Figure 16.22 The field of Exercise P16.13 at $t = 0$.

P16.12 Derive the wave equation for **B** in empty space, the second of (16.12), by taking the curl of the Ampère–Maxwell law.

P16.13 An electric field in empty space is $\mathbf{E} = E_0\cos(kz - \omega t)\,\hat{\mathbf{x}}$ (Figure 16.22).

(a) Use Faraday's law to find **B**, taking it to have no part that is constant in time.

(b) Show that the Ampère–Maxwell law is then satisfied only if $\omega = ck$.

(c) Show that **E** and **B** satisfy both divergence equations, that they are perpendicular to each other and to the direction of travel, that $B = E/c$, and that $\mathbf{E} \times \mathbf{B}$ points in the direction of travel.

P16.14 Suppose that the Ampère–Maxwell law had a minus sign, $\nabla \times \mathbf{B} = -\mu_0 \varepsilon_0 \partial \mathbf{E}/\partial t$, in empty space.

(a) Show that **E** would obey $\nabla^2 \mathbf{E} = -\mu_0 \varepsilon_0 \partial^2 \mathbf{E}/\partial t^2$.

(b) Show that $f = \sin kz\, e^{\pm kvt}$ satisfies the one-dimensional version of this equation, and that a traveling pulse $g(z - vt)$ does not, unless g is a straight line.

(c) Why would such a world be unstable?

P16.15 In a linear dielectric with $\varepsilon = \varepsilon_r \varepsilon_0$, which is not magnetic ($\mathbf{B} = \mu_0 \mathbf{H}$) and has no free charges or currents, Maxwell's equations are $\nabla \cdot \mathbf{D} = 0$, $\nabla \cdot \mathbf{B} = 0$, $\nabla \times \mathbf{E} = -\partial \mathbf{B}/\partial t$, and $\nabla \times \mathbf{H} = \partial \mathbf{D}/\partial t$.

(a) Derive the wave equation for **E**, and show that the waves travel at $c/\sqrt{\varepsilon_r}$.

(b) Your introductory physics course probably said that light travels at c/n in a material, where n is the index of refraction. What is n in terms of ε_r?

(c) Water has $\varepsilon_r = 80$ for static fields, but its index of refraction for visible light is 1.33. What does that tell you about ε_r?

P16.16 Fields that depend only on z and t, with $\mathbf{E} = E(z,t)\,\hat{\mathbf{x}}$ and $\mathbf{B} = B(z,t)\,\hat{\mathbf{y}}$, obey

$$\frac{\partial B}{\partial t} = -\frac{\partial E}{\partial z}, \qquad \frac{\partial E}{\partial t} = -\frac{1}{\mu_0 \varepsilon}\frac{\partial B}{\partial z},$$

in a medium of permittivity ε that is not magnetic.

(a) Show that these follow from Faraday's law and (16.5).

(b) Solve them numerically by the *finite-difference time-domain* method. Keep E at grid points $z_j = j\,\Delta z$ and B at the points halfway between (Figure 16.23), and update them in turn:

$$B_{j+1/2} \leftarrow B_{j+1/2} - \frac{\Delta t}{\Delta z}\left(E_{j+1} - E_j\right),$$
$$E_j \leftarrow E_j - \frac{\Delta t}{\mu_0 \varepsilon_j \Delta z}\left(B_{j+1/2} - B_{j-1/2}\right).$$

In vacuum, start with a Gaussian pulse of E and $B = 0$, and show that it splits into two pulses that travel in opposite directions at the speed of light. Use $c\,\Delta t = \Delta z/2$. (What happens if $c\,\Delta t > \Delta z$?)

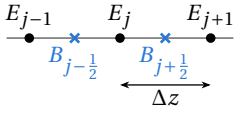


Figure 16.23 The staggered grid of Exercise P16.16: E at the grid points, B halfway between them.

(c) Now send a single pulse, with $B = E/c$, toward a layer of glass ($\varepsilon_r = 2.25$) of thickness $1.0\,\mu$m. Find the fraction of the pulse's energy that is reflected, for a Gaussian pulse $E \propto e^{-(z-z_0)^2/w^2}$ with $w = 0.05\,\mu$m, much less than the thickness of the glass, and with $w = 2$, 5, and $10\,\mu$m. Compare the first with $2R/(1+R)$, the fraction expected from two surfaces that each reflect $R = 4\%$ of the energy, counting all of the reflections between them. Why does a layer much thinner than the pulse reflect so little?

P16.17 Show that the potentials $V = 0$ and $\mathbf{A} = (E_0/\omega)\sin(kz - \omega t)\,\hat{\mathbf{x}}$, with $\omega = ck$, give the fields of Exercise P16.13. Is this $\mathbf{A}$ in the Coulomb gauge?

Exercises for 16.6 Summary

P16.18 True or false? Explain your reasoning.

(a) An electric field points along $+\hat{\mathbf{z}}$ and is increasing. The magnetic field curls around it in the $+\hat{\boldsymbol{\phi}}$ direction.

(b) Lenz's law gives the direction in which the magnetic field curls around a displacement current.

(c) The displacement current between the plates of a charging capacitor equals the current in the wires that charge it.

(d) Inside a copper wire carrying a 60 Hz current, the displacement current is about as large as the current.

P16.19 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as the corrections to the quasistatic fields of a capacitor driven at high frequency, an electromagnetic pulse traveling through a layer of glass, the magnetic field of a charging capacitor found from the Biot–Savart law of the currents in its wires and plates, the fringing field and the displacement current of a capacitor whose plates have finite size, or the magnetic field of the polarization current in a dielectric between the plates of a charging capacitor.

Part V

Units, Circuits, and Conservation Laws

I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind.

— William Thomson (Lord Kelvin), "Electrical Units of Measurement" (1883)*

*W. Thomson (Lord Kelvin), "Electrical units of measurement," lecture at the Institution of Civil Engineers, 3 May 1883, in *Popular Lectures and Addresses*, Vol. 1 (Macmillan, London, 1889), p. 73.

Chapter 17

Unit Systems

Every equation you have met so far has been written in SI units, and you have probably accepted the constants in them — ε_0, and that peculiar $\mu_0 = 4\pi \times 10^{-7}\,\mathrm{N/A^2}$ — without asking where they came from. This chapter asks. By now you have met the equations it works with — Coulomb's law, the Lorentz force law, the force between two wires, Maxwell's equations, and the energy stored in the fields — so you can watch what happens to each of them. The answer turns out to be more interesting than bookkeeping, because it exposes a genuine logical problem in defining electrical units at all, and because the other common solution to that problem — Gaussian units — is still what you will meet in graduate courses and in a good deal of the research literature. Later sections look at the units researchers use for atoms and for elementary particles, and at the 2019 redefinition of the SI.

17.1 SI versus Gaussian units

17.1.1 Where base units come from

The SI system singles out a few quantities as **base units** — the meter, the kilogram, the second, and four others — and derives everything else from them. A newton is not fundamental; it is a $\mathrm{kg\cdot m/s^2}$, because Newton's second law says force is mass times acceleration.

The base units began life tied to tangible things. A meter was originally one ten-millionth of the distance from the equator to the North Pole. A kilogram was the mass of a reference block of platinum kept in France, chosen to match a cube of water ten centimeters on a side, so that water would come out at $1\,\mathrm{g/cm^3}$. Both have since been replaced by sharper definitions as measurement improved.

Now put yourself in the position of the first people trying to do the same job for electricity and magnetism. You have discovered that two charges push on each other with a force that grows with the charges and falls off with distance (Figure 17.1),

$$F \sim \frac{q_1 q_2}{r^2},$$

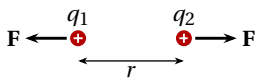


Figure 17.1 Two charges a distance r apart push on each other, with a force that grows with the charges and falls off with the distance.

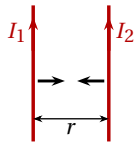


Figure 17.2 Two parallel wires carrying currents in the same direction attract each other.

and you have discovered that moving charges constitute currents, which produce a different sort of field that pushes on other moving charges. You combine the two into the Lorentz force law, (10.2),

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}).$$

You have also found that two nearby parallel currents attract (Figure 17.2), with a force we will call the *Ampère force law*:[1]

$$F \sim \frac{I_1 I_2}{r}.$$

What you do *not* have is a definition of the coulomb — and therefore none of the volt, the unit of electric field, or the tesla, since each is defined in terms of the one before it. Nor do you have numerical values for the constants in the two force laws, ε_0 and μ_0. So you set out to pin one of them down.

17.1.2 Two attempts, both of which fail

Attempt 1: use Coulomb's law. Write it precisely,

$$F = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r^2},$$

set up two charged balls, and measure the force between them.

This does not work. Without a definition of the coulomb there is no way to measure ε_0 — its value depends entirely on how large you decide a coulomb is, so it could come out to be anything at all. And running the argument the other way, without a value for ε_0 there is no way to say what a coulomb is. You are stuck.

Attempt 2: use the Ampère force law. If you could define an ampere you would get the coulomb free, since an ampere is a coulomb per second. So measure the force between two long straight wires instead,

$$F = \left(\frac{\mu_0 L}{2\pi}\right)\frac{I_1 I_2}{r}.$$

This fails in exactly the same way. Without a definition of the ampere there is no measuring μ_0, and without μ_0 there is no defining the ampere.

17.1.3 The vicious cycle

The trouble is that seven quantities are tangled together: the coulomb, the ampere, the unit of electric field, the volt, the tesla, ε_0, and μ_0. Fix *any one* of them and all six others follow. Know the coulomb, and the ampere is C/s, the electric field

[1] I have never heard this equation called by that name, but it seems a fair analog of Coulomb's law. Example 10.2 derived it: two parallel wires of length L, a distance $r \ll L$ apart, attract with $F = (\mu_0 L/2\pi)\, I_1 I_2 / r$.

unit is N/C, the volt is electric field times distance, the tesla is electric field over velocity, and ε_0 and μ_0 drop out of the two force laws by measurement. Know the ampere instead, and you get the other six by the same chain. The same is true starting from any of the seven.

But knowing none of them, there is no way in.

It is like a crime drama with seven guilty suspects, none of whom can be convicted until one of them flips. And — this is the point worth holding on to — it does not greatly matter which one does. The physics is identical either way. All that changes is which quantity is handed a defined value and which ones are then left to be measured.

! Seven quantities, each definable from any other, and none of them independently known. One has to be *fixed by decree* before the rest can be measured. Which one you pick is what separates SI from Gaussian units.

17.1.4 How SI breaks the cycle

The SI solution, in the form that held until 2019, was to **make μ_0 a defined quantity**:

$$\mu_0 \equiv 4\pi \times 10^{-7}\,\mathrm{N/A^2}.$$

That is where the odd-looking constant you have been carrying since your introductory electricity and magnetism course comes from. The 4π is historical; the 10^{-7} is chosen to make an ampere come out at a convenient size for everyday currents.

With μ_0 fixed, the Ampère force law gives you a way to realize the ampere,[2] the ampere gives the coulomb, and everything else follows. The price is that ε_0 becomes something you measure rather than choose,

$$\varepsilon_0 = 8.854 \times 10^{-12}\,\mathrm{C^2/N\,m^2},$$

and that a coulomb ends up being an enormous amount of charge. (From 1983, when the meter was redefined so that the speed of light has an exact value, $\varepsilon_0 = 1/\mu_0 c^2$ was exact as well, until the change described in Section 17.4.)

17.1.5 How Gaussian units break it instead

Given the argument above, there are at least six other ways to break the cycle. The Gaussian system takes a different one: it **makes ε_0 a defined quantity**,

$$\boxed{\varepsilon_0 = \frac{1}{4\pi}}, \qquad \text{(Gaussian units)}$$

and the reason for that particular value is to make Coulomb's law come out clean. In fact the symbol ε_0 is never written at all in Gaussian equations — you simply see factors of 4π where it would have been.

Gaussian units make a second, logically independent choice: centimeters, grams and seconds as base units rather than meters, kilograms and seconds. Keep

[2] For much of the twentieth century the ampere was realized with current balances, which weigh the force between current-carrying coils rather than between two long, straight wires. The principle is the same, and that is all the argument needs.

that in mind separately, because it affects every equation containing a length or a mass. It also means the derived units of force and energy are the dyne and the erg:

$$1\,\text{dyne} = 1\,\text{g}\cdot\text{cm}/\text{s}^2 = 10^{-5}\,\text{N}, \qquad 1\,\text{erg} = 1\,\text{g}\cdot\text{cm}^2/\text{s}^2 = 10^{-7}\,\text{J}.$$

17.1.6 What follows from the Gaussian choice

Putting $\varepsilon_0 = 1/4\pi$ into Coulomb's law collapses it to

$$\boxed{F = \frac{q_1 q_2}{r^2}}, \qquad \text{(Coulomb's law, Gaussian units)}$$

with force in dynes and distance in centimeters. Reading off the units gives dyne = [charge]^2/cm^2, so [charge] = cm $\cdot$ dyne$^{1/2}$. That unit has its own name — in fact three names, since there is no general agreement: the **statcoulomb**, the **franklin**, or the **electrostatic unit of charge** (esu). They are interchangeable; I will use esu. One esu is the charge that would exert a force of one dyne on an equal charge one centimeter away (Figure 17.3), and

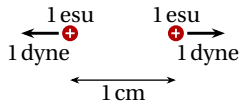


Figure 17.3 The definition of the esu, or statcoulomb: two charges of one esu, one centimeter apart, repel each other with a force of one dyne.

$$1\,\text{esu} = 3.3356 \times 10^{-10}\,\text{C}.$$

Current is measured in esu per second, sometimes called the *statampere* — and, as the trade-off for making charges come out at a usable size, one esu per second is a very small current indeed.

Since ε_0 was fixed, μ_0 is now the measured one. But you never write it either, because Maxwell's equations predict waves traveling at $1/\sqrt{\varepsilon_0\mu_0}$, which is the measured speed of light. Setting that equal to c with $\varepsilon_0 = 1/4\pi$ gives

$$\boxed{\mu_0 = \frac{4\pi}{c^2}}. \qquad \text{(Gaussian units)}$$

The unit of electric field comes from $\mathbf{F} = q\mathbf{E}$, so it is the dyne per esu. It has no name of its own. The Gaussian unit of potential, the **statvolt**, is one erg per esu, so a dyne per esu is a statvolt per centimeter, and that is how the unit of electric field is usually written.

17.1.7 Why E and B end up with the same units

Here is the part that is genuinely illuminating rather than merely conventional.

In SI, the units of **B** are forced by the Lorentz force law: the two terms inside the parentheses must match, so tesla = (N/C)/(m/s). But there is no law of nature saying the magnetic force must be written $q\mathbf{v}\times\mathbf{B}$. It could equally well be written

$$\mathbf{F} = q\left(\mathbf{E} + \frac{\mathbf{v}}{c}\times\mathbf{B}\right), \qquad \text{(Lorentz force, Gaussian units)}$$

which contains exactly the same physics. And now look at the units: inside the parentheses, v and c cancel. So in Gaussian units **E** and **B** have *the same units*.

The unit of magnetic field is called the **gauss**, and a gauss is the same unit as a statvolt per centimeter, although by custom the name gauss is used only for magnetic fields. For reference,

$$1\,\text{gauss} = 10^{-4}\,\text{tesla}, \qquad 1\,\text{statV/cm} = 2.998\times 10^{4}\,\text{V/m}.$$

The two SI values differ by a factor of c (2.998×10^4 V/m is c times 10^{-4} T), because in SI the magnetic field is the Gaussian one divided by c; that is the third translation rule below.

17.1.8 Three translation rules

There are subtleties I am passing over, particularly once materials are involved, but these three rules will get you from SI to Gaussian about ninety percent of the time. The first is the definition, the second follows from $c = 1/\sqrt{\varepsilon_0\mu_0}$, and the third from the rewritten Lorentz force:

! **SI to Gaussian.** Replace ε_0 by $1/4\pi$, μ_0 by $4\pi/c^2$, and $\mathbf{B}$ by $\mathbf{B}/c$.

1. Wherever you see ε_0, replace it with $1/4\pi$.
2. Wherever you see μ_0, replace it with $4\pi/c^2$.
3. Wherever you see $\mathbf{B}$, replace it with $\mathbf{B}/c$.

17.2 Maxwell's equations in Gaussian units

Applying those three rules to Maxwell's equations (Section 2.1) gives the Gaussian forms in Table 17.1.

SI	GAUSSIAN
$\nabla\cdot\mathbf{E} = \dfrac{\rho}{\varepsilon_0}$	$\boxed{\nabla\cdot\mathbf{E} = 4\pi\rho}$
$\nabla\times\mathbf{E} = -\dfrac{\partial\mathbf{B}}{\partial t}$	$\boxed{\nabla\times\mathbf{E} = -\dfrac{1}{c}\dfrac{\partial\mathbf{B}}{\partial t}}$
$\nabla\cdot\mathbf{B} = 0$	$\boxed{\nabla\cdot\mathbf{B} = 0}$
$\nabla\times\mathbf{B} = \mu_0\mathbf{J} + \mu_0\varepsilon_0\dfrac{\partial\mathbf{E}}{\partial t}$	$\boxed{\nabla\times\mathbf{B} = \dfrac{4\pi}{c}\mathbf{J} + \dfrac{1}{c}\dfrac{\partial\mathbf{E}}{\partial t}}$

Table 17.1 Maxwell's equations in SI and in Gaussian units.

Notice that c now appears explicitly in the equations themselves. In SI it is hidden inside $\mu_0\varepsilon_0$ and only emerges when you derive the wave equation; in Gaussian units the fact that electromagnetic waves travel at c is written on the face of the theory.

17.2.1 Another example: energy stored in the fields

The same translation works on anything. Take the energies stored in the fields, (5.23) and (12.35):

$$U_{\text{electric}} = \frac{\varepsilon_0}{2}\int E^2\,d\tau \quad\longrightarrow\quad \boxed{U_{\text{electric}} = \frac{1}{8\pi}\int E^2\,d\tau},$$

$$U_{\text{magnetic}} = \frac{1}{2\mu_0}\int B^2\,d\tau \quad\longrightarrow\quad \boxed{U_{\text{magnetic}} = \frac{1}{8\pi}\int B^2\,d\tau},$$

where the magnetic case uses rules 2 and 3 together: $\mu_0 \to 4\pi/c^2$ and $B \to B/c$, and the two factors of c^2 cancel.

The two expressions are identical. The electric and magnetic fields contribute *equally* to the stored energy — an electric field of one statvolt per centimeter holds exactly as much energy as a magnetic field of one gauss in the same volume. That symmetry is entirely invisible in the SI forms, where one carries $\varepsilon_0/2$ and the other $1/2\mu_0$.

17.2.2 Why anyone still uses Gaussian units

SI has taken over completely at the high-school and undergraduate level, and this book uses it throughout. But Gaussian units persist in research and in graduate courses, for reasons that the above should now make concrete:

1. Coulomb's law loses its $1/4\pi\varepsilon_0$ out front.
2. There is no need for the strange defined constant $\mu_0 = 4\pi \times 10^{-7}\,\text{N/A}^2$, which surely puzzles every introductory student who meets it without explanation.
3. The equations carry fewer symbols — neither ε_0 nor μ_0 appears anywhere.
4. That electromagnetic waves travel at c is built into the equations rather than derived from them.
5. **E** and **B** share units.
6. Symmetries between **E** and **B** become visible that SI hides, the stored-energy expressions being the clearest case.

Atomic physics	$4\pi\varepsilon_0 = \hbar = m_e = e = 1$
Particle physics	$\hbar = c = 1;\ \varepsilon_0 = 1$
Relativity	$G = c = 1$
Spectroscopy	energy in cm^{-1}
Plasma physics	temperature in eV
Astronomy	parsecs, solar masses

Table 17.2 Some unit habits of research fields: the constants each sets equal to one, or the units in which it quotes quantities (Sections 17.3.1 and 17.3.2).

17.3 Atomic units and other systems used in research

Gaussian units are not the only alternative to SI. In research, people use whatever units make the equations of their own field simplest and the numbers in their calculations a convenient size, and there are many such systems. This section describes the one you are most likely to meet if you do computations on atoms, molecules, or solids, and then a few others (Table 17.2).

17.3.1 Hartree atomic units

Atomic units were introduced by Douglas Hartree in 1928, for his calculations of the structure of atoms. They make the Gaussian choice for electrostatics, $4\pi\varepsilon_0 = 1$, and then go further: the reduced Planck constant, the mass of the electron, and the elementary charge are all set equal to one as well,

$$\boxed{\hbar = m_e = e = 4\pi\varepsilon_0 = 1}. \qquad \text{(Hartree atomic units)}$$

With those four constants fixed, every other quantity has a natural unit built out of them. The unit of length is the **Bohr radius**,

$$a_0 = \frac{4\pi\varepsilon_0\hbar^2}{m_e e^2} = 5.29177\times10^{-11}\,\mathrm{m} \approx 0.529\,\text{Å},$$

the radius of the smallest electron orbit in the Bohr model of hydrogen (Figure 17.4). The unit of energy is the **hartree**,

$$E_h = \frac{\hbar^2}{m_e a_0^2} = \frac{m_e e^4}{(4\pi\varepsilon_0)^2\hbar^2} = 27.2114\,\mathrm{eV},$$

which is also the potential energy of two electrons one bohr apart, $e^2/4\pi\varepsilon_0 a_0$. In atomic units Coulomb's law has its Gaussian form, $F = q_1 q_2 / r^2$, with charges measured in multiples of e and distances in bohrs.

The payoff is in the equations. The Schrödinger equation for the electron in a hydrogen atom, which in SI units reads

$$-\frac{\hbar^2}{2m_e}\nabla^2\psi - \frac{e^2}{4\pi\varepsilon_0 r}\psi = E\psi,$$

becomes, in atomic units,

$$-\frac{1}{2}\nabla^2\psi - \frac{1}{r}\psi = E\psi.$$

Its ground state is $\psi = e^{-r}/\sqrt{\pi}$, with $E = -1/2$: the electron's energy is $-\frac{1}{2}$ hartree, or $-13.6\,\mathrm{eV}$, and its most probable distance from the proton is one bohr.[3] (Exercise P17.4 checks this.)

This is why atomic units are widely used in computational chemistry and computational condensed matter physics. The equations carry no constants, so there are none to mistype; the numbers that come out are of order one, rather than of order $10^{-18}\,\mathrm{J}$ and $10^{-11}\,\mathrm{m}$; and converting a result is a single multiplication at the end. A computed energy of -0.5 means $-0.5\,E_h = -13.6\,\mathrm{eV}$, and a computed distance of 2.0 means $2.0\,a_0 = 1.06\,\text{Å}$.

The one constant you cannot choose is the speed of light. With $\hbar$, m_e, e, and $4\pi\varepsilon_0$ already fixed, there is nothing left to adjust, and c comes out as

$$c = \frac{1}{\alpha} \approx 137.036, \qquad \text{(atomic units)}$$

[3] Strictly, the mass in the hydrogen equation is the reduced mass of the electron and proton, which changes these numbers by about 0.05 percent.



Douglas Hartree (1897–1958, English) Douglas Rayner Hartree was born in Cambridge. He went from Bedales School to St John's College, Cambridge, in 1915, and after a year he left to work on anti-aircraft gunnery with A. V. Hill's group. He received his PhD in 1926, and in 1928 he introduced the self-consistent field method, in which each electron moves in the average field of the nucleus and the other electrons, defining atomic units for the purpose. He moved to Manchester in 1929, as professor of applied mathematics and from 1937 of theoretical physics, and he was elected a Fellow of the Royal Society in 1932. In 1934, with Arthur Porter, he built a differential analyser out of Meccano. In 1946 he advised on the use of ENIAC, one of the first electronic computers, and became Plummer Professor of Mathematical Physics at Cambridge, a post he held until his death. (Wikipedia)

☞ *Caution*: In atomic units the speed of light is *not* one; it is $1/\alpha \approx 137$. The system that sets $c = 1$ is the natural units of particle physics, Section 17.3.2, which measures lengths in inverse energies rather than in bohrs.

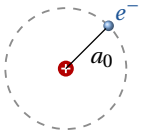


Figure 17.4 The smallest electron orbit of the Bohr model of hydrogen, whose radius is the Bohr radius a_0, the atomic unit of length.

where

$$\alpha = \frac{e^2}{4\pi\varepsilon_0\hbar c} \approx \frac{1}{137.036}$$

is the dimensionless **fine-structure constant**. The atomic unit of velocity is the speed of the electron in the first Bohr orbit, so $c \approx 137$ says that this electron moves at about $c/137$, which is one reason relativistic effects in hydrogen are small. Table 17.3 gives the atomic units of the quantities you are most likely to need.

Quantity	Atomic unit	SI value
length	a_0 (bohr)	$5.29177 \times 10^{-11}\,\mathrm{m}$
energy	E_h (hartree)	$4.35974 \times 10^{-18}\,\mathrm{J} = 27.2114\,\mathrm{eV}$
time	$\hbar/E_h$	$2.41888 \times 10^{-17}\,\mathrm{s}$
velocity	$a_0 E_h/\hbar = \alpha c$	$2.18769 \times 10^{6}\,\mathrm{m/s}$
electric field	$E_h/(e\,a_0)$	$5.14221 \times 10^{11}\,\mathrm{V/m}$

Table 17.3 Hartree atomic units of some quantities, and their values in SI units.

The atomic unit of electric field is the field of a proton at a distance of one bohr. Laboratory fields, apart from those of intense lasers, are many orders of magnitude smaller, which is why an applied field only slightly perturbs an atom.

A variant, **Rydberg atomic units**, sets $\hbar = 2m_e = e^2/2 = 1$ (again with $4\pi\varepsilon_0 = 1$), so that the unit of energy is the rydberg, 13.6057 eV, half a hartree; lengths are still in bohrs. Computer programs differ in their choices. Quantum ESPRESSO, a widely used package for calculations on solids, works in Rydberg atomic units; the quantum chemistry program Gaussian reports energies in hartrees; VASP, another solid-state package, uses electron volts and ångströms. Before you compare numbers from two programs, or from a program and a paper, check which units each one uses.

17.3.2 Natural units, and others

Particle physicists set $\hbar = c = 1$. Mass, energy, and momentum then all have the same units, as do length and time, and everything is measured in powers of an energy, usually MeV or GeV. The electron's mass is 0.511 MeV; a length is an inverse energy, converted to meters with

$$\hbar c = 197.327\,\mathrm{MeV}\cdot\mathrm{fm},$$

so that $1\,\mathrm{GeV}^{-1}$ is a length of 0.197 fm. For electromagnetism they usually also set $\varepsilon_0 = 1$ (and so $\mu_0 = 1/c^2 = 1$), a choice called **Heaviside–Lorentz units**. Maxwell's equations then read

$$\nabla\cdot\mathbf{E} = \rho, \qquad \nabla\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t}, \qquad \nabla\cdot\mathbf{B} = 0, \qquad \nabla\times\mathbf{B} = \mathbf{J} + \frac{\partial\mathbf{E}}{\partial t},$$

with no 4π anywhere; the 4π shows up in Coulomb's law instead, $F = q_1 q_2/4\pi r^2$, the opposite of the Gaussian trade. The elementary charge is the pure number $e = \sqrt{4\pi\alpha} \approx 0.303$.

Other fields have their own habits. In general relativity it is common to set $G = c = 1$, so that masses and times are measured in meters: the mass of the Sun is $GM_\odot/c^2 = 1.48\,\text{km}$. Spectroscopists quote photon energies as wavenumbers in cm^{-1} ($1\,\text{cm}^{-1}$ corresponds to $1.24 \times 10^{-4}\,\text{eV}$), plasma physicists quote temperatures in electron volts ($1\,\text{eV}$ corresponds to $11\,600\,\text{K}$), and astronomers measure distances in parsecs and masses in solar masses.

None of these systems contains different physics. Each is SI with some constants set equal to one and the units rescaled to match, and you can always get back to SI by putting the constants back in where the units require them. What you do need, before using a number from a paper or a program, is to know which system it was written in.

17.4 The 2019 change to SI units

Sections 17.1 and 17.2 describe SI as it stood until 2019, and that is deliberate: many textbooks still quote $\mu_0 = 4\pi \times 10^{-7}$, and the pre-2019 story is the cleaner illustration of the vicious cycle. But the system has since changed, and it changed in a way that bears directly on this chapter's argument.

In 2019 four of the seven SI base units were redefined — the kilogram, the ampere, the kelvin and the mole — so that Planck's constant, the elementary charge, the Boltzmann constant and the Avogadro constant all take *defined* rather than measured values. The elementary charge in particular is now exactly

$$e \equiv 1.602176634 \times 10^{-19}\,\text{C}.$$

Read that against the vicious cycle and you can see what happened. The suspect who flipped has changed. It is no longer μ_0 that is fixed by decree; it is effectively the coulomb itself, defined as exactly $1/(1.602176634 \times 10^{-19})$ elementary charges. An ampere is correspondingly a current of exactly 10^{19} elementary charges per 1.602176634 seconds.

And so μ_0 has changed sides. It is now measured, at

$$\mu_0 = 1.25663706127(20) \times 10^{-6}\,\text{N/A}^2,$$

which agrees with $4\pi \times 10^{-7}$ to about one part in 10^{10} — far too small a difference to matter for anything in this book, which is why the old value is still quoted everywhere and why you should not worry about it.

! Since 2019, μ_0 is a *measured* quantity in SI, not a defined one. Its value is still almost exactly $4\pi \times 10^{-7}$, which is why nearly every textbook still prints that number.

It is worth being clear about what did *not* change. Not one equation in this chapter, or anywhere else in the book, is altered by the 2019 revision. The numbers you put into them are the same to every precision you will ever use. What changed is only the logical structure underneath — which of the seven suspects was made to flip first. That the choice is available at all, and that physics is indifferent to it, is the real lesson of this chapter.

17.5 Summary

Almost everything you will meet in the rest of this book is written in SI, and that is how it should be. But the Gaussian system has not gone away, and when you open a paper written in it, these are the pieces you need.

The three translation rules. Apply them to any SI equation and you get its Gaussian form.

$\varepsilon_0 \to \dfrac{1}{4\pi}$	the definition: Coulomb's law is written with no constant out front
$\mu_0 \to \dfrac{4\pi}{c^2}$	follows, because $c = 1/\sqrt{\varepsilon_0\mu_0}$ must still hold
$\mathbf{B} \to \dfrac{\mathbf{B}}{c}$	follows from writing the Lorentz force as $\mathbf{F} = q\left(\mathbf{E} + \frac{\mathbf{v}}{c} \times \mathbf{B}\right)$

What the rules buy you. Coulomb's law loses its constant, c appears on the face of Maxwell's equations instead of hiding inside $\mu_0\varepsilon_0$, and $\mathbf{E}$ and $\mathbf{B}$ end up in the same units — which is the system's real argument for itself, since the two fields are one object seen from different frames.

Conversions worth having.

$1\,\text{esu} = 3.3356 \times 10^{-10}\,\text{C}$	charge (also called statcoulomb or franklin)
$1\,\text{gauss} = 10^{-4}\,\text{T}$	magnetic field
$1\,\text{statV/cm} = 2.998 \times 10^{4}\,\text{V/m}$	electric field; the same unit as the gauss
$1\,\text{dyne} = 10^{-5}\,\text{N}, \quad 1\,\text{erg} = 10^{-7}\,\text{J}$	force and energy, from cgs

Units in research. Other systems set more constants equal to one, to make the equations of a particular field simple.

$\hbar = m_e = e = 4\pi\varepsilon_0 = 1$	Hartree atomic units: lengths in bohrs ($a_0 = 0.529\,\text{Å}$), energies in hartrees ($E_h = 27.2\,\text{eV}$), and $c = 1/\alpha \approx 137$
$\hbar = c = \varepsilon_0 = 1$	natural (Heaviside–Lorentz) units of particle physics: everything in powers of an energy, with $\hbar c = 197.3\,\text{MeV}\cdot\text{fm}$

Finally, remember that since 2019 the SI is defined by fixing seven constants rather than by any physical artifact, so μ_0 is now a measured quantity rather than an exact one. The change is far below experimental relevance for anything in this book, but the logic of which quantities are definitions and which are measurements has been inverted, and that is worth knowing.

Exercises

Exercises for 17.1 SI versus Gaussian units

Warm-up questions Read Section 17.1, then answer before class

W17.1 Suppose the coulomb were redefined to be twice as large as it is now. What would the measured value of ε_0 become? What does your answer say about measuring ε_0 with Coulomb's law before a unit of charge has been chosen?

W17.2 Suppose the volt, rather than μ_0, were the quantity fixed by decree. Outline how you would then obtain the coulomb, the ampere, and μ_0, naming the law or definition that each step uses.

W17.3 In Gaussian units a gauss is the same unit as a statvolt per centimeter. Does an electric field of 1 statV/cm push on a charge as hard as a magnetic field of 1 gauss does? Explain with the Gaussian form of the Lorentz force.

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P17.1 The Poynting vector, which you will meet properly in Section 20.2.2, is the energy flux carried by an electromagnetic field. In SI units it is

$$\mathbf{S} = \frac{1}{\mu_0}\,\mathbf{E}\times\mathbf{B}.$$

Use the three translation rules to write it in Gaussian units.

P17.2 One esu is defined as the charge that exerts a force of one dyne on an equal charge one centimeter away.

(a) Verify, directly from Coulomb's law in Gaussian units, that this is consistent with the definition.

(b) Now check it the other way around. Using $1\,\text{esu} = 3.3356\times10^{-10}\,\text{C}$ and SI Coulomb's law, compute the force in newtons between two 1-esu charges 1 cm apart, and confirm it comes to one dyne.

Exercises for 17.2 Maxwell's equations in Gaussian units

Warm-up questions Read Section 17.2, then answer before class

W17.4 Use the three translation rules to explain the factors in the Gaussian Ampère–Maxwell law of Table 17.1: why does $\mathbf{J}$ get $4\pi/c$, but $\partial\mathbf{E}/\partial t$ only $1/c$?

W17.5 This section says that an electric field of one statvolt per centimeter holds exactly as much energy as a magnetic field of one gauss. Check the claim in SI, where one statvolt per centimeter is $2.998\times10^4\,\mathrm{V/m}$ and one gauss is $10^{-4}\,\mathrm{T}$.

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P17.3 The force per unit length between two long parallel wires carrying currents I_1 and I_2 a distance d apart is $\mu_0 I_1 I_2/2\pi d$ in SI units, (10.4). Convert it to Gaussian units. Only one of the three translation rules is needed — say which, and explain why the other two do not come into it.

Exercises for 17.3 Atomic units and other systems used in research

Warm-up questions Read Section 17.3, then answer before class

W17.6 Two programs both report an energy of -1.000: Quantum ESPRESSO, which works in Rydberg atomic units, and the quantum chemistry program Gaussian, which reports hartrees. Are the two energies the same? Give each in electron volts.

W17.7 Heaviside–Lorentz units take the 4π out of Maxwell's equations and put it into Coulomb's law; Gaussian units do the opposite. Why can no choice of units remove it from both? (Apply Gauss's law to a sphere around a point charge.)

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P17.4 In Hartree atomic units the Schrödinger equation for the electron in a hydrogen atom is

$$-\frac{1}{2}\nabla^2\psi-\frac{1}{r}\psi=E\psi.$$

(a) Show that $\psi=e^{-r}/\sqrt{\pi}$ satisfies it, and find E. (Use the Laplacian in spherical coordinates, (1.32).)

(b) Show that this ψ is normalized, $\int|\psi|^2\,d\tau=1$ over all space, and that the probability density for finding the electron at a distance r from the proton, $4\pi r^2|\psi|^2$, is largest at $r=1$.

(c) Convert E to electron volts and the most probable distance to meters.

P17.5 This exercise connects the atomic and natural units of Section 17.3.

(a) Show that the Bohr radius can be written $a_0 = \hbar/(\alpha\, m_e c)$.

(b) In natural units the mass of the electron is 0.511 MeV. Find $\hbar/m_e c$, the reduced Compton wavelength of the electron, in femtometers, using $\hbar c = 197.327\,\text{MeV}\cdot\text{fm}$; then find a_0 in ångströms.

(c) Show that in Hartree atomic units the speed of light is $1/\alpha$.

Exercises for 17.4 The 2019 change to SI units

Warm-up questions Read Section 17.4, then answer before class

W17.8 Could the 2019 SI have fixed the elementary charge exactly and kept $\mu_0 = 4\pi \times 10^{-7}\,\text{N/A}^2$ exact as well? Explain with the vicious cycle of Section 17.1.

W17.9 A classmate says that, since μ_0 is no longer exactly $4\pi \times 10^{-7}\,\text{N/A}^2$, the answers to this book's exercises are now slightly wrong and should be recomputed. Do you agree?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

Chapter 18

Mathematics Review 3: Complex Numbers

Much of the rest of this book is about fields and currents that oscillate: the alternating currents of circuits (Chapter 19), electromagnetic waves (Chapters 21 to 23), and the radiation of oscillating charges and currents (Chapters 27 and 28). Complex numbers are the tool for all of them. A sine and a cosine turn into each other when they are differentiated, and adding or multiplying them takes trigonometric identities. A complex exponential stays a complex exponential when it is differentiated, and the derivative is just a multiplication.

This chapter reviews complex numbers and then develops the tools the later chapters use: complex exponentials for oscillations and waves, complex wave numbers, products of oscillating quantities, and the two sign conventions for writing oscillations, both of which this book uses.

18.1 Complex numbers

18.1.1 Rectangular and polar forms

A **complex number** $z = x + iy$, with $i = \sqrt{-1}$, has two parts: its **real part** $x = \mathrm{Re}[z]$ and its **imaginary part** $y = \mathrm{Im}[z]$, both of them real numbers. In the **complex plane** the number z is the point (x, y), with the real part measured along the horizontal axis and the imaginary part along the vertical axis (Figure 18.1). So a complex number can be written in rectangular or polar form, just like coordinates in the xy plane. The **rectangular form** gives the point by its coordinates (x, y). The **polar form** (A, θ) gives it by its distance A from the origin, which is the **magnitude** $|z|$ of z, and by the angle θ between the positive real axis and the line from the origin to the point, which is the **phase** of z (also called its argument):

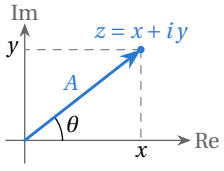


Figure 18.1 The complex number $z = x + iy$ in the complex plane, with its magnitude A and its phase θ.

$$A = |z| = \sqrt{x^2 + y^2}, \qquad x = A\cos\theta, \qquad y = A\sin\theta. \tag{18.1}$$

The rectangular form is most useful for adding and subtracting complex numbers: the real parts add, and the imaginary parts add, as the components of vectors do.



Caspar Wessel (1745–1818, Danish–Norwegian) Wessel was born in Vestby, Norway, then part of Denmark–Norway. From 1764 to 1805 he worked on the Royal Danish Academy's survey of Denmark, Schleswig, Holstein and Oldenburg, from 1798 as Royal Inspector of Surveying. His paper of 1797, "Om Directionens analytiske Betegning" (On the analytical representation of direction), printed in 1799, treated line segments with both a length and a direction as numbers, added by joining them so that one begins where the other ends. No portrait of Wessel survives; the picture is a map of southeastern Zealand that he drew in 1770. (Wikipedia)



Jean-Robert Argand (1768–1822, Genevan) Little is known about Argand's life, and historians are not certain that the records of birth and death usually given for him are his. His *Essai sur une manière de représenter les quantités imaginaires dans les constructions géométriques* (1806) did not carry his name. Legendre sent a copy to François Français, whose brother Jacques found it among his papers after his death and published its ideas in 1813, asking the unknown author to come forward. Argand did, and published a version of his essay in the same journal. No portrait of Argand is known; the picture shows a copy of his *Essai*, inscribed to Gergonne. (Wikipedia)

The polar form is most useful for multiplying and dividing complex numbers, as the next subsection shows.

18.1.2 Euler's formula

The polar form becomes most useful when it is written with an exponential. The Taylor series of the exponential function, $e^u = 1 + u + u^2/2! + u^3/3! + \cdots$, holds for complex u as well as for real u. Put $u = i\theta$, with θ real, and use $i^2 = -1$, $i^3 = -i$, $i^4 = 1$, and so on. The even powers of $i\theta$ are real and the odd powers are imaginary, so

$$e^{i\theta} = \left(1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \cdots\right) + i\left(\theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \cdots\right).$$

The two series are the Taylor series of the cosine and the sine, and so

$$\boxed{e^{i\theta} = \cos\theta + i\sin\theta}. \tag{18.2}$$

This is **Euler's formula**. By (18.1), $z = A\cos\theta + iA\sin\theta = Ae^{i\theta}$: a complex number of magnitude A and phase θ is A times a complex exponential. For example, consider the complex number $3+4i$. It is $(3,4)$ in rectangular form, $(5, 53.13°)$ in polar form, and $5e^{i\,53.13°}$ or $5e^{0.9273i}$ in complex exponential form, since $53.13° = 0.9273$ rad. The number $e^{i\theta}$ has the magnitude 1, so it lies on the circle of radius 1 around the origin; as θ increases, it moves counterclockwise around that circle, and it comes back to the same point each time θ increases by 2π. Two values worth remembering are $e^{i\pi/2} = i$ and $e^{i\pi} = -1$.

By the rules of exponents, when you multiply two complex numbers in polar form you multiply their magnitudes and add their phases, and when you divide them you divide the magnitudes and subtract the phases:

$$A_1 e^{i\theta_1}\, A_2 e^{i\theta_2} = A_1 A_2\, e^{i(\theta_1+\theta_2)}, \qquad \frac{A_1 e^{i\theta_1}}{A_2 e^{i\theta_2}} = \frac{A_1}{A_2}\, e^{i(\theta_1-\theta_2)}. \tag{18.3}$$

In the complex plane, multiplying z by $Ae^{i\phi}$ stretches the line from the origin to z by the factor A and turns it counterclockwise through the angle ϕ. Multiplying by $i = e^{i\pi/2}$ turns it through 90°, and multiplying by $-1 = e^{i\pi}$ turns it through 180°.

I like to write the polar form using this notation: $A\angle\theta$, which is also the angle notation of electrical engineers. The "$\angle$" symbol is read as "at an angle of." Thus you can write

$$(3+4i)(5+12i) = (5\angle 53.13°)(13\angle 67.38°) = 65\angle 120.51°,$$

since $65 = 5\times 13$ and $120.51° = 53.13° + 67.38°$. Multiplying out the rectangular forms gives the same number, $15 + 36i + 20i + 48i^2 = -33 + 56i$, which has the magnitude $\sqrt{33^2 + 56^2} = 65$ and lies at 120.51°.

Leonhard Euler, whose portrait is in Section 1.1.5, published (18.2) in 1748, in his *Introductio in analysin infinitorum*. The picture of complex numbers as points in a plane came half a century later. Caspar Wessel, a surveyor, presented it

to the Royal Danish Academy of Sciences and Letters in 1797, in a paper written in Danish, and the paper went unnoticed for nearly a century. Jean-Robert Argand, an amateur mathematician, found the same picture on his own and set it out in an essay of 1806 that did not carry his name. The complex plane is often called the Argand plane, and a drawing of complex numbers in it an Argand diagram.

18.1.3 The complex conjugate

The **complex conjugate** of $z = x + iy = Ae^{i\theta}$ is

$$z^* = x - iy = Ae^{-i\theta}. \tag{18.4}$$

In the complex plane, z^* is the mirror image of z across the real axis (Figure 18.2). To take the complex conjugate of any expression built from real numbers and i by adding, subtracting, multiplying, dividing and raising to powers, just switch every i to $-i$. For example, the complex conjugate of $(1+3i)/(4-5i)$ is $(1-3i)/(4+5i)$. The rule works because i and $-i$ are equally good square roots of -1 (Section 18.5).

The product of a complex number and its conjugate is real, and it is the square of the magnitude:

$$\begin{aligned} zz^* &= (x+iy)(x-iy) \\ &= x^2 + y^2 \\ &= |z|^2. \end{aligned} \tag{18.5}$$

This is the way to divide in rectangular form: multiply the numerator and the denominator by the complex conjugate of the denominator, which makes the denominator real. The conjugate also gives the real and imaginary parts:

$$\mathrm{Re}[z] = \frac{z+z^*}{2}, \qquad \mathrm{Im}[z] = \frac{z-z^*}{2i}. \tag{18.6}$$

When you convert from rectangular to polar form, always draw a picture. The phase is not simply $\arctan(y/x)$: the arctangent of a calculator gives an angle between $-90°$ and $90°$, and y/x is the same for (x, y) and $(-x, -y)$. For example, $2-i$ and $-2+i$ both have $y/x = -1/2$, and $\arctan(-1/2) = -26.57°$. That is the phase of $2-i$, but $-2+i$ is in the second quadrant, at $-26.57° + 180° = 153.43°$ (Figure 18.3). Whenever $x < 0$, add $180°$ to the calculator's answer (or subtract it, to keep the phase between $-180°$ and $180°$).

18.1.4 Roots

The equation $z^2 = 1$ has two solutions, $z = \pm 1$, and so does $z^2 = -1$: $z = \pm i$. In general a nonzero complex number has n different nth **roots**. To find them, write the number in polar form, remembering that adding a multiple of 2π to the phase changes nothing, $Ae^{i\theta} = Ae^{i(\theta+2\pi m)}$ for any integer m. Then

$$\left(Ae^{i\theta}\right)^{1/n} = A^{1/n} e^{i(\theta+2\pi m)/n}, \qquad m = 0, 1, \ldots, n-1. \tag{18.7}$$

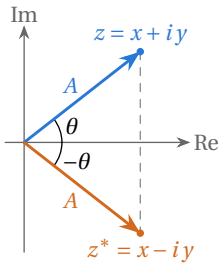


Figure 18.2 A complex number $z = x + iy$ and its complex conjugate $z^* = x - iy$, the mirror image of z across the real axis. The two have the same magnitude A and opposite phases.

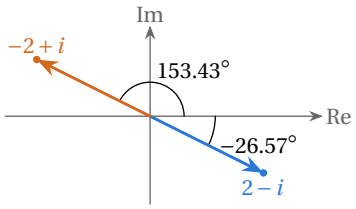


Figure 18.3 The numbers $2-i$ and $-2+i$ have the same ratio $y/x = -1/2$, so the arctangent gives $-26.57°$ for both. Only $2-i$ is at that angle; $-2+i$ is at $153.43°$.

The n roots have the same magnitude, and they are spaced $360°/n$ apart around a circle. For example, the three cube roots of 1 are 1, $e^{2\pi i/3} = -\frac{1}{2} + i\frac{\sqrt{3}}{2}$ and $e^{4\pi i/3} = -\frac{1}{2} - i\frac{\sqrt{3}}{2}$.

A calculator or a computer gives only one of the roots, the **principal root**: it takes the phase in the range $-180° < \theta \le 180°$ and divides it by n. So the principal square root of a complex number has a positive real part, or lies on the positive imaginary axis; the principal square root of -4 is $2i$. Which root a calculation needs is a question of physics, not of convention, as Section 18.3 shows.

18.2 Oscillations and waves as complex exponentials

18.2.1 Phase and the complex amplitude

Suppose that something oscillates at the angular frequency ω: a displacement, a voltage, a current, or one component of a field at a point. Its time dependence can be written

$$f(t) = A\cos(\omega t + \phi), \tag{18.8}$$

where A is its amplitude and ϕ its phase. With $\phi = 0$ the oscillation reaches its peak at $t = 0$. With a positive ϕ it reaches its peak earlier, at $t = -\phi/\omega$: a positive ϕ shifts the graph of $f(t)$ to the left along the time axis, and the oscillation leads $A\cos\omega t$ by the angle ϕ (Figure 18.4). A negative ϕ shifts the graph to the right. The graph of any function shifts in the same way. Replacing x by $x+3$ shifts a graph to the left by 3, and replacing x by $x-3$ shifts it to the right by 3: $(x+3)^2$ has its minimum at $x = -3$, and $(x-3)^2$ has its minimum at $x = 3$ (Figure 18.5).

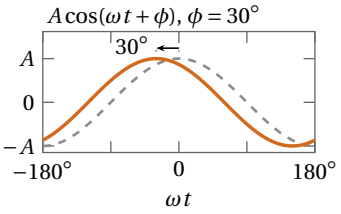


Figure 18.4 $A\cos(\omega t+\phi)$ with $\phi = 30°$ (orange) and $A\cos\omega t$ (dashed). A positive ϕ shifts the curve to the left, to earlier times: the oscillation leads.

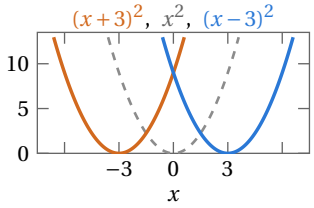


Figure 18.5 The graphs of $(x+3)^2$, x^2 (dashed) and $(x-3)^2$. Adding 3 to x shifts the graph to the left, and subtracting 3 shifts it to the right.

There is an immediate complication: some people define the phase the opposite way, and start with

$$f(t) = A\cos(\omega t - \phi). \tag{18.9}$$

With this definition a positive ϕ shifts the graph to the right. The oscillation reaches its peak later, at $t = \phi/\omega$, and it lags $A\cos\omega t$ by the angle ϕ (Figure 18.6). In this book the two choices are called the **phase-lead convention**, (18.8), in which a positive phase is a lead, and the **time-delay convention**, (18.9), in which a positive phase is a delay. Be careful to find out which convention a source is using. This can be particularly confusing, because in my experience it is roughly 50–50 as to which convention is used.

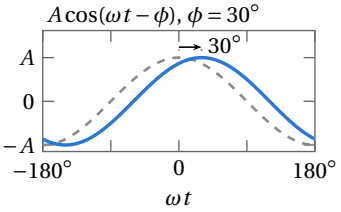


Figure 18.6 $A\cos(\omega t-\phi)$ with $\phi = 30°$ (blue) and $A\cos\omega t$ (dashed). With this convention a positive ϕ shifts the curve to the right, to later times: the oscillation lags.

The two conventions tend to be used in different fields. Books on circuits usually use the phase-lead convention, and so, most likely, did the treatment of alternating currents in your introductory electricity and magnetism course. Physics books usually write a wave that travels toward $+z$ as $A\cos(kz-\omega t+\phi)$, which at $z = 0$ is $A\cos(\omega t-\phi)$: the time-delay convention. This book follows the same pattern. For now it uses the phase-lead convention (18.8), in the rest of this section and in Sections 18.2.2 and 18.2.3, and it will continue to use it for the circuits of Chapter 19. For electromagnetic fields and waves it switches to the time-delay convention, starting in Section 18.2.4 and continuing in Chapter 21 and beyond. The text says clearly wherever the switch happens.

Now we turn to complex numbers. By Euler's formula, $\cos(\omega t+\phi)$ is the real part of $e^{i(\omega t+\phi)}$, so the oscillation (18.8) can be represented in several equivalent ways:

$$\begin{aligned} A\cos(\omega t+\phi) &\longleftrightarrow Ae^{i(\omega t+\phi)} \\ &\longleftrightarrow Ae^{i\phi} \qquad \text{(with an assumed } e^{i\omega t} \text{ time dependence)} \\ &\longleftrightarrow A\angle\phi \longleftrightarrow \tilde{A}. \end{aligned}$$

The complex number $\tilde{A} = Ae^{i\phi} = A\angle\phi$ is the **complex amplitude** of the oscillation. Its magnitude is the actual amplitude A, and its phase is the actual phase ϕ of the oscillation $A\cos(\omega t+\phi)$. In this book a tilde marks a complex quantity. The factor $e^{i\omega t}$ is the same for every quantity that oscillates at the frequency ω, so it is usually not written but understood: that is what an assumed $e^{i\omega t}$ time dependence means. To get the real oscillation back, put the factor back and take the real part:

$$\boxed{f(t) = \operatorname{Re}\left[\tilde{A}\,e^{i\omega t}\right], \qquad \tilde{A} = Ae^{i\phi}}. \tag{18.10}$$

☞ **Caution: two conventions.** In $A\cos(\omega t+\phi)$ a positive phase ϕ is a lead; in $A\cos(\omega t-\phi)$ it is a delay. Before you use a formula from another source, find out which convention it uses (Section 18.5).

The words for the parts of a complex number vary. A is called the amplitude, magnitude, or modulus of the complex number $Ae^{i\phi}$. ϕ is called the phase angle, phase, angle, argument, or amplitude of the complex number, and $e^{i\phi}$ is called its phase factor. Yes, unfortunately the word "amplitude" can at times refer to either A or ϕ! I myself will reserve "amplitude" for A, though.

It is understood that this is just a temporary mathematical substitution. You calculate with $\tilde{A}e^{i\omega t}$, and if you want to know the real oscillation, you take the real part of the complex exponential: you turn it back into a cosine. The substitution works because taking the real part commutes with every linear operation: adding, multiplying by a real number, and differentiating or integrating with respect to time. The time derivative of $\tilde{A}e^{i\omega t}$ is $i\omega\tilde{A}e^{i\omega t}$, so

$$\frac{d}{dt} \longrightarrow i\omega, \qquad \frac{d^2}{dt^2} \longrightarrow -\omega^2 : \tag{18.11}$$

a linear differential equation in time becomes an algebraic equation for the complex amplitude.

In the complex plane a complex amplitude is drawn as a vector, an arrow from the origin of length A at the angle ϕ from the positive real axis (Figure 18.7). The arrows stand still, since the factor $e^{i\omega t}$ is the same for all of them and is left out. The angle between two arrows is the difference between the phases of their oscillations, and with an assumed $e^{i\omega t}$ time dependence the arrow at the larger angle leads.

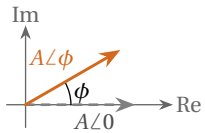


Figure 18.7 The complex amplitudes of the two oscillations of Figure 18.4, drawn as arrows that stand still: $A\angle 0$ (dashed) for $A\cos\omega t$, and $A\angle\phi$ (orange) for $A\cos(\omega t+\phi)$, which leads it by ϕ.

18.2.2 Adding oscillations

When you add two sines or cosines with the same frequency, you get a sinusoidal oscillation with the same frequency. With complex amplitudes the proof takes one line:

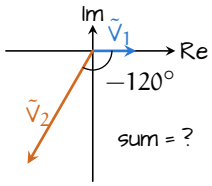
$$\operatorname{Re}\left[\tilde{x}_1 e^{i\omega t}\right] + \operatorname{Re}\left[\tilde{x}_2 e^{i\omega t}\right] = \operatorname{Re}\left[(\tilde{x}_1+\tilde{x}_2)\,e^{i\omega t}\right]. \tag{18.12}$$

The complex amplitude of the sum is the sum of the complex amplitudes. In the complex plane, adding them is just like adding vectors.

Example 18.1

Two sources of alternating voltage with the same frequency are connected in series, so that their voltages add. They are $V_1 = V_0 \cos \omega t$ and $V_2 = 3V_0 \cos(\omega t - 120°)$. What are the amplitude and the phase of the total voltage?

Solution:



The first step is a sketch of the two complex amplitudes as arrows in the complex plane: $\tilde{V}_1$ along the real axis, and $\tilde{V}_2$, three times as long, at $-120°$. The total voltage's complex amplitude is their sum. By (18.10) the complex amplitudes are $\tilde{V}_1 = V_0$ and $\tilde{V}_2 = 3V_0 e^{-i120°}$.

In rectangular form,

$$\tilde{V}_2 = 3V_0 \left(\cos 120° - i \sin 120°\right) = V_0(-1.5 - 2.598i),$$

so the total is

$$\tilde{V} = \tilde{V}_1 + \tilde{V}_2 = V_0(-0.5 - 2.598i).$$

Its magnitude is

$$\begin{aligned} |\tilde{V}| &= V_0 \sqrt{0.5^2 + 2.598^2} \\ &= \sqrt{7}\, V_0 \\ &= 2.646 V_0. \end{aligned}$$

A calculator gives $\arctan\left((-2.598)/(-0.5)\right) = 79.11°$, but the sketch shows that $\tilde{V}$ is in the third quadrant, so its phase is $79.11° - 180° = -100.89°$ (Figure 18.8). The total voltage is $V = 2.646 V_0 \cos(\omega t - 100.89°)$, which lags V_1 by $100.89°$.

Reality check: The three arrows make a triangle with sides V_0 and $3V_0$ and the angle $60°$ between them, so by the law of cosines $|\tilde{V}|^2 = V_0^2 + 9V_0^2 - 6V_0^2 \cos 60° = 7V_0^2$.

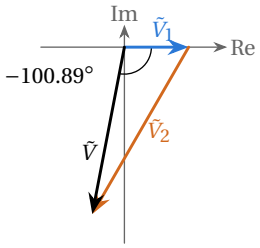


Figure 18.8 Example 18.1: the complex amplitudes $\tilde{V}_1$ and $\tilde{V}_2$ added tip to tail, and their sum $\tilde{V}$.

18.2.3 The driven oscillator

The rule (18.11) does the most work in a differential equation, as in the following example of a **driven oscillator**, which will come back in several later chapters.

Example 18.2

A mass m on a spring is pushed by the force $F_0 \cos\omega t$ and slowed by a damping force $-m\Gamma\, dx/dt$, proportional to its velocity, so that its displacement x from its position at rest obeys

$$\frac{d^2x}{dt^2} + \Gamma\frac{dx}{dt} + \omega_0^2 x = \frac{F_0}{m}\cos\omega t,$$

where ω_0 is the frequency at which it would oscillate with no damping and no driving force. Find the amplitude and the phase of its steady oscillation as functions of ω.

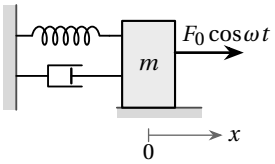
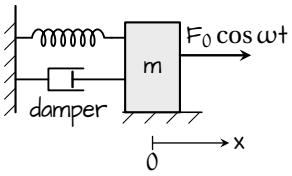


Figure 18.9 Example 18.2: the mass m, joined to a wall by a spring and a damper, and the driving force $F_0 \cos\omega t$.

Solution:



The first step is a sketch of the oscillator: the mass, the spring, the damping, and the driving force, with the displacement x that we want, measured from the position of the mass at rest. Write the force as $\mathrm{Re}[F_0 e^{i\omega t}]$, and look for a steady oscillation $x = \mathrm{Re}[\tilde{x}e^{i\omega t}]$ at the same frequency.

By (18.11) the differential equation becomes an algebraic one,

$$\left(-\omega^2 + i\omega\Gamma + \omega_0^2\right)\tilde{x} = \frac{F_0}{m},$$

and so

$$\tilde{x} = \frac{F_0/m}{\omega_0^2 - \omega^2 + i\omega\Gamma}.$$

The magnitude of a quotient is the quotient of the magnitudes, so the amplitude is

$$|\tilde{x}| = \frac{F_0/m}{\sqrt{(\omega_0^2 - \omega^2)^2 + \omega^2\Gamma^2}}.$$

The phase of a quotient is the difference of the phases. The numerator is real, and the denominator lies above the real axis, at the angle δ, with

$$\tan\delta = \frac{\omega\Gamma}{\omega_0^2 - \omega^2}, \qquad 0 < \delta < 180^\circ,$$

so $\tilde{x} = |\tilde{x}|e^{-i\delta}$. The phase of the displacement is $\phi = -\delta$: the mass lags the force by δ. Draw a picture to choose the quadrant. Below resonance, $\omega < \omega_0$, the denominator is in the first quadrant and δ is between 0 and 90°. Above resonance it is in the second quadrant and δ is between 90° and 180°, where the arctangent of a calculator would be wrong by 180°.

Reality check: At low frequencies the mass follows the force, in phase, with the amplitude $F_0/m\omega_0^2$ that a constant force would give.

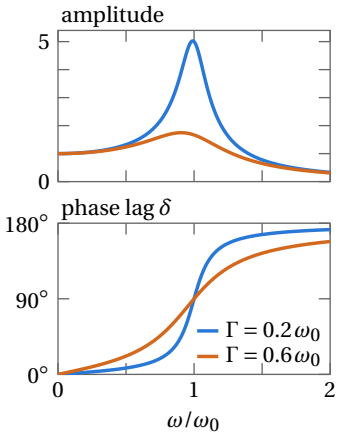


Figure 18.10 The amplitude (top) and the phase lag (bottom) of the oscillator of Example 18.2 against the driving frequency, for two damping rates. The amplitude is in units of $F_0/m\omega_0^2$, its value for a constant force.

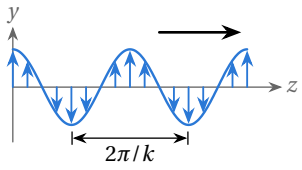


Figure 18.11 A sinusoidal wave traveling in the $+z$ direction, at one instant: here the electric field of an electromagnetic wave, along $\hat{\mathbf{y}}$. The wavelength is $2\pi/k$.

At resonance, $\omega = \omega_0$, the amplitude is $F_0/m\omega_0\Gamma$, and the mass lags the force by exactly 90°. (Why does the mass lag the force by exactly 90° at resonance? Answer in footnote.[1]) At high frequencies the amplitude falls as $F_0/m\omega^2$, and the mass moves opposite to the force, $\delta \to 180°$ (Figure 18.10).

The general solution of the equation also contains a solution of the equation with no driving force, which dies away (as $e^{-\Gamma t/2}$ when $\Gamma < 2\omega_0$). What is left after it has died away is the steady oscillation found here.

The same equation describes many oscillators: the current in a circuit made of a resistor, an inductor and a capacitor (Example 19.3), and an electron bound in an atom and driven by the electric field of a light wave, whose damping by its own radiation Example 27.3 finds.

18.2.4 Waves

As mentioned in Section 18.2.1, for wave propagation we are going to use the time-delay convention, in which the phase is defined by $\cos(\omega t - \phi)$. Since $\cos x = \cos(-x)$, that is the same as $\cos(-\omega t + \phi)$. When we talk about traveling waves, we need to include the dependence on position as well. In one dimension, a sinusoidal wave that travels in the positive z direction (Figure 18.11) is

$$f(z,t) = A\cos(kz - \omega t + \phi),$$

which we can represent as the real part of

$$Ae^{i(kz-\omega t+\phi)} \longleftrightarrow Ae^{i\phi}e^{i(kz-\omega t)} \longleftrightarrow \tilde{A}e^{i(kz-\omega t)}, \tag{18.13}$$

and at times we will have an assumed $e^{-i\omega t}$ time dependence. The complex amplitude $\tilde{A} = Ae^{i\phi}$ again holds the amplitude and the phase. We still take the *real* part of the wave when we want to know what the *real* wave is doing.

The main reason we do this is so that the sign of k can now be used to indicate which spatial direction the wave is traveling. For example, using numbers to make things more concrete,

$$(2\angle 5)\, e^{i(3z-4t)} = 2e^{i(3z-4t+5)}$$

becomes $2\cos(3z-4t+5)$ when we take the real part. It is a wave traveling to the right, with a phase shift of 5 radians toward later time relative to a reference wave $2\cos(3z-4t)$. And

$$2e^{i(-3z-4t+5)}$$

is a wave traveling to the *left*, with that same phase shift of 5 radians toward later time. As you will see, using k to specify the direction of travel will become hugely

[1] At $\omega = \omega_0$ the real part of the denominator, $\omega_0^2 - \omega^2$, is zero, so the denominator, $i\omega_0\Gamma$, is on the positive imaginary axis. Then the spring's force alone gives the mass its acceleration, and the driving force balances the damping force, so it is in phase with the velocity, which is a quarter cycle ahead of the displacement.

important in Section 21.4, where we allow for travel in more than just the $+z$ or $-z$ direction.

Derivatives turn into multiplications as before, but the time derivative now has the opposite sign from (18.11):

$$\frac{\partial}{\partial t} \longrightarrow -i\omega, \qquad \frac{\partial}{\partial z} \longrightarrow ik. \tag{18.14}$$

18.3 Complex wave numbers

In a material that absorbs energy from a wave, such as a metal (Section 22.2), solving the wave equation gives a complex value of k^2, and so a **complex wave number**, $\tilde{k} = k_{\text{real}} + ik_{\text{imag}}$. With it,

$$e^{i(\tilde{k}z-\omega t)} = e^{-k_{\text{imag}}z}\, e^{i(k_{\text{real}}z-\omega t)}. \tag{18.15}$$

The second factor is an ordinary wave with the wavelength $2\pi/k_{\text{real}}$, traveling in the $+z$ direction. The first is a real exponential that makes its amplitude decay with distance, by a factor of $1/e$ in each distance $1/k_{\text{imag}}$ (Figure 18.12). The intensity, which is proportional to the square of the amplitude, decays twice as fast, as $e^{-2k_{\text{imag}}z}$. So the real part of $\tilde{k}$ sets the wavelength, and the imaginary part sets the decay.

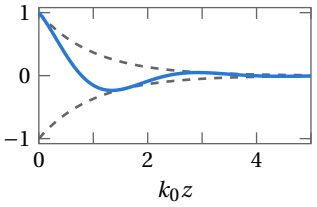


Figure 18.12 The real part of $e^{i(\tilde{k}z-\omega t)}$ at $t = 0$ for $\tilde{k} = (2+i)k_0$, and its envelope $\pm e^{-k_0 z}$ (dashed).

The wave equation gives $\tilde{k}^2$, so $\tilde{k}$ is one of its two square roots, which differ in sign. Choose the root by the physics: in a material that absorbs, a wave traveling toward $+z$ must decay in the direction it travels, so $k_{\text{imag}} > 0$, and $k_{\text{real}} > 0$ as well. For example, if $\tilde{k}^2 = (3+4i)k_0^2$, where $k_0 = \omega/c$ is the wave number in vacuum, the roots are $\pm(2+i)k_0$, since $(2+i)^2 = 3+4i$. The root $(2+i)k_0$ describes a wave traveling toward $+z$ with half the wavelength it would have in vacuum, whose amplitude falls by $1/e$ in the distance $1/k_0 = \lambda_0/2\pi$; in one of its own wavelengths, π/k_0, its amplitude falls by the factor $e^{-\pi} = 0.043$. The wave traveling toward $-z$ is $e^{i(-\tilde{k}z-\omega t)}$, with the same $\tilde{k}$, and it decays toward $-z$.

Here the principal square root happens to be the right one. It is not always, and Exercise P18.8 gives a case where the principal root describes a wave that grows. Why $\tilde{k}$ is complex in a material, and how it is related to the material's complex index of refraction $\tilde{n} = n + i\kappa$ and its complex permittivity, is the subject of Chapter 22.

18.4 Products and time averages

The real part of a sum is the sum of the real parts, but the real part of a product is *not* the product of the real parts: $\text{Re}[a]\,\text{Re}[b] \neq \text{Re}[ab]$. For example, $\text{Re}[e^{i\omega t}]\,\text{Re}[e^{i\omega t}] = \cos^2\omega t$, which is never negative, while $\text{Re}[e^{2i\omega t}] = \cos 2\omega t$, which is negative half of the time. So a product of two oscillating quantities, such as the power delivered by a voltage and a current or the Poynting vector of a wave, must be found from the real quantities, not from the complex ones.

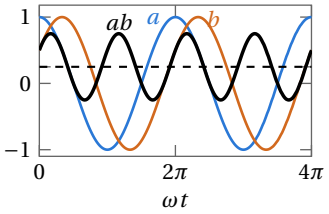


Figure 18.13 Two oscillations, $a = \cos\omega t$ and $b = \cos(\omega t - 60°)$, and their product ab (thick), which oscillates at 2ω about its average, $\frac{1}{2}\cos 60° = \frac{1}{4}$ (dashed).

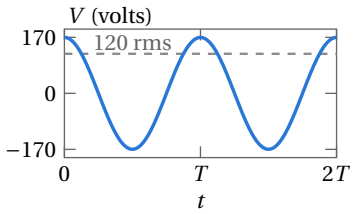


Figure 18.14 The voltage of a household outlet in the United States: an amplitude of 170 V, and an rms value of $170\,\text{V}/\sqrt{2} = 120\,\text{V}$ (dashed).

Take two quantities that oscillate at the same frequency, $a(t) = \text{Re}[\tilde{a}e^{i\omega t}]$ and $b(t) = \text{Re}[\tilde{b}e^{i\omega t}]$, in the phase-lead convention. With (18.6) for each real part,

$$\begin{aligned} a(t)\,b(t) &= \frac{\tilde{a}e^{i\omega t} + \tilde{a}^* e^{-i\omega t}}{2} \cdot \frac{\tilde{b}e^{i\omega t} + \tilde{b}^* e^{-i\omega t}}{2} \\ &= \frac{\tilde{a}\tilde{b}e^{2i\omega t} + \tilde{a}^*\tilde{b}^* e^{-2i\omega t}}{4} + \frac{\tilde{a}\tilde{b}^* + \tilde{a}^*\tilde{b}}{4} \\ &= \frac{1}{2}\text{Re}\left[\tilde{a}\tilde{b}\,e^{2i\omega t}\right] + \frac{1}{2}\text{Re}\left[\tilde{a}\tilde{b}^*\right]. \end{aligned}$$

The product has two parts: one that oscillates at twice the frequency, and one that is constant (Figure 18.13). The first averages to zero over a cycle, so the **time average** of the product, written with angle brackets, is

$$\boxed{\langle a(t)\,b(t)\rangle = \frac{1}{2}\text{Re}\left[\tilde{a}\tilde{b}^*\right]}. \tag{18.16}$$

The same steps with $e^{-i\omega t}$ in the place of $e^{i\omega t}$ give the same result, so (18.16) holds in both conventions. With $\tilde{a} = A_a e^{i\phi_a}$ and $\tilde{b} = A_b e^{i\phi_b}$ it is $\frac{1}{2}A_a A_b\cos(\phi_a - \phi_b)$: the average depends on the difference of the phases. Two oscillations in phase give $\frac{1}{2}A_a A_b$, and two that are 90° apart give zero.

For $b = a$, (18.16) gives the average of a square, $\langle a^2\rangle = \frac{1}{2}|\tilde{a}|^2$, which includes $\langle\cos^2\omega t\rangle = \frac{1}{2}$. The square root of $\langle a^2\rangle$ is the **root-mean-square** (rms) value of a, $a_{\text{rms}} = |\tilde{a}|/\sqrt{2}$. Household outlets in the United States are rated at 120 V, an rms value, so their voltage swings between +170 and −170 V (Figure 18.14). In this book a complex amplitude is a peak value. Many engineering books use rms values for their complex amplitudes, which removes the factor $\frac{1}{2}$ from (18.16), so find out which a source uses before you borrow a formula for power from it. Section 19.6 uses (18.16) for the power in circuits, and Section 20.2.4 for the average of the Poynting vector.

As an example, take the oscillator of Example 18.2. The power that the driving force delivers to the mass is $P = Fv$, with the velocity $v = \text{Re}[\tilde{v}e^{i\omega t}]$ and, by (18.11), $\tilde{v} = i\omega\tilde{x}$. Its average is

$$\begin{aligned} \langle P\rangle &= \frac{1}{2}\text{Re}\left[F_0\tilde{v}^*\right] \\ &= \frac{1}{2}\text{Re}\left[-i\omega F_0\tilde{x}^*\right] \\ &= -\frac{1}{2}\omega F_0\,\text{Im}[\tilde{x}]. \end{aligned}$$

Only the part of the motion that lags the force by 90°, so that its velocity is in phase with the force, takes in power on average. With $\tilde{x}$ from Example 18.2,

$$\langle P\rangle = \frac{F_0^2}{2m}\,\frac{\omega^2\Gamma}{(\omega_0^2 - \omega^2)^2 + \omega^2\Gamma^2}. \tag{18.17}$$

In the steady state this is the power that the damping force takes away, $m\Gamma\langle v^2\rangle$. It is largest at resonance, where it is $F_0^2/2m\Gamma$, and the width of the peak is set by Γ (Exercise P18.9).

18.5 The two sign conventions

This book uses both of the conventions of Section 18.2.1. For circuits (Chapter 19) it uses the phase-lead convention, $A\cos(\omega t+\phi) = \mathrm{Re}[\tilde{A}e^{i\omega t}]$, as most books on circuits do. For electromagnetic fields and waves it always uses the time-delay convention, with an assumed $e^{-i\omega t}$ time dependence and a wave traveling in the $+z$ direction written as $e^{i(kz-\omega t)}$, as most physics books do. The two differ in the sign of the exponent in $e^{\pm i\omega t}$, and so they are also called the two **sign conventions** for complex time dependence.

Most electrical engineers use the phase-lead convention, and they write j for $\sqrt{-1}$, since they use i for current. Engineers who use $e^{+i\omega t}$ nearly always write j instead of i. The letter j therefore does two jobs. It keeps $\sqrt{-1}$ apart from the current, which is the reason everyone will tell you about, and it also marks the convention: when j stands for $\sqrt{-1}$, you can be nearly certain that the equations use $e^{j\omega t}$ and the phase-lead convention. Unfortunately the same is not true of i, as this book shows: it uses i in both conventions, as is standard in physics.

So to use an engineering formula for a circuit in Chapter 19, just write i for j: an inductor's impedance $j\omega L$ is $i\omega L$ there, and a capacitor's $1/j\omega C$ is $1/i\omega C$. To change an engineering formula for electromagnetic waves into this book's convention for waves, replace every j by $-i$:

$$j \longrightarrow -i \qquad \text{(for waves)}. \tag{18.18}$$

For example, an engineering book writes a wave that travels toward $+z$ as $e^{j(\omega t-kz)}$, which becomes $e^{-i(\omega t-kz)} = e^{i(kz-\omega t)}$, and it writes the complex permittivity of a material that absorbs as $\varepsilon' - j\varepsilon''$, which becomes $\varepsilon' + i\varepsilon''$. The signs are written that way so that ε'' itself is *the same* in both conventions: a positive number for a material that absorbs. In the same way an index of refraction written $n - j\kappa$ becomes $n + i\kappa$, with the same κ. A physics book that writes its waves as $e^{i(\omega t-kz)}$ is translated the same way, by changing the sign of every i.

The rule (18.18) is actually not just a rule: it is a formal swap of which square root of -1 the two conventions use. Both i and $-i$ are square roots of -1, so either one is perfectly fine to use for any and all complex-number work. Swapping them turns every complex number $a+ib$ into its complex conjugate $a-ib$, so the change of convention is equivalent to a mirror reflection of the complex plane across the real axis (Figure 18.15). Real parts, and with them the physical fields, voltages and currents, are unchanged; every imaginary part and every phase angle changes sign, $\phi \to -\phi$. A complex amplitude at the angle ϕ in the phase-lead convention is at $-\phi$ in the time-delay convention, and both describe the same oscillation, since $A\cos(\omega t+\phi) = A\cos\big(\omega t-(-\phi)\big)$. So an arrow drawn at a larger angle leads in the phase-lead convention and lags in the time-delay convention.

The driven oscillator of Example 18.2 shows the reflection in a formula. With an assumed $e^{-i\omega t}$ time dependence the time derivative becomes $-i\omega$ (18.14), and the displacement becomes

$$\tilde{x} = \frac{F_0/m}{\omega_0^2 - \omega^2 - i\omega\Gamma},$$

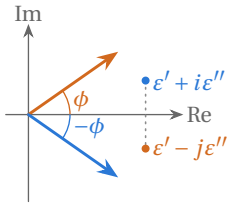


Figure 18.15 Changing the convention reflects the complex plane across the real axis. A complex amplitude at the angle ϕ in the phase-lead convention (orange) is at $-\phi$ in the time-delay convention (blue), and a permittivity written $\varepsilon' - j\varepsilon''$ in an engineering book (orange) is $\varepsilon' + i\varepsilon''$ in the time-delay convention (blue), with the same ε''.

the complex conjugate of the result of Example 18.2. Chapter 22 uses the oscillator in this form for the electrons in a material.

The arithmetic is the same in both conventions, but a choice that a convention makes need not be. The principal root is one: a calculator gives a negative real number the phase +180°, in either convention, and so it picks the same root in both, not mirror-image roots. For a wave in a material that absorbs nothing and has a negative permittivity, the principal root gives a wave that decays in the time-delay convention but one that grows in the engineers' (Exercise P18.8). Choose roots by the physics, then translate.

To tell which convention a source uses, look at how it writes a wave, an impedance, or a material that absorbs. The forms $e^{j(\omega t-kz)}$ and $e^{i(\omega t-kz)}$, an inductor's impedance $j\omega L$ or $i\omega L$, a permittivity $\varepsilon' - j\varepsilon''$, and an index of refraction $n - i\kappa$ with $\kappa > 0$ mark the phase-lead convention; $e^{i(kz-\omega t)}$, $-i\omega L$, $\varepsilon' + i\varepsilon''$ and $n + i\kappa$ mark the time-delay convention.

Electrical engineers have calculated with complex numbers since the 1890s. In 1893 Charles Proteus Steinmetz showed how complex quantities reduce the calculation of alternating-current circuits to algebra, as Example 18.2 did for the oscillator.



Charles Proteus Steinmetz (1865–1923, Prussian-American) Born Karl August Rudolph Steinmetz in Breslau (now Wrocław, Poland), he emigrated to the United States in 1889 and went to work for Rudolf Eickemeyer in Yonkers, New York. In 1893 the newly formed General Electric Company bought Eickemeyer's company, and Steinmetz with it. His paper of that year, "Complex Quantities and Their Use in Electrical Engineering," turned the calculation of alternating-current circuits into what he called "a simple problem of algebra." His textbook, *Theory and Calculation of Alternating Current Phenomena* (1897), is built on the method. He also found an empirical law for the energy lost to magnetic hysteresis (1892), and from 1902 to 1913 he chaired electrical engineering at Union College in Schenectady. (Wikipedia)

18.5.1 Tips for Mathematica and Python

Writing complex numbers. Mathematica's imaginary unit is `I`; Python's is `1j`, as in `3 + 4j` or `complex(3, 4)`. The real part, the imaginary part, the magnitude, the phase (in radians) and the complex conjugate are `Re`, `Im`, `Abs`, `Arg` and `Conjugate` in Mathematica, and `z.real`, `z.imag`, `abs(z)`, `numpy.angle(z)` and `z.conjugate()` in Python. Mathematica's `ComplexExpand` splits an expression into its real and imaginary parts, assuming that every variable in it is real.

Phases. `Arg` and `numpy.angle` take the quadrant into account. So does the arctangent with two arguments, whose arguments come in the opposite order in the two languages: `ArcTan[x, y]` in Mathematica and `numpy.arctan2(y, x)` in Python. With one argument, `ArcTan[y/x]` and `numpy.arctan(y/x)` fall into the trap of Section 18.1.3.

Square roots. NumPy keeps a real argument real: `numpy.sqrt(-1)` returns `nan` with a warning, while `numpy.sqrt(-1 + 0j)` and `cmath.sqrt(-1)` return `1j`. These functions, and Mathematica's `Sqrt`, return the principal root. On the negative real axis Python looks at the sign of a zero imaginary part, so that the square root of `complex(-4, 0.0)` is `2j`, but the square root of `complex(-4, -0.0)` is `-2j`. Choose roots by the physics.

18.6 Summary

An oscillation or a wave is the real part of a complex exponential. The substitution works for everything linear; products need the rule for time averages. This book writes the oscillations of circuits with an assumed $e^{i\omega t}$ (the phase-lead conven-

tion) and electromagnetic fields and waves with an assumed $e^{-i\omega t}$ (the time-delay convention).

Complex numbers

$z = x + iy$	rectangular form for adding,
$z = Ae^{i\theta} = A\angle\theta$	polar form for multiplying
$A = \lvert z\rvert = \sqrt{x^2+y^2}, \quad x = A\cos\theta,$ $y = A\sin\theta$	magnitude and phase; draw a picture to find θ
$e^{i\theta} = \cos\theta + i\sin\theta$	Euler's formula
$A_1 e^{i\theta_1} A_2 e^{i\theta_2} = A_1 A_2 e^{i(\theta_1+\theta_2)}$	multiply the magnitudes and add the phases
$z^* = x - iy = Ae^{-i\theta}, \quad zz^* = \lvert z\rvert^2$	complex conjugate: switch every i to $-i$
$\left(Ae^{i\theta}\right)^{1/n} = A^{1/n} e^{i(\theta+2\pi m)/n}$	n roots; a computer gives the principal one

Oscillations, waves and products

$A\cos(\omega t + \phi) = \mathrm{Re}\left[\tilde{A}e^{i\omega t}\right],$ $\tilde{A} = Ae^{i\phi} = A\angle\phi$	phase-lead convention (circuits): a positive ϕ is a lead
$A\cos(\omega t - \phi)$	time-delay convention: a positive ϕ is a delay
$d/dt \to i\omega$	with an assumed $e^{i\omega t}$: derivatives become multiplications
$\tilde{x}_1 + \tilde{x}_2$	oscillations of the same frequency add like vectors
$\tilde{x} = \dfrac{F_0/m}{\omega_0^2 - \omega^2 + i\omega\Gamma}$	the driven, damped oscillator, with an assumed $e^{i\omega t}$
$\mathrm{Re}\left[\tilde{A}e^{i(kz-\omega t)}\right], \quad \mathrm{Re}\left[\tilde{A}e^{i(-kz-\omega t)}\right]$	waves toward $+z$ and toward $-z$, in the time-delay convention
$\partial/\partial t \to -i\omega, \quad \partial/\partial z \to ik$	derivatives of waves
$\tilde{k} = k_{\mathrm{real}} + ik_{\mathrm{imag}}$: $e^{-k_{\mathrm{imag}}z} e^{i(k_{\mathrm{real}}z-\omega t)}$	wavelength from k_{real}, decay from k_{imag}; choose the root by the physics
$\langle ab\rangle = \frac{1}{2}\mathrm{Re}\left[\tilde{a}\tilde{b}^*\right], \quad a_{\mathrm{rms}} = \lvert\tilde{a}\rvert/\sqrt{2}$	time average of a product, in either convention; never $\mathrm{Re}\left[\tilde{a}\tilde{b}\right]$
$j \to i$ (circuits), $\quad j \to -i$ (waves)	engineering formulas into this book's conventions; $j \to -i$ changes the sign of imaginary parts and phases, not of real parts

Exercises

Exercises for 18.1 Complex numbers

Warm-up questions Read Section 18.1, then answer before class

W18.1 A classmate says that multiplying a complex number by $1+i$ moves its point in the complex plane one unit to the right and one unit up. What is wrong with this? What does multiplying by $1+i$ do to the point? (Write $1+i$ in polar form.)

W18.2 Use the polar form to explain why zz^* is a positive real number for every nonzero complex number z, while z^2 is real only for some z. For which z is z^2 real?

W18.3 Adding 2π to the phase of a complex number does not change the number. Why, then, does (18.7) give a different cube root for $m=1$ than for $m=0$? Would $m=3$ give yet another cube root?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P18.1 The complex conjugate of a quotient.

(a) Find the complex conjugate of $(1+3i)/(4-5i)$ by switching every i to $-i$, and then write it in rectangular form.

(b) Now do it the long way. First write $(1+3i)/(4-5i)$ in rectangular form, by multiplying the numerator and the denominator by the complex conjugate of the denominator; then take the complex conjugate. Do the two answers agree?

(c) Use the polar form to show that the shortcut works for every quotient: $(z_1/z_2)^* = z_1^*/z_2^*$.

P18.2 Rectangular and polar forms. Draw a picture for each part.

(a) Write $2-i$, $-2+i$, $-3-4i$ and $-5i$ in the form $A\angle\theta$, with θ between $-180°$ and $180°$. For which of them does $\arctan(y/x)$ give the right angle?

(b) Write $10e^{i\pi/6}$, $2\angle -135°$ and $e^{i\pi}$ in rectangular form. Show that $10e^{i\pi/6}$ and $10\angle 30°$ are the same number.

(c) Describe the set of all the points in the complex plane with $|z|=5$, the set with $|z-(1+i)|=2$, and the set with the phase $30°$.

P18.3 Roots.

(a) Find all five solutions of $z^5 = 1$, in polar and in rectangular form, and draw them in the complex plane.

(b) Show that the five solutions add up to zero.

(c) Find both square roots of i. Which one is the principal root?

P18.4 Using Euler's formula.

(a) Show that $\cos\theta = (e^{i\theta} + e^{-i\theta})/2$ and $\sin\theta = (e^{i\theta} - e^{-i\theta})/2i$.

(b) Derive the formulas for $\cos(\alpha+\beta)$ and $\sin(\alpha+\beta)$ from $e^{i(\alpha+\beta)} = e^{i\alpha}e^{i\beta}$.

(c) Show that $\cos 3\theta = 4\cos^3\theta - 3\cos\theta$. (Hint: $e^{3i\theta} = (e^{i\theta})^3$.)

(d) Here is a proof of Euler's formula without series. Show that $f(\theta) = e^{-i\theta}(\cos\theta + i\sin\theta)$ has $df/d\theta = 0$, and use the value of $f(0)$ to finish the proof.

Exercises for 18.2 Oscillations and waves as complex exponentials

Warm-up questions Read Section 18.2, then answer before class

W18.4 A source says that the current in a circuit has a phase of 40° relative to the voltage $V_0\cos\omega t$, and a classmate concludes that the current reaches its peaks before the voltage does. What else must you know before you can agree?

W18.5 At high frequencies the mass of Example 18.2 moves opposite to the force that drives it. Which term on the left of its differential equation is largest at high frequencies? Use (18.11) to explain why that term makes the mass move opposite to the force.

W18.6 A classmate says that the wave $\mathrm{Re}[e^{i(3z+4t)}]$ travels toward $+z$, since the coefficient of z is positive. Write the wave as a cosine, and follow one of its crests for a small time change to decide whether the classmate is right. When does the sign of k tell you which way a wave travels?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P18.5 Three-phase power. Electric power is often distributed on three wires whose voltages, measured from a common neutral wire, are $V_1 = V_0\cos\omega t$, $V_2 = V_0\cos(\omega t - 120°)$ and $V_3 = V_0\cos(\omega t - 240°)$.

(a) Draw the three complex amplitudes in the complex plane, and show that $V_1 + V_2 + V_3 = 0$ at every instant.

(b) Find the amplitude and the phase of the voltage between two of the wires, $V_1 - V_2$. Does it lead or lag V_1?

(c) The rms voltage between each wire and the neutral wire (Section 18.4) is often 120 V. What is the rms voltage between two of the wires?

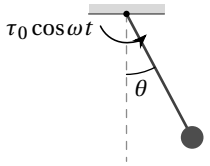


Figure 18.16 The pendulum of Exercise P18.6, with its angle θ from the vertical and the driving torque $\tau_0 \cos\omega t$.

P18.6 A driven pendulum. A pendulum with damping is driven by a torque that oscillates at the frequency ω, so that its angle θ from the vertical (Figure 18.16) obeys

$$\frac{d^2\theta}{dt^2} + \Gamma\frac{d\theta}{dt} + \omega_0^2 \sin\theta = \frac{\tau_0}{I}\cos\omega t,$$

where I is its moment of inertia. For small angles, $\sin\theta \cong \theta$, and this is the equation of Example 18.2. Take $\Gamma = 0.1\omega_0$ and $\omega = 0.9\omega_0$, and use the units $\omega_0 = 1$.

(a) Solve the equation numerically for $\tau_0/I = 0.005\omega_0^2$, starting from rest at $\theta = 0$, and plot $\theta(t)$. After the start-up oscillation has died away, measure the amplitude and the phase lag of the motion, and compare them with Example 18.2. About how long does the start-up oscillation take to die away?

(b) Repeat (a) with the drive forty times as strong, $\tau_0/I = 0.2\omega_0^2$. Compare the amplitude and the phase lag of the steady motion with the predictions of Example 18.2.

(c) Why can the method of complex amplitudes not describe the motion of (b)?

Exercises for 18.3 Complex wave numbers

Warm-up questions Read Section 18.3, then answer before class

W18.7 Figure 18.12 shows the wave for $\tilde{k} = (2+i)k_0$. How would the figure change for $\tilde{k} = (2+0.1i)k_0$? For $\tilde{k} = (4+i)k_0$? Sketch or describe each. (Consider k_{real} and k_{imag} separately.)

W18.8 The amplitude of the wave in Figure 18.12 falls as the wave travels toward $+z$. Do its crests slow down as well? Explain with (18.15).

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P18.7 A strongly absorbing material. Solving the wave equation in a material gives $\tilde{k}^2 = (-3+4i)k_0^2$, where $k_0 = \omega/c$.

(a) Find both square roots, and choose the one that describes a wave traveling toward $+z$.

(b) Find the wavelength in the material, in units of the wavelength in vacuum $\lambda_0 = 2\pi/k_0$, and the distance in which the intensity falls by a factor of e.

(c) By what factor does the amplitude fall in one wavelength?

P18.8 The principal-root trap. In a material that absorbs nothing and has the negative relative permittivity $\varepsilon_r = -4$, the wave number is $\tilde{k} = k_0\sqrt{\varepsilon_r}$, where $k_0 = \omega/c$.

(a) Take $\tilde{k}$ to be the principal square root. Does the wave $e^{i(\tilde{k}z-\omega t)}$ decay or grow toward $+z$?

(b) An engineering book writes the same wave as $e^{j(\omega t-\tilde{k}z)}$ and also takes the principal root, $\sqrt{-4} = 2j$. Does its wave decay or grow toward $+z$?

(c) Now let the material absorb a little: $\tilde{\varepsilon}_r = -4+0.01i$ in the time-delay convention, which is $-4-0.01j$ in the engineers'. Find the principal root in each convention. Do both waves decay now?

(d) What rule should you follow instead of trusting the principal root?

Exercises for 18.4 Products and time averages

Warm-up questions Read Section 18.4, then answer before class

W18.9 The time average (18.16) depends only on the difference of the two phases, $\phi_a - \phi_b$, and not on each phase separately. Why must that be so?

W18.10 At low frequencies the mass of Example 18.2 moves in phase with the force. Is the average power that the force delivers to the mass then large, or nearly zero? During part of each cycle the force does positive work on the mass; what becomes of that energy?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P18.9 The resonance of Example 18.2.

(a) Show that the amplitude $|\tilde{x}|$ is largest at $\omega = \sqrt{\omega_0^2 - \Gamma^2/2}$ (if $\Gamma < \sqrt{2}\omega_0$), slightly below ω_0, but that the average power (18.17) is largest at exactly ω_0.

(b) Show that the average power falls to half of its largest value at two frequencies ω_+ and ω_- whose difference is exactly $\omega_+ - \omega_- = \Gamma$. (Hint: solve $(\omega_0^2 - \omega^2)^2 = \omega^2\Gamma^2$ as two quadratic equations.)

(c) Find the phase lag δ of the oscillator at ω_- and at ω_+.

P18.10 Power in an alternating-current circuit. The voltage across a device is $V(t) = V_0 \cos\omega t$, and the current into it is $I(t) = I_0 \cos(\omega t + \phi)$.

(a) Find the average power $\langle VI \rangle$ by averaging the product of the two cosines over one period. Then find it again from (18.16).

(b) A student instead takes the real part of the product of the complex quantities, $\mathrm{Re}[\tilde{V}e^{i\omega t}\,\tilde{I}e^{i\omega t}]$, and averages that. What does the student get?

(c) For an ideal inductor, $V = L\,dI/dt$. Use (18.11) to find ϕ and the average power. Does the current lead or lag the voltage?

(d) Show that the average power into a resistor R is V_{rms}^2/R.

Exercises for 18.5 The two sign conventions

Warm-up questions Read Section 18.5, then answer before class

W18.11 Why does the time average $\langle ab \rangle = \frac{1}{2}\mathrm{Re}[\tilde{a}\tilde{b}^*]$ of Section 18.4 hold in both conventions? Explain with the mirror picture of Figure 18.15.

W18.12 In the time-delay convention the displacement $\tilde{x}$ of the driven oscillator is the complex conjugate of the one in Example 18.2, so its phase is $+\delta$ rather than $-\delta$. Does the mass then lead the force? Explain.

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

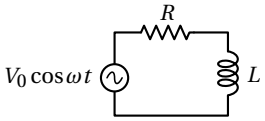


Figure 18.17 The resistor and the inductor of Exercise P18.11, in series with the source of the voltage $V_0 \cos\omega t$.

P18.11 Translating an engineering formula. A resistor R and an inductor L in series are driven by the voltage $V_0 \cos\omega t$ (Figure 18.17). An engineering book gives the current as $I(t) = \mathrm{Re}[\tilde{I}e^{j\omega t}]$, with $\tilde{I} = V_0/(R + j\omega L)$.

(a) Write the formula as Chapter 19 would, in the phase-lead convention with i for j, and find $I(t)$ from it.

(b) Translate the formula into the time-delay convention, with an assumed $e^{-i\omega t}$ time dependence, by the rule (18.18), and find $I(t)$ from it. Show that it is the same function of time as in (a). What is the phase of $\tilde{I}$ in each convention?

(c) Derive $\tilde{I}$ directly in the time-delay convention, from the circuit equation $RI + L\,dI/dt = V_0 \cos\omega t$.

(d) Repeat (a) and (b) for a resistor in series with a capacitor C, whose impedance an engineering book gives as $1/j\omega C$.

Exercises for 18.6 Summary

P18.12 True or false? Explain your reasoning.

(a) $|e^{i\theta}| = 1$ for every complex number θ.

(b) The real part of the product of two complex numbers is the product of their real parts.

(c) The sum of two sinusoidal oscillations with the same frequency is a sinusoidal oscillation with that frequency, or zero.

(d) With an assumed $e^{i\omega t}$ time dependence, an oscillation with the complex amplitude $2i$ leads one with the complex amplitude 2 by 90°.

(e) $\sqrt{z^2} = z$ for every complex number z, if the square root is the principal root.

(f) If $\tilde{k}$ is purely imaginary, $e^{i(\tilde{k}z-\omega t)}$ oscillates in time but does not travel.

(g) Changing every i to $-i$ in a correct equation gives another correct equation, if every variable in it is real.

(h) The average of $\cos\omega t \sin\omega t$ over a cycle is zero.

(i) The rms value of $V_0 \cos\omega t$ is $V_0/2$.

P18.13 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as how the graph of $A\cos(\omega t + \phi)$ shifts as ϕ changes, the sum of many oscillations with random phases, the response of a driven oscillator and its start-up oscillation, the motion of a driven pendulum at large amplitudes, waves with complex wave numbers, or the roots of a complex number and the branch cut of the square root.

Chapter 19

Ohm's Law, Induction, and Circuits

This chapter follows charges as they move through matter and around circuits. Section 19.1 completes Maxwell's equations inside materials for fields that change, with the current that a changing polarization carries. Sections 19.2 and 19.3 take up conduction: Ohm's law, which says how a current flows through a material in response to a field, and Drude's model of the electrons in a metal, which explains it. Section 19.4 treats inductance, the property of a circuit that resists a change in its current, and shows that the energy of the magnetic field is the work needed to set up the currents. Sections 19.5 through 19.7 then solve circuits driven by sources that oscillate, with the complex numbers of Chapter 18: impedances, the power that such circuits take in, and filters that pass some frequencies and block others.

19.1 The Ampère–Maxwell law in matter

Chapters 9 and 13 described matter with the fields **D** and **H**, whose sources are the free charges and currents alone, and found the boundary conditions that static fields obey at a surface. Section 16.1.4 then stated the Ampère–Maxwell law in matter, (16.5), which includes a current that static fields do not have: the polarization current. This section derives that current, collects Maxwell's equations in the form used inside materials, and shows that the boundary conditions still hold when the fields change. These are the equations that Chapters 21 through 23 solve.

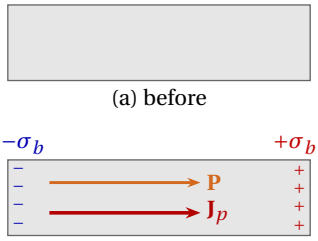


Figure 19.1 A block of dielectric (a) before it is polarized and (b) while its polarization **P** grows. The bound surface charges $\pm\sigma_b$ on its ends were not there before, so charge must be crossing the block: the polarization current $\mathbf{J}_p$.

19.1.1 The polarization current

Take a block of dielectric that is not polarized, and polarize it along its length, for example by placing it between the plates of a capacitor and charging the capacitor (Figure 19.1). Before, the block has no bound charge anywhere. Afterward, its ends carry the bound surface charges $\sigma_b = \mathbf{P}\cdot\hat{\mathbf{n}} = \pm P$ of (9.13). The block as

a whole is still neutral, so charge has moved from one end to the other: while the polarization was growing, a current flowed through the block. This is the **polarization current**. Its density follows from the charge that collects on the end where σ_b is positive. If that end has area A, the charge on it is $q = \sigma_b A = PA$, and all of it arrived through the block, so the current density across the block is

$$\begin{aligned} J_p &= \frac{1}{A}\frac{dq}{dt} \\ &= \frac{d\sigma_b}{dt} \\ &= \frac{dP}{dt}. \end{aligned}$$

The same result holds however the polarization varies from place to place. The bound charge density $\rho_b = -\nabla\cdot\mathbf{P}$, (9.14), holds at every instant, so it changes at the rate

$$\frac{\partial\rho_b}{\partial t} = -\nabla\cdot\frac{\partial\mathbf{P}}{\partial t}.$$

This is a continuity equation, (10.12), with $\partial\mathbf{P}/\partial t$ in the place of the current density: bound charge is conserved, and $\partial\mathbf{P}/\partial t$ is the current that carries it. The same holds at a surface, where (9.13) gives $\partial\sigma_b/\partial t = (\partial\mathbf{P}/\partial t)\cdot\hat{\mathbf{n}}$, the rate at which this current arrives at the surface from inside. The polarization current density is therefore

$$\mathbf{J}_p = \frac{\partial\mathbf{P}}{\partial t}. \tag{19.1}$$

A continuity equation by itself fixes a current density only up to a part with zero divergence, since such a part carries no charge anywhere. The picture of a dielectric as a collection of molecular dipoles settles the question: as each dipole stretches or turns, its two charges move, and their motion carries exactly the current density (19.1) (Figure 16.4 and Exercise P19.1). The bound current $\nabla\times\mathbf{M}$ of a magnetized material, (13.13), comes from charges that circulate inside the atoms, and it is a separate current.

How large is the polarization current? Suppose that the field in the block of Figure 19.1 is switched on quickly and then held steady. The polarization grows with the field, and the polarization current flows only while it grows: once the field is steady, $\partial\mathbf{P}/\partial t = 0$ and the current stops (Figure 19.2). The charge per unit area that it carries across the block, the area under the curve of J_p against the time, is the final polarization $P = \varepsilon_0\chi_e E$ of the linear dielectric, however quickly the field was switched on. So after the switching the polarization current is gone, but the faster the switching, the larger the current while it lasts. For a field that oscillates, $E = E_0\cos\omega t$, the polarization current density is $J_p = \partial P/\partial t = -\omega\varepsilon_0\chi_e E_0\sin\omega t$, whose amplitude grows in proportion to the frequency: the faster the field changes, the more the polarization current matters. Its size depends on the frequency in a second way as well, since χ_e itself depends on how quickly the field changes. A material does not respond to a field that oscillates quickly in the same way as to a steady one (Section 22.5).

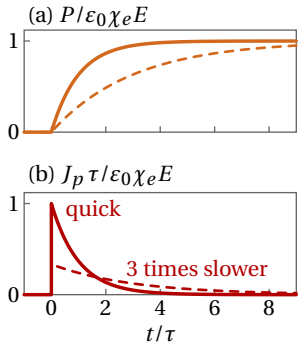


Figure 19.2 The block of Figure 19.1 when the field is switched on at $t = 0$, rising as $1 - e^{-t/\tau}$, and then held steady (solid), and when it is switched on three times more slowly (dashed). (a) The polarization, as a fraction of its final value $\varepsilon_0\chi_e E$. (b) The polarization current density. The areas under the two curves in (b) are the same.

In a capacitor filled with a dielectric of large ε_r, nearly all of the displacement current $\partial\mathbf{D}/\partial t$ between the plates is polarization current. Pure water at 20°C, for example, keeps its static dielectric constant, $\varepsilon_r = 80.1$ (Table 9.2), at frequencies up to far above 1 MHz, and in a water-filled capacitor driven at 1 MHz, 98.8% of $\partial\mathbf{D}/\partial t$ between the plates is polarization current (Exercise P16.3).

19.1.2 Maxwell's equations in matter

Inside matter the total current density has three parts: the free current $\mathbf{J}_f$, the bound current $\nabla\times\mathbf{M}$, and the polarization current (19.1). Section 16.1.4 put all three into the Ampère–Maxwell law and found (16.5),

$$\nabla\times\mathbf{H} = \mathbf{J}_f + \frac{\partial\mathbf{D}}{\partial t}, \qquad \text{with} \qquad \frac{\partial\mathbf{D}}{\partial t} = \varepsilon_0\frac{\partial\mathbf{E}}{\partial t} + \frac{\partial\mathbf{P}}{\partial t}.$$

The polarization current is part of this displacement current, so it must not be added a second time. Gauss's law gains no new term. Both $\nabla\cdot\mathbf{E} = \rho/\varepsilon_0$, with $\rho = \rho_f + \rho_b$, and $\rho_b = -\nabla\cdot\mathbf{P}$ hold at every instant, so $\nabla\cdot\mathbf{D} = \rho_f$, (9.25), holds whether or not the fields change. The other two equations contain no charges or currents, so they are the same in matter as in empty space. Together, these are **Maxwell's equations in matter**:

$$\boxed{\begin{aligned} \nabla\cdot\mathbf{D} &= \rho_f, & \nabla\times\mathbf{E} &= -\frac{\partial\mathbf{B}}{\partial t},\\ \nabla\cdot\mathbf{B} &= 0, & \nabla\times\mathbf{H} &= \mathbf{J}_f + \frac{\partial\mathbf{D}}{\partial t}, \end{aligned}} \tag{19.2}$$

where $\mathbf{D} = \varepsilon_0\mathbf{E} + \mathbf{P}$ and $\mathbf{H} = \mathbf{B}/\mu_0 - \mathbf{M}$, (9.21) and (13.18). They hold in every material.

By themselves they are not enough to find the fields, since besides $\mathbf{E}$ and $\mathbf{B}$ they contain $\mathbf{D}$ and $\mathbf{H}$. The material supplies the rest: how its polarization and magnetization, and its free current, depend on the fields. These are called **constitutive relations**. In linear materials they are (9.22) and (13.20),

$$\mathbf{D} = \varepsilon\mathbf{E}, \qquad \mathbf{B} = \mu\mathbf{H}, \tag{19.3}$$

and in a conductor that obeys Ohm's law the free current is $\mathbf{J}_f = \sigma\mathbf{E}$ (Section 19.2). When the fields change quickly, ε, μ, and σ depend on how quickly (Section 22.4). Free charge, like bound charge, is conserved by itself (Exercise P19.2).

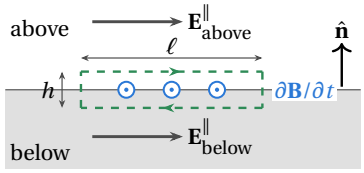


Figure 19.3 A squat loop (dashed) that straddles a surface, with a changing magnetic field through it (⊙). As the loop is flattened, $h\to 0$, the flux through it goes to zero, so $\mathbf{E}^{\parallel}$ is the same just above and just below the surface, as in electrostatics.

19.1.3 Boundary conditions for changing fields

The boundary conditions of the static chapters came from the divergence equations applied to a thin pillbox, and the curl equations applied to a squat loop (Figure 19.3), that straddles a surface (Sections 4.4, 5.2, 9.8, 11.3, and 13.9). When the fields change, the two curl equations gain the time derivatives $-\partial\mathbf{B}/\partial t$ and

$\partial\mathbf{D}/\partial t$, and neither changes the result. Take Faraday's law, (15.4), around a squat loop with long sides of length ℓ just above and just below the surface (Figure 19.3):

$$\oint \mathbf{E}\cdot d\boldsymbol{\ell} = -\frac{d}{dt}\int \mathbf{B}\cdot d\mathbf{a}.$$

As the loop is flattened, its short sides contribute nothing to the left side, and the flux on the right, and with it its rate of change, goes to zero with the loop's area, as long as $\mathbf{B}$ stays finite. Fields may jump at a surface, but they do not become infinite there, so the parallel component of $\mathbf{E}$ is continuous, exactly as in electrostatics (as Section 15.2.3 noted). In the same way, the flux of $\partial\mathbf{D}/\partial t$ through the loop goes to zero, so only the free surface current makes $\mathbf{H}^{\parallel}$ jump. The divergence equations contain no time derivatives at all. So at a surface at rest, the boundary conditions are the same at every instant as in the static case (Table 19.1). (At a moving surface, such as the surface of a moving conductor, they need more care.)

Equation	Boundary condition
$\nabla\cdot\mathbf{D} = \rho_f$	$D^{\perp}_{\text{above}} - D^{\perp}_{\text{below}} = \sigma_f$
$\nabla\times\mathbf{E} = -\partial\mathbf{B}/\partial t$	$\mathbf{E}^{\parallel}_{\text{above}} = \mathbf{E}^{\parallel}_{\text{below}}$
$\nabla\cdot\mathbf{B} = 0$	$B^{\perp}_{\text{above}} = B^{\perp}_{\text{below}}$
$\nabla\times\mathbf{H} = \mathbf{J}_f + \partial\mathbf{D}/\partial t$	$\mathbf{H}^{\parallel}_{\text{above}} - \mathbf{H}^{\parallel}_{\text{below}} = \mathbf{K}_f\times\hat{\mathbf{n}}$

Table 19.1 The boundary conditions at a surface at rest, each with the equation of (19.2) that it comes from. They hold at every instant when the fields change. "Above" is the side that $\hat{\mathbf{n}}$ points toward.

In terms of the total surface charge and current, the first and last conditions are the jumps of σ/ε_0 in $E^{\perp}$ and of $\mu_0\mathbf{K}\times\hat{\mathbf{n}}$ in $\mathbf{B}^{\parallel}$, (5.5) and (11.11). At an interface between two linear materials, (19.3) turns the four conditions into conditions on $\mathbf{E}$ and $\mathbf{B}$:

$$\begin{aligned}
\varepsilon_{\text{above}}E^{\perp}_{\text{above}} - \varepsilon_{\text{below}}E^{\perp}_{\text{below}} &= \sigma_f,\\
\mathbf{E}^{\parallel}_{\text{above}} - \mathbf{E}^{\parallel}_{\text{below}} &= 0,\\
B^{\perp}_{\text{above}} - B^{\perp}_{\text{below}} &= 0,\\
\frac{1}{\mu_{\text{above}}}\mathbf{B}^{\parallel}_{\text{above}} - \frac{1}{\mu_{\text{below}}}\mathbf{B}^{\parallel}_{\text{below}} &= \mathbf{K}_f\times\hat{\mathbf{n}}.
\end{aligned} \tag{19.4}$$

At most interfaces between two materials there is no free charge or current on the surface, and the right sides are all zero.

A free surface current needs a perfect conductor. In a material with a finite conductivity σ, the current density $\sigma\mathbf{E}$ is finite wherever $\mathbf{E}$ is, so no current can flow in a layer of zero thickness, and $\mathbf{K}_f = 0$ at its surface. When the fields oscillate quickly, the current crowds into a thin layer at the surface, the skin effect of Section 22.2.2. A **perfect conductor**, with $\sigma\to\infty$, is the limit in which that layer has zero thickness and the current in it becomes a true surface current. Section 23.1 uses this limit for the walls of waveguides.

The potentials are continuous too, since a jump in V or in a parallel component of $\mathbf{A}$ would make $\mathbf{E}$ or $\mathbf{B}$ infinite at the surface (Sections 5.3.4 and 12.1.3). The

perpendicular component of **A** is not tied to the fields in this way. It is continuous in the Coulomb gauge, (12.5), and in the Lorenz gauge (Section 24.1), in which $\nabla \cdot \mathbf{A}$ is finite.

19.2 Ohm's law and conductivity

Apply an electric field to a piece of metal, and each of its conduction electrons feels a force. Does an electron then speed up all the way to the far end of the metal? It does not. The electrons collide with the atoms of the metal, and the collisions slow them down, much as the air slows a skydiver in free fall. A skydiver soon stops gaining speed and falls at a steady speed, and in the same way the conduction electrons soon settle into a steady average velocity, the **drift velocity** $\mathbf{v}_d$, which is proportional to the field. (Section 19.3 gives a model that explains why.) Moving charge carries the current density (10.10), the charge density times the velocity. If there are N conduction electrons per unit volume, each with charge $-e$, their charge density is $-Ne$, so

$$\mathbf{J} = -Ne\mathbf{v}_d. \tag{19.5}$$

The electrons drift opposite to the field, so **J** points along **E**, and since $\mathbf{v}_d$ is proportional to **E**, so is **J**:

$$\mathbf{J} = \sigma\mathbf{E}. \qquad \text{(Ohm's law)} \tag{19.6}$$

This is **Ohm's law**. The constant σ is the **conductivity** of the material, and its reciprocal,

$$\rho = \frac{1}{\sigma}, \tag{19.7}$$

is the **resistivity**. Resistivity is measured in ohm meters and conductivity in **siemens** per meter, where $1\,\mathrm{S} = 1\,\Omega^{-1} = 1\,\mathrm{A/V}$. Table 19.2 lists the resistivities and conductivities of some pure metals at 20°C.[1] The resistivity of a metal rises with its temperature. Between 273 K and 293 K, that of copper rises from 1.543 to $1.678 \times 10^{-8}\,\Omega\,\mathrm{m}$, and that of aluminum from 2.417 to $2.650 \times 10^{-8}\,\Omega\,\mathrm{m}$ (same source), about 0.4 percent for each kelvin in both. Section 19.3 explains why.

Ohm's law is not a law in the sense that Maxwell's equations are. It is a constitutive relation (Section 19.1.2): it describes how one kind of material responds to a field, and other materials need not obey it. Metals obey it very well. A material that does is called **ohmic**. Many materials and devices are not ohmic: in a semiconductor diode, for example, the current is far larger for one direction of the field than for the other.

[1] From "Electrical Resistivity of Pure Elemental Metals," in *CRC Handbook of Chemistry and Physics*, 107th Edition (Internet Version 2026), Leah R. McEwen, ed., CRC Press/Taylor & Francis, Boca Raton, FL. The values for copper, silver, and gold are R. A. Matula's (Section 16.4.1).



Georg Simon Ohm (1789–1854, German) Ohm's father, a locksmith in Erlangen, taught his sons mathematics and physics himself. Ohm received his doctorate from the University of Erlangen in 1811, and in 1817 he became a teacher of mathematics and physics at the Jesuit Gymnasium in Cologne, whose well-equipped laboratory let him begin experiments with electric currents. In papers of 1825 and 1826 he described conduction in a circuit by analogy with Joseph Fourier's theory of the conduction of heat, and in 1827 he published the book *Die galvanische Kette, mathematisch bearbeitet* (The Galvanic Circuit Investigated Mathematically), which contains his law. It was received with little enthusiasm at first, and he left his post at Cologne. He taught at the Polytechnic School in Nuremberg from 1833, and in 1852 he became professor of experimental physics at the University of Munich. The Royal Society gave him its Copley Medal in 1841. The unit of resistance is named for him. (Wikipedia)

☞ *Caution*: many references write n for the number of conduction electrons per unit volume. This book writes N, because it uses n for the index of refraction, from Chapter 21 on.

☞ *Caution*: in this chapter σ is the conductivity and ρ the resistivity, not the surface and volume charge densities. There are not enough Greek letters to go around, and the context shows which is meant.

Metal	ρ ($10^{-8}\,\Omega$m)	σ (10^{7} S/m)
Silver	1.587	6.30
Copper	1.678	5.96
Gold	2.214	4.52
Aluminum	2.650	3.77
Sodium	4.77	2.10
Tungsten	5.28	1.89
Nickel	6.93	1.44
Iron	9.61	1.04
Platinum	10.5	0.952
Lead	20.8	0.481

Table 19.2 Resistivities and conductivities of some pure metals at 20°C (source in the footnote).

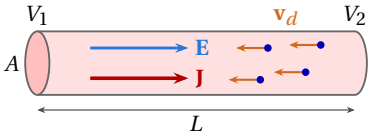


Figure 19.4 A length L of uniform wire with cross section A. The field $\mathbf{E}$ inside drives the current density $\mathbf{J} = \sigma\mathbf{E}$ from the end at V_1 toward the end at the lower potential V_2, while the conduction electrons drift the other way with the drift velocity $\mathbf{v}_d$.

19.2.1 The field inside a wire

Section 4.5 showed that the field inside a conductor is zero. That holds only when the charges are at rest. A wire that carries a current is not static: its conduction electrons move, and by Ohm's law it takes a field inside the wire, $\mathbf{E} = \mathbf{J}/\sigma$, to keep them moving. Where does that field come from? When the current is steady, charge does not pile up anywhere, so $\nabla\cdot\mathbf{J} = 0$ by the continuity equation (10.12). In a material with uniform σ, Ohm's law then gives $\nabla\cdot\mathbf{E} = 0$, and by Gauss's law there is no charge inside the wire. The charges that produce the field sit on the wire's surface (and at the terminals of the battery). At the surface of the wire, no current flows out into the air, so the component of $\mathbf{J}$, and so of $\mathbf{E}$, perpendicular to the surface is zero just inside it: the field inside runs along the wire (Figure 19.4). The surface charges arrange themselves so that the field follows every bend in the wire. When the battery is first connected, they rearrange within a tiny fraction of a second, and the electrons all around the circuit begin to drift at nearly the same time.

19.2.2 Resistance

Take a uniform wire of length L and cross section A, with the potential difference $\Delta V = V_1 - V_2$ between its ends (Figure 19.4). The field inside is uniform and along the wire, $E = \Delta V/L$, and the current density is uniform too, $J = I/A$. Ohm's law, $J = \sigma E = E/\rho$, then gives

$$\frac{I}{A} = \frac{1}{\rho}\frac{\Delta V}{L}, \qquad \text{or} \qquad \Delta V = I\left(\frac{\rho L}{A}\right).$$

The quantity in parentheses is the **resistance** of the wire,

$$R = \frac{\rho L}{A}, \tag{19.8}$$

measured in ohms, and

$$\Delta V = IR \tag{19.9}$$

is Ohm's law in the form you met in your introductory course. With words: a wire's resistance grows in proportion to its length and falls in inverse proportion to its cross section, and the material enters only through ρ. For a conductor of any other shape the resistance is still $R = \Delta V/I$, but finding the current requires knowing the field everywhere inside (Section 19.2.4).

19.2.3 Ohmic heating

The field does work on the drifting electrons. In a time dt the charge dq in a small volume moves the distance $\mathbf{v}_d\,dt$, and the field does the work $dq\,\mathbf{E}\cdot\mathbf{v}_d\,dt$ on it. Per unit volume, the charge is $-Ne$, so the field does work at the rate $-Ne\mathbf{v}_d\cdot\mathbf{E} = \mathbf{J}\cdot\mathbf{E}$ per unit volume (Section 20.2.1 finds the same result for any moving charges). In an ohmic material this is

$$\mathbf{J}\cdot\mathbf{E} = \sigma E^2 = \frac{J^2}{\sigma}. \tag{19.10}$$

The electrons' drift does not speed up, so this energy is not stored as their kinetic energy. They give it to the atoms in collisions, and it becomes heat: **ohmic heating**, also called Joule heating. For the uniform wire, multiplying by the volume AL gives the power $(JA)(EL)$, or

$$\begin{aligned} P &= I\Delta V \\ &= I^2R \\ &= \frac{(\Delta V)^2}{R}. \end{aligned} \tag{19.11}$$

The energy reaches the wire through the fields outside it, as Example 20.1 shows.

19.2.4 Numerical solutions: current in a conductor of any shape

To find the resistance of a conductor of any shape, we need the potential everywhere inside it. Section 19.2.1 showed that a steady current in a material with uniform σ leaves no charge inside, and when nothing changes, $\mathbf{E} = -\nabla V$. So inside the conductor the potential obeys Laplace's equation,

$$\nabla^2 V = 0,$$

exactly as in the empty regions of Chapter 7. What is new is a boundary condition. Where the conductor touches an electrode, a good conductor connected to a battery, V is given, as before. On every other part of its surface the conductor meets an insulator, no current crosses, and so the component of ∇V perpendicular to the surface is zero:

$$\frac{\partial V}{\partial n} = 0 \qquad \text{(insulating surfaces)}, \tag{19.12}$$

where $\partial/\partial n$ is the derivative along the normal. A condition on the normal derivative, rather than on V itself, is called a **Neumann boundary condition**. With V given on the electrodes and (19.12) on the rest of the surface, the potential is unique (Exercise P19.7).

The relaxation method of Section 7.3 handles the new condition with one change. Put the grid points at the centers of the squares of the grid, so that each insulating edge lies halfway between a point inside and a point just outside. Before each pass, set the value at every point just outside an insulating edge equal to its mirror image just inside; then the two values are the same, there is no difference of potential across the edge, and no current crosses it. At an electrode held at V_0, set the value just outside to $2V_0$ minus the value just inside, so that the two average to V_0 on the edge. When the grid has relaxed, add up the current that crosses the electrode, and divide the potential difference by it.

A resistor made of a thin film of thickness t shows how this works. A square of the film, with the current flowing between two opposite edges, has L and the width w equal, so by (19.8) its resistance is $\rho w/(tw) = \rho/t$, whatever the size of the square. This is the **sheet resistance** $R_s = \rho/t$, often quoted in "ohms per square." A straight strip of length L and width w is L/w squares in a row, with the resistance R_sL/w. Now bend the strip through a right angle, as in Figure 19.5.

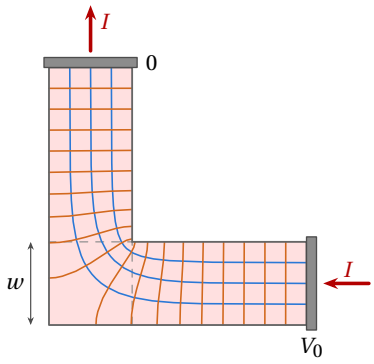


Figure 19.5 An L-shaped strip of resistive film between two electrodes (dark gray). Equipotentials (orange) at equal steps and current lines (blue) that divide the current into four equal parts, computed by the relaxation method. Each region between neighboring lines is nearly a square of the film, so it has the resistance R_s. The corner square (dashed) counts as only 0.56 square.

How many squares is the corner square worth? The current takes the shorter path around the inside of the bend: there the current lines crowd together and the equipotentials crowd together too. So the corner square offers less resistance than a square in a straight run. With 20 grid squares across the strip, the relaxation method gives the corner as 0.565 square, with 40 as 0.561, and with 80 as 0.560; as the grid is made finer the value approaches 0.559.[2] The figure shows the result in another way. Its current lines and equipotentials divide the strip into regions that are nearly squares, each of which has the resistance R_s of a square of film. Each current tube is a row of 19 of them in series, and the four tubes are in parallel, so the whole strip has the resistance $19R_s/4 = 4.75R_s$: it counts as 4.75 squares.

19.3 Microscopic view of conduction

Why is the drift velocity proportional to the field? In 1900 Paul Drude proposed a model of conduction in metals in which the conduction electrons are a gas of free particles that collide with the atoms of the metal.[3] Between collisions the field speeds an electron up, and each collision sends it off in a random direction, so that it loses the extra velocity that the field gave it. On average, the collisions remove that extra velocity in a time τ, and so they act on the average velocity $\mathbf{v}$ of the conduction electrons like a damping force, $-m\mathbf{v}/\tau$. With the force of the field added, Newton's second law for the average electron is

$$m\frac{d\mathbf{v}}{dt} = -e\mathbf{E} - \frac{m}{\tau}\mathbf{v}. \tag{19.13}$$

This is **Drude's scattering model**, usually called simply the **Drude model**. The **damping time** τ is of the order of the average time between collisions.

In a steady field the average velocity stops changing, just as a skydiver's speed does, and with $d\mathbf{v}/dt = 0$ the model gives the drift velocity

$$\mathbf{v}_d = -\frac{e\tau}{m}\mathbf{E}. \tag{19.14}$$

It is proportional to $\mathbf{E}$, as Section 19.2 said it must be. Putting it into (19.5) gives $\mathbf{J} = (Ne^2\tau/m)\mathbf{E}$, which is Ohm's law with the conductivity

$$\sigma_0 = \frac{Ne^2\tau}{m}. \tag{19.15}$$

The subscript 0 marks this as the **DC conductivity**, the conductivity for a field that does not change (DC stands for direct current). When the field oscillates in a time comparable to τ, the term $m\,d\mathbf{v}/dt$ can no longer be dropped, and



Paul Drude (1863–1906, German) Drude's doctoral thesis, written under Woldemar Voigt at Göttingen in 1887, was on the reflection and refraction of light in absorbing crystals, and his measurements of the optical constants of solids were the most precise of their time. In 1900 he proposed the model of conduction in metals that bears his name, (19.13), in which free electrons move like the molecules of a gas and lose their extra velocity in collisions, and published an influential textbook of optics. From 1900 he edited the *Annalen der Physik*, then the leading physics journal. (Wikipedia)

[2] The exact value, found by conformal mapping, is $1-(2/\pi)\ln 2 = 0.5587$: "Resistance calculations for thin film patterns," P. M. Hall, Thin Solid Films 1, 277–295 (1968). https://doi.org/10.1016/0040-6090(68)90046-1

[3] "Zur Elektronentheorie der Metalle," P. Drude, Ann. Phys. 306, 566–613 (1900). https://doi.org/10.1002/andp.19003060312

the conductivity becomes a complex number that depends on the frequency (Section 22.4.2).

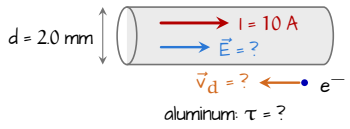
How quickly does the drift settle down? If the field is switched on at $t = 0$, the solution of (19.13) that starts from $\mathbf{v} = 0$ is $\mathbf{v} = \mathbf{v}_d\left(1 - e^{-t/\tau}\right)$. The average velocity approaches the drift velocity exponentially, in the time τ, as a skydiver's speed would approach its final value if the drag of the air were proportional to the speed. For a metal at room temperature τ is of the order of 10^{-14} s (Example 19.1), so the drift is steady almost at once.

The model also explains why the resistivity of a metal rises with its temperature (Section 19.2). At a higher temperature the atoms vibrate more strongly, an electron is scattered more often, τ is shorter, and by (19.15) the resistivity $\rho = m/Ne^2\tau$ is larger. Impurities and defects in the crystal scatter electrons too, which is why an alloy usually has a larger resistivity than the pure metals it is made of.

Example 19.1

An aluminum wire 2.0 mm in diameter carries a steady current of 10 A. Aluminum has three conduction electrons per atom, so that $N = 1.81 \times 10^{29}\,\mathrm{m}^{-3}$. Find the field in the wire, the drift speed of the electrons, and the damping time τ.

Solution:



We sketch a piece of the wire with what is given and what is asked. The drift velocity points opposite to the current, because the electrons' charge is negative.

The current is spread uniformly over the cross section, of radius $a = 1.0\,\mathrm{mm}$, so the current density is

$$\begin{aligned} J &= \frac{I}{\pi a^2} \\ &= \frac{10\,\mathrm{A}}{\pi\left(1.0 \times 10^{-3}\,\mathrm{m}\right)^2} \\ &= 3.18 \times 10^6\,\mathrm{A/m^2}. \end{aligned}$$

The resistivity of aluminum is $2.650 \times 10^{-8}\,\Omega\mathrm{m}$ (Table 19.2), so by Ohm's law the field along the wire is

$$\begin{aligned} E &= \rho J \\ &= \left(2.650 \times 10^{-8}\,\Omega\mathrm{m}\right)\left(3.18 \times 10^6\,\mathrm{A/m^2}\right) \\ &= 0.0843\,\mathrm{V/m}. \end{aligned}$$

By (19.5), the drift speed is

$$\begin{aligned} v_d &= \frac{J}{Ne} \\ &= \frac{3.18\times10^{6}\,\mathrm{A/m^2}}{\left(1.81\times10^{29}\,\mathrm{m^{-3}}\right)\left(1.602\times10^{-19}\,\mathrm{C}\right)} \\ &= 1.10\times10^{-4}\,\mathrm{m/s}, \end{aligned}$$

about a tenth of a millimeter per second. At that speed an electron takes about two and a half hours to drift 1 m. (The lights still come on at once when you flip a switch, because the field is set up all along the wire almost at once, Section 19.2.1.) Finally, by (19.15) with $\sigma_0 = 1/\rho$,

$$\begin{aligned} \tau = \frac{m}{Ne^2\rho} &= \frac{9.109\times10^{-31}\,\mathrm{kg}}{\left(1.81\times10^{29}\,\mathrm{m^{-3}}\right)\left(1.602\times10^{-19}\,\mathrm{C}\right)^2\left(2.650\times10^{-8}\,\Omega\,\mathrm{m}\right)} \\ &= 7.40\times10^{-15}\,\mathrm{s}. \end{aligned}$$

Reality check: The units of $m/Ne^2\rho$ are $\mathrm{kg/(m^{-3}\cdot C^2\cdot\Omega\cdot m)} = \mathrm{kg\cdot m^2/(C^2\cdot\Omega)}$. An ohm is a volt per ampere, or $\mathrm{V\cdot s/C}$, so $\mathrm{C^2\cdot\Omega = C\cdot V\cdot s = J\cdot s}$, and $\mathrm{kg\cdot m^2/(J\cdot s) = s}$, a time, as τ must be.

The drift is a tiny addition to the much faster random motion of the electrons.[4]

19.4 Inductance

Chapter 12 defined the mutual inductance M and the self-inductance L, (12.30) and (12.32), and Chapter 15 found the EMFs they produce when currents change, (15.5). This section treats the inductor as a part of a circuit, shows that the energy (12.35) is the work needed to set up a distribution of currents (which Section 12.7.2 could not do without Faraday's law), and takes up two circuits that share their flux, as the coils of a transformer do.

19.4.1 The inductor in a circuit

A coil made to have a particular inductance is called an **inductor**. When the current I through it changes, the back EMF $-L\,dI/dt$, (15.5), opposes the change:

[4] In a metal that random motion is not thermal. By quantum mechanics, the conduction electrons move at about the Fermi speed, which for a gas of free electrons is $(\hbar/m)(3\pi^2N)^{1/3}$, or 2.0×10^6 m/s for aluminum, at ordinary temperatures. The distance an electron goes between collisions, its mean free path, is then about $(2.0\times10^6\,\mathrm{m/s})\tau = 15\,\mathrm{nm}$, dozens of times the spacing of the atoms. In the quantum picture the electrons are scattered not by the atoms themselves but by their vibrations and by imperfections of the crystal.

inductors resist a change in current. In a circuit, the potential difference between the inductor's terminals is

$$V_L = L\frac{dI}{dt}, \tag{19.16}$$

with the terminal where the current enters at the higher potential while the current grows (Figure 19.6a).[5]

As an example, connect a battery of EMF $\mathscr{E}$, a resistor R, and an inductor L in series, and close the switch at $t = 0$ (Figure 19.6). Going around the circuit, the EMF of the battery equals the potential drops across the resistor and the inductor,

$$\mathscr{E} = IR + L\frac{dI}{dt}.$$

The solution with $I = 0$ at $t = 0$ is

$$I(t) = \frac{\mathscr{E}}{R}\left(1 - e^{-Rt/L}\right). \tag{19.17}$$

Without the inductor the current would jump to $\mathscr{E}/R$ at once. With it, the back EMF holds the current back, and it approaches $\mathscr{E}/R$ exponentially, with the **time constant** L/R: after one time constant it has reached $1 - 1/e = 63\%$ of its final value.

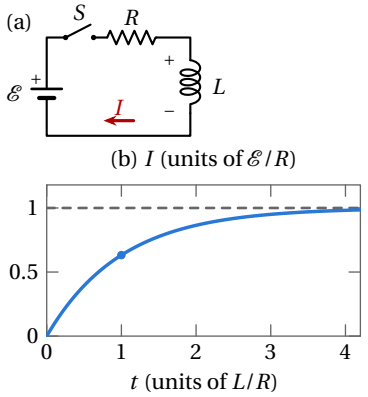


Figure 19.6 (a) A battery, a switch S, a resistor, and an inductor in series. While the current grows, the end of the inductor where it enters is at the higher potential. (b) The current after the switch is closed at $t = 0$, (19.17). Dashed: its final value $\mathscr{E}/R$. The dot marks $t = L/R$.

19.4.2 The energy stored in an inductor

Starting a current in an inductor takes work, because the inductor resists the change. While the current grows at the rate dI/dt, whatever drives it must push charge through the potential difference (19.16), so it delivers energy to the inductor at the rate

$$P = V_L I = \left(L\frac{dI}{dt}\right) I.$$

Building the current up from zero to I therefore takes the work

$$\begin{aligned} W &= \int P\,dt \\ &= \int_0^I LI'\,dI' \\ &= \frac{1}{2}LI^2, \end{aligned} \tag{19.18}$$

which is (12.37), now shown to be the work needed to set up the current. Compare $\frac{1}{2}CV^2$ for a capacitor. In the circuit of Figure 19.6, the battery supplies energy at the rate $\mathscr{E}I$. Part of it becomes heat in the resistor, and the rest goes into the inductor until the current stops growing (Exercise P19.10).

[5] This is the potential difference that a voltmeter reads when its leads stay outside the coil's magnetic field. If the coil's wire has no resistance, the field inside the wire is zero: the induced field there is canceled by the field of charges that collect on the coil, and those charges produce the potential difference between its terminals. A voltmeter whose leads loop through the coil's own field reads something else, because the induced field is not a gradient field.

The same is true for currents spread through space in any way. While the currents grow, the magnetic field changes, and it induces an electric field; whatever drives the currents must keep them flowing against that field. Take the conductors to have no resistance, so that no energy is lost as heat. By (20.10), the field does work on all of the moving charges at the rate $\int \mathbf{E}\cdot\mathbf{J}\,d\tau$, integrated over all of space. Where this integrand is negative, as inside a battery, the charges give energy to the field, so the sources supply energy at the rate

$$\frac{dW}{dt} = -\int \mathbf{E}\cdot\mathbf{J}\,d\tau = \int \mathbf{J}\cdot\nabla V\,d\tau + \int \mathbf{J}\cdot\frac{\partial \mathbf{A}}{\partial t}\,d\tau,$$

using $\mathbf{E} = -\nabla V - \partial\mathbf{A}/\partial t$, (15.11). If the currents grow slowly, they do not pile up anywhere, so $\nabla\cdot\mathbf{J} = 0$, and the first integral is $\int \nabla\cdot(V\mathbf{J})\,d\tau = \oint V\mathbf{J}\cdot d\mathbf{a} = 0$, on a surface outside all of the currents. In the second, put $\mu_0\mathbf{J} = \nabla\times\mathbf{B}$ (Ampère's law, which holds for slowly changing currents, Section 16.4) and use the product rule of Section 12.7.2 with $\partial\mathbf{A}/\partial t$ in the place of $\mathbf{A}$:

$$\frac{dW}{dt} = \frac{1}{\mu_0}\int \frac{\partial\mathbf{A}}{\partial t}\cdot(\nabla\times\mathbf{B})\,d\tau = \frac{1}{\mu_0}\int \mathbf{B}\cdot\frac{\partial\mathbf{B}}{\partial t}\,d\tau - \frac{1}{\mu_0}\oint\left(\frac{\partial\mathbf{A}}{\partial t}\times\mathbf{B}\right)\cdot d\mathbf{a},$$

since $\nabla\times(\partial\mathbf{A}/\partial t) = \partial\mathbf{B}/\partial t$. On a large sphere the surface term goes to zero, as in Section 12.7.2, and $\mathbf{B}\cdot\partial\mathbf{B}/\partial t = \frac{1}{2}\partial B^2/\partial t$. So

$$\frac{dW}{dt} = \frac{d}{dt}\left(\frac{1}{2\mu_0}\int B^2\,d\tau\right),$$

and the work needed to build the currents up from zero is $(1/2\mu_0)\int B^2\,d\tau$, (12.35). With words: the energy stored in the magnetic field is the work done against the induced electric field while the field was being built. With magnetic materials present, the same steps with $\nabla\times\mathbf{H} = \mathbf{J}_f$ give $\int \mathbf{H}\cdot\partial\mathbf{B}/\partial t\,d\tau$, which for a linear material is the rate of change of $\frac{1}{2}\int \mathbf{B}\cdot\mathbf{H}\,d\tau$, (13.35).

19.4.3 Two circuits

Now take two circuits, with self-inductances L_1 and L_2 and mutual inductance M. Build up the current in circuit 1 to I_1 first, with no current in circuit 2; that takes the work $\frac{1}{2}L_1I_1^2$. Then hold I_1 fixed while the current in circuit 2 grows to I_2. That takes the work $\frac{1}{2}L_2I_2^2$ in circuit 2. But while I_2 grows it induces the EMF $-M\,dI_2/dt$ in circuit 1, (15.5), and to keep I_1 from changing, the source in circuit 1 must push against it, delivering energy at the rate $MI_1\,dI_2/dt$, or MI_1I_2 in all. The total is

$$U = \frac{1}{2}L_1I_1^2 + MI_1I_2 + \frac{1}{2}L_2I_2^2. \tag{19.19}$$

Setting up I_2 first gives the same total only because $M_{12} = M_{21}$, which Section 12.7.1 found from the Neumann formula. Exercise P19.11 follows the term MI_1I_2 when two loops with fixed currents are brought together.

The energy $U = (1/2\mu_0)\int B^2\,d\tau$ cannot be negative, whatever the currents are, and that limits M. If I_1 and I_2 circulate so that their fields oppose each

other, the middle term of (19.19) is negative. For a given I_1, U is smallest when $I_2 = -MI_1/L_2$, and then

$$U = \frac{I_1^2}{2L_2}\left(L_1L_2 - M^2\right).$$

Since this cannot be negative,

$$M^2 \le L_1L_2. \tag{19.20}$$

The ratio $k = M/\sqrt{L_1L_2}$, between -1 and 1, is called the **coupling coefficient**. It is close to ± 1 when nearly all of the flux of each circuit passes through the other, as when two coils are wound on one iron core, and close to zero for two coils far apart or at right angles (Exercise P19.12).

19.4.4 The transformer

An alternating current in one coil makes an alternating magnetic field, which makes an alternating flux through a second coil, which induces an alternating EMF in the second coil. This is basically how transformers work. In a **transformer** the two coils, the primary with N_1 turns and the secondary with N_2, are wound on one core of iron, which guides nearly all of the flux of each coil through every turn of the other (Figure 19.7). If every turn of both coils encloses the same flux Φ, Faraday's law gives the EMFs $-N_1\,d\Phi/dt$ in the primary and $-N_2\,d\Phi/dt$ in the secondary, so the voltages across the two coils (with no resistance in their windings) are in the ratio of their turns:

$$\frac{V_2}{V_1} = \frac{N_2}{N_1}. \tag{19.21}$$

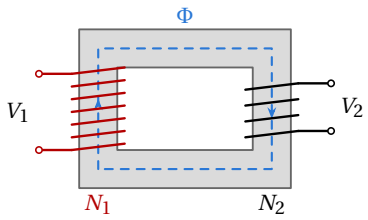


Figure 19.7 A transformer: the primary coil of N_1 turns (red) and the secondary coil of N_2 turns (black) on one iron core (gray), which carries the flux Φ (blue) through every turn of both.

A transformer with many more turns on its secondary than on its primary steps a voltage up, and one with fewer steps it down. The same flux through every turn also means that, for a core of linear material, the flux per turn is proportional to $N_1I_1 + N_2I_2$, so L_1, M, and L_2 are proportional to N_1^2, N_1N_2, and N_2^2 with one constant in common: then $M^2 = L_1L_2$, and $k = 1$. A real transformer comes close to this ideal. Some flux leaks out of the core, and some energy is lost to the resistance of the windings and to the eddy currents (Section 15.2.5) and hysteresis (Section 13.8.2) of the core. If no energy is lost, the average power that goes into the primary equals the average power that the secondary delivers to whatever is connected to it. Section 19.5 takes up a transformer that drives a load.

19.5 Complex impedance

So far this chapter has followed the current inside a conductor. This section and the next two treat whole circuits, made of resistors, capacitors and inductors and driven by a source whose voltage oscillates sinusoidally, as the voltage of a household outlet does: alternating-current, or AC, circuits. Complex amplitudes (Section 18.2.1) make such circuits nearly as easy to solve as circuits driven by a battery. Circuits in this book are written in the phase-lead convention

of Section 18.2.1, $A\cos(\omega t+\phi) = \mathrm{Re}[\tilde{A}e^{i\omega t}]$, with an assumed $e^{i\omega t}$ time dependence. The fields and waves of the other chapters use the time-delay convention (Section 18.5).

☞ *Caution*: books on circuits write j for $\sqrt{-1}$, and an inductor's impedance as $j\omega L$. In the phase-lead convention used here the formulas are the same with i in place of j (Section 18.5).

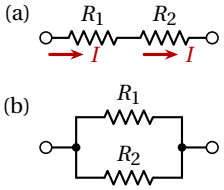


Figure 19.8 Two resistors (a) in series, carrying the same current, and (b) in parallel, with the same potential difference between their ends.

19.5.1 Circuits with steady currents

Start with what you saw in your introductory course. A battery drives a steady current through a network of resistors, each of which obeys Ohm's law in the form $\Delta V = IR$, (19.9). Two resistors in series, one after the other, carry the same current, and their potential differences add, so together they act as a single resistor of resistance $R_1 + R_2$ (Figure 19.8a). Two resistors in parallel, side by side, have the same potential difference across them, and their currents add, so together they act as a single resistor whose resistance R obeys $1/R = 1/R_1 + 1/R_2$ (Figure 19.8b):

$$R_{\text{series}} = R_1 + R_2, \qquad R_{\text{parallel}} = \left(\frac{1}{R_1} + \frac{1}{R_2}\right)^{-1} \equiv R_1 // R_2. \tag{19.22}$$

The symbol $R_1 // R_2$ is a convenient shorthand for the parallel combination. A network made of resistors in series and in parallel can be reduced, step by step, to one equivalent resistance.

A network that cannot be reduced in this way, such as one with two batteries in different branches, is solved with two rules known as **Kirchhoff's laws**:

1. The **junction law**: the currents flowing into any junction of the circuit add up to the currents flowing out of it.

2. The **loop law**: the potential differences around any closed loop of the circuit add up to zero, with the EMF of each battery counted as a rise in potential.

The junction law is the conservation of charge. When the currents are steady, no charge piles up at a junction, so $\nabla\cdot\mathbf{J} = 0$ there, (10.12), and as much current flows out as flows in. The loop law holds because a static electric field has $\oint \mathbf{E}\cdot d\boldsymbol{\ell} = 0$ around every closed path (Section 5.1). With an inductor in the circuit the electric field is no longer static: its curl is $-\partial\mathbf{B}/\partial t$, and its line integral around a loop through the inductor is the inductor's EMF, $-L\,dI/dt$. But the changing magnetic field is nearly all inside the inductor's coil, so outside the coil the curl of $\mathbf{E}$ is nearly zero. There $\mathbf{E}$ is still a gradient field, and the potential difference between the coil's terminals, along any path that stays outside the coil, is $V_L = L\,dI/dt$, (19.16). With V_L counted among the potential differences, the loop law holds as before, as Section 19.4.1 already used it. Both laws fail when a circuit is not small compared with the wavelength of electromagnetic waves at its frequency. Then charge does pile up along the wires, and the changing fields are not confined to the inductors and capacitors; Section 23.8 treats one such case, the transmission line.

19.5.2 Complex amplitudes and impedance

Now drive a circuit with a source whose voltage oscillates, $V(t) = V_0\cos\omega t$, choosing the zero of time so that the voltage is largest at $t = 0$. When the source is

switched on, the currents go through a **transient** that dies away. What is left is the **steady state**, in which every current and every potential difference in a circuit of resistors, capacitors and inductors oscillates at the same frequency ω as the source, though not necessarily in phase with it.[6] Each of these oscillations is described by its complex amplitude, (18.10): a current $I_0\cos(\omega t+\phi_I)$ is $\mathrm{Re}[\tilde{I}e^{i\omega t}]$ with $\tilde{I}=I_0e^{i\phi_I}=I_0\angle\phi_I$, and the source's voltage has the real complex amplitude $\tilde{V}=V_0$.

For a resistor, a capacitor or an inductor, the complex amplitude of the potential difference across it is proportional to the complex amplitude of the current through it. The ratio is its **impedance** $\tilde{Z}$:

$$\tilde{V}=\tilde{Z}\tilde{I}. \tag{19.23}$$

For a resistor, $V=IR$ at every instant, so its impedance is R. For a capacitor, the charge on the plates is $Q=CV$, (5.19), and the current that flows onto them is $I=dQ/dt=C\,dV/dt$. With an assumed $e^{i\omega t}$ time dependence, d/dt becomes $i\omega$, (18.11), so $\tilde{I}=i\omega C\tilde{V}$, and the capacitor's impedance is $1/i\omega C$. For an inductor, $V_L=L\,dI/dt$, (19.16), so $\tilde{V}=i\omega L\tilde{I}$. In summary,

$$\tilde{Z}_R=R,\qquad \tilde{Z}_C=\frac{1}{i\omega C}=-\frac{i}{\omega C},\qquad \tilde{Z}_L=i\omega L. \tag{19.24}$$

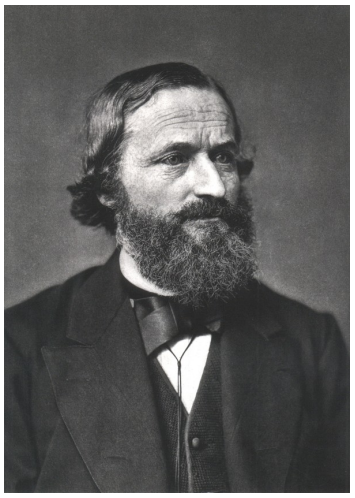
With these, (19.23) is Ohm's law for capacitors and inductors as well as for resistors. Impedances are measured in ohms. The real part of an impedance is its resistance, and its imaginary part is its **reactance**, positive for an inductor and negative for a capacitor. The word *impedance* is Oliver Heaviside's, from 1886 (his portrait is in Section 2.4).

The phases of (19.24) show how each element's voltage and current line up in time (Figure 19.9). In a resistor they are in phase. An inductor's impedance is $i\omega L=\omega L\angle 90^\circ$, so its voltage leads its current by 90°: the voltage is largest when the current is changing fastest. A capacitor's impedance is $(1/\omega C)\angle -90^\circ$, so its voltage lags its current by 90°: charge keeps flowing onto the plates as long as the current is positive, and the voltage is largest when the current reverses.

Kirchhoff's laws hold at every instant, and they are linear, so they hold for the complex amplitudes too: the $\tilde{I}$'s at a junction add up to zero, and so do the $\tilde{V}$'s around a loop. The series and parallel rules (19.22) follow from Kirchhoff's laws and Ohm's law alone, so with (19.23) in place of Ohm's law they carry over unchanged:

$$\tilde{Z}_{\text{series}}=\tilde{Z}_1+\tilde{Z}_2,\qquad \tilde{Z}_{\text{parallel}}=\left(\frac{1}{\tilde{Z}_1}+\frac{1}{\tilde{Z}_2}\right)^{-1}=\tilde{Z}_1/\!/\tilde{Z}_2. \tag{19.25}$$

[6] When a source is switched on, the equation of the circuit also has solutions with no source at all (complementary solutions), which add the transient, just as for the driven oscillator of Example 18.2. Finding them takes the methods for differential equations of particular and complementary solutions, which this book uses only in the simplest cases, such as the LR circuit of Section 19.4.1. Impedances describe the steady state that is left after the transient has died away.



Gustav Robert Kirchhoff (1824–1887, German) Kirchhoff was born in Königsberg, Prussia, and studied at its university, in the seminar in mathematical physics of Franz Neumann and Carl Jacobi. He formulated his two circuit laws in 1845, as an exercise for the seminar, and the work later became his doctoral thesis under Neumann. He taught at Berlin from 1847 and at Breslau from 1850. In 1854 he moved to Heidelberg, where he worked with the chemist Robert Bunsen on spectroscopy: in 1859 Kirchhoff showed from its spectrum that the Sun contains sodium, and with Bunsen he discovered the elements cesium and rubidium. In 1857 he calculated that an electric signal travels along a wire with no resistance at the speed of light, and in 1859 he stated his law of thermal radiation; in 1860 he coined the term *black body*. He returned to Berlin in 1875. (Wikipedia)

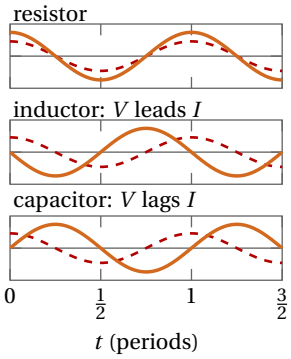


Figure 19.9 The current (dashed red) and the voltage (orange) of a resistor, an inductor and a capacitor, against time. The current is drawn smaller, so that both curves show.

These rules work for any combination of resistors, capacitors and inductors. What adds is the complex amplitudes, not the amplitudes themselves: two voltages that are out of phase can add up to less than either one.

19.5.3 Solving a circuit with impedances

To find the current that a source of the voltage $V_0 \cos \omega t$ supplies to a circuit:

1. Find the circuit's equivalent impedance $\tilde{Z}$ with the series and parallel rules (19.25).
2. Write it in polar form, $\tilde{Z} = |\tilde{Z}| \angle \phi$.
3. Divide: $\tilde{I} = \tilde{V}/\tilde{Z} = (V_0/|\tilde{Z}|) \angle -\phi$.
4. Convert back to a function of time, if you want one:

$$I(t) = \frac{V_0}{|\tilde{Z}|} \cos(\omega t - \phi).$$

If the phase ϕ of the impedance is positive, as for an inductor, the current lags the source's voltage; if ϕ is negative, as for a capacitor, the current leads (Figure 19.9). Once $\tilde{I}$ is known, the voltages and currents of the parts of the circuit follow from (19.23) applied to each part.

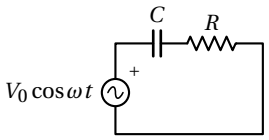
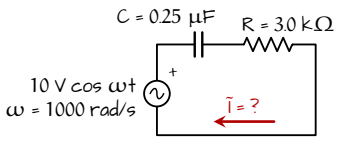


Figure 19.10 The circuit of Example 19.2.

Example 19.2

A source of the voltage $V_0 \cos \omega t$, with $V_0 = 10\,\text{V}$ and $\omega = 1000\,\text{rad/s}$, drives a capacitor of $C = 0.25\,\mu\text{F}$ and a resistor of $R = 3.0\,\text{k}\Omega$ in series. What current does the source supply? What are the voltages across the capacitor and across the resistor?

Solution:



The first step is a sketch of the circuit, with the value of each element and the current we want. The capacitor and the resistor are in series, so the same current flows through both, and their impedances add.

The capacitor's impedance is

$$\begin{aligned} \tilde{Z}_C &= -\frac{i}{\omega C} \\ &= -\frac{i}{(1000\,\text{s}^{-1})(0.25 \times 10^{-6}\,\text{F})} \\ &= -4.0i\,\text{k}\Omega, \end{aligned}$$

so the circuit's impedance is $\tilde{Z} = R + \tilde{Z}_C = (3.0 - 4.0i)\,\mathrm{k\Omega}$. In polar form, $|\tilde{Z}| = \sqrt{3.0^2 + 4.0^2}\,\mathrm{k\Omega} = 5.0\,\mathrm{k\Omega}$, and $\tilde{Z}$ is in the fourth quadrant, at $\phi = -\arctan(4.0/3.0) = -53.13°$. The current is

$$\begin{aligned}\tilde{I} &= \frac{\tilde{V}}{\tilde{Z}} \\ &= \frac{10\,\mathrm{V}}{5.0\,\mathrm{k\Omega}\angle -53.13°} \\ &= 2.0\,\mathrm{mA}\angle 53.13°,\end{aligned}$$

or $I(t) = (2.0\,\mathrm{mA})\cos(\omega t + 53.13°)$: the current leads the source's voltage by 53.13°, as it must when the impedance's phase is negative.

The voltages follow from (19.23):

$$\begin{aligned}\tilde{V}_R &= R\tilde{I} = 6.0\,\mathrm{V}\angle 53.13°, \\ \tilde{V}_C &= \tilde{Z}_C\tilde{I} \\ &= (4.0\,\mathrm{k\Omega}\angle -90°)(2.0\,\mathrm{mA}\angle 53.13°) \\ &= 8.0\,\mathrm{V}\angle -36.87°,\end{aligned}$$

so $V_R(t) = (6.0\,\mathrm{V})\cos(\omega t + 53.13°)$ and $V_C(t) = (8.0\,\mathrm{V})\cos(\omega t - 36.87°)$.

Reality check: The amplitudes 6.0 V and 8.0 V add up to more than the source's 10 V, but the two voltages are 90° out of phase, so they are never largest at the same time. It is their complex amplitudes that add up to the source's:

$$\tilde{V}_R + \tilde{V}_C = (3.6 + 4.8i)\,\mathrm{V} + (6.4 - 4.8i)\,\mathrm{V} = 10\,\mathrm{V}.$$

Example 19.3

A source of the voltage $V_0 \cos\omega t$ drives a capacitor C, a resistor R and an inductor L in series. Find the current, the voltage across the resistor, and the voltage across the inductor as functions of time. At what frequency is the current largest?

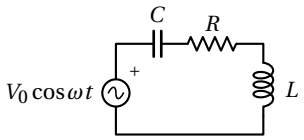
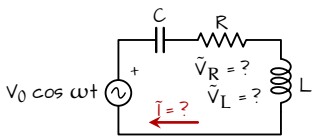


Figure 19.11 The series circuit of Example 19.3.

Solution:



The first step is a sketch of the circuit, with the current that all three elements share and the two voltages we want.

The three elements are in series, so their impedances add:

$$\tilde{Z} = R + i\omega L + \frac{1}{i\omega C} = R + i\left(\omega L - \frac{1}{\omega C}\right).$$

Its magnitude and phase are

$$|\tilde{Z}| = \sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}, \qquad \phi = \arctan\left[\frac{1}{R}\left(\omega L - \frac{1}{\omega C}\right)\right].$$

The real part of $\tilde{Z}$ is R, which is positive, so $\tilde{Z}$ is in the first or the fourth quadrant, and the arctangent gives the right angle, between $-90°$ and $90°$. The current is $\tilde{I} = \tilde{V}/\tilde{Z} = (V_0/|\tilde{Z}|)\angle -\phi$, or

$$I(t) = \frac{V_0}{|\tilde{Z}|}\cos(\omega t - \phi).$$

Reality check: With no inductor, $L = 0$, the circuit is that of Example 19.2, and the magnitude and the phase of $\tilde{Z}$ become $|\tilde{Z}| = \sqrt{R^2 + (1/\omega C)^2}$ and $\phi = -\arctan(1/\omega RC)$. With $R = 3.0\,\text{k}\Omega$ and $1/\omega C = 4.0\,\text{k}\Omega$ they are $5.0\,\text{k}\Omega$ and $-53.13°$, as found there.

The resistor's voltage, $\tilde{V}_R = R\tilde{I}$, is in phase with the current:

$$V_R(t) = \frac{V_0 R}{|\tilde{Z}|}\cos(\omega t - \phi).$$

For the inductor, $i = 1\angle 90°$, so

$$\begin{aligned}\tilde{V}_L &= i\omega L\tilde{I} \\ &= (\omega L\angle 90°)\frac{V_0}{|\tilde{Z}|}\angle -\phi \\ &= \frac{V_0\omega L}{|\tilde{Z}|}\angle(90° - \phi),\end{aligned}$$

and

$$V_L(t) = \frac{V_0\omega L}{|\tilde{Z}|}\cos(\omega t - \phi + 90°),$$

which leads the current by $90°$ (Figure 19.12).

The current is largest when $|\tilde{Z}|$ is smallest, which is when $\omega L = 1/\omega C$, at

$$\omega_0 = \frac{1}{\sqrt{LC}}.$$

This is the circuit's resonance. There $\tilde{Z} = R$, and the current, with the amplitude V_0/R, is in phase with the source. The voltages across the inductor and the capacitor are always opposite in phase. (Why are $\tilde{V}_L$ and $\tilde{V}_C$ always opposite in phase? Answer in footnote.[7]) At resonance they also have the same amplitude, so they cancel each other exactly.

The circuit is the driven oscillator of Example 18.2 in another form. With the charge q on the capacitor, $I = dq/dt$, the loop law reads

$$L\frac{d^2q}{dt^2} + R\frac{dq}{dt} + \frac{q}{C} = V_0\cos\omega t,$$

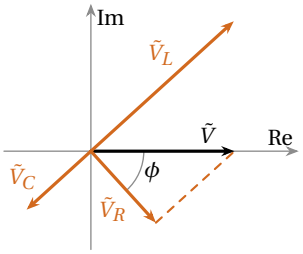


Figure 19.12 The complex amplitudes of the voltages in the circuit of Example 19.3 at a frequency above resonance ($\omega L = 2.0R$, $1/\omega C = 0.90R$). $\tilde{V}_L$ and $\tilde{V}_C$ point in opposite directions, at right angles to $\tilde{V}_R$, which points along the current; their sum (dashed) completes the source's $\tilde{V}$. The current lags the source by ϕ.

[7] The two elements carry the same current, and their impedances, $i\omega L$ and $-i/\omega C$, point in opposite directions along the imaginary axis. So $\tilde{V}_L = i\omega L\tilde{I}$ is at $+90°$ from the current and $\tilde{V}_C = -(i/\omega C)\tilde{I}$ at $-90°$, at every frequency.

which is the oscillator's equation with q for x, L for m, R/L for Γ, $1/LC$ for ω_0^2, and V_0 for F_0. The inductor gives the circuit its inertia, the capacitor acts as its spring, and the resistor damps it.

Tips for Mathematica and Python. A computer does the steps of Section 19.5.3 and avoids the quadrant trap of Section 18.1.3. Mathematica's `Abs` and `Arg`, and Python's `abs` and `numpy.angle`, give the magnitude and the phase of an impedance directly (Section 18.5.1), and Mathematica's `ComplexExpand` splits an impedance into its resistance and its reactance. A circuit that the series and parallel rules cannot reduce is solved from Kirchhoff's laws, one linear equation for each junction and each loop, with complex coefficients: Mathematica's `Solve` and `LinearSolve` and Python's `numpy.linalg.solve` handle complex coefficients just as they do real ones. To see how a circuit responds at all frequencies, compute $\tilde{I}$ for an array of values of ω at once, and plot its magnitude and its phase.

19.6 Power in AC circuits

The power that a source delivers to a circuit, or that a part of the circuit takes in, is $P(t) = V(t)I(t)$, the product of the real voltage and the real current. As Section 18.4 showed, it must be found from the real quantities, never as a product of complex amplitudes. With $V(t) = V_0 \cos\omega t$ and $I(t) = I_0 \cos(\omega t - \phi)$, where ϕ is the phase of the circuit's impedance, the power oscillates at twice the frequency, and what usually matters is its average over a cycle. By (18.16), with $\tilde{V} = V_0$ and $\tilde{I} = I_0 e^{-i\phi}$,

$$\begin{aligned} \langle P \rangle &= \frac{1}{2}\mathrm{Re}\left[\tilde{V}\tilde{I}^*\right] \\ &= \frac{1}{2}V_0 I_0 \cos\phi \\ &= V_{\mathrm{rms}} I_{\mathrm{rms}} \cos\phi, \end{aligned} \tag{19.26}$$

where $V_{\mathrm{rms}} = V_0/\sqrt{2}$ and $I_{\mathrm{rms}} = I_0/\sqrt{2}$ are the rms values of Section 18.4. The factor $\cos\phi$ is the circuit's **power factor**, and ϕ itself is sometimes called the power angle. With $\tilde{V} = \tilde{Z}\tilde{I}$, the average power can also be written

$$\langle P \rangle = \frac{1}{2}|\tilde{I}|^2 \mathrm{Re}[\tilde{Z}] = \frac{V_0^2}{2|\tilde{Z}|}\cos\phi\,: \tag{19.27}$$

only the resistance, the real part of the impedance, takes in power on average.

An ideal inductor or capacitor by itself has $\phi = \pm 90°$, and it takes in no power on average: it stores energy for a quarter of a cycle, as $\frac{1}{2}LI^2$ or $\frac{1}{2}CV^2$, and gives it back during the next quarter (Figure 19.13). In a resistor $\phi = 0$, and $\langle P \rangle = V_{\mathrm{rms}} I_{\mathrm{rms}} = I_{\mathrm{rms}}^2 R$, which is (19.11) with rms values in place of steady ones. That is why alternating voltages and currents are usually quoted as rms values. A household outlet in the United States, rated at 120 V, has the amplitude 170 V (Figure 18.14), and it delivers the same average power to a resistor as a steady

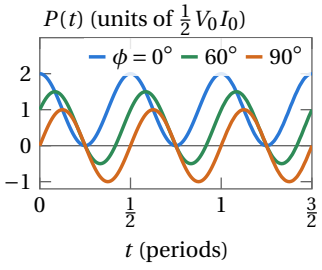


Figure 19.13 The power $P(t) = V(t)I(t)$ delivered to a load with the power factor $\cos\phi$, for $\phi = 0$ (a resistor), 60°, and 90° (an ideal inductor). Its average is $\cos\phi$ in these units.

120 V would. A load whose power factor is less than 1 draws more current for the same average power than a resistor would. The extra current carries energy back and forth between the source and the load, and it heats the wires that connect them without delivering any energy on average.

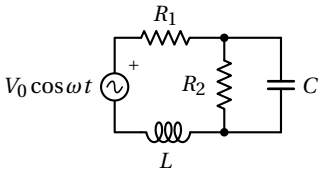
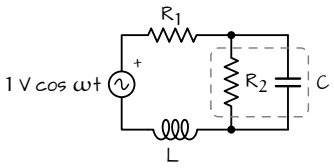


Figure 19.14 The circuit of Example 19.4.

Example 19.4

In the circuit of Figure 19.14, $V_0 = 1.0\,\text{V}$, $R_1 = 10\,\Omega$, $R_2 = 20\,\Omega$, $C = 0.10\,\mu\text{F}$ and $L = 0.10\,\text{mH}$. Find the average power that the source supplies, as a function of ω, and plot it.

Solution:



$R_1 = 10\ \Omega$, $R_2 = 20\ \Omega$,
$C = 0.1\ \mu\text{F}$, $L = 0.1\ \text{mH}$
$\langle P \rangle$ as a function of ω = ?

The first step is a sketch of the circuit, with its values and the way its elements are grouped: R_2 and C are in parallel, and that pair is in series with R_1 and L.

The circuit's impedance is

$$\tilde{Z} = R_1 + \left(R_2 // \tilde{Z}_C\right) + i\omega L = R_1 + \left(\frac{1}{R_2} + i\omega C\right)^{-1} + i\omega L,$$

and by (19.27), with $|\tilde{I}| = V_0/|\tilde{Z}|$,

$$\langle P \rangle = \frac{1}{2}|\tilde{I}|^2 \,\text{Re}[\tilde{Z}] = \frac{V_0^2 \,\text{Re}[\tilde{Z}]}{2|\tilde{Z}|^2}.$$

The algebra of writing out $\text{Re}[\tilde{Z}]$ and $|\tilde{Z}|^2$ is best left to a computer, which can also plot the result (Figure 19.15).

Reality check: The limits make sense. As $\omega \to 0$, the capacitor carries no current and the inductor has no impedance, so $\tilde{Z} = R_1 + R_2 = 30\,\Omega$, and $\langle P \rangle = V_0^2/2(R_1 + R_2) = 16.7\,\text{mW}$. At high frequencies the capacitor shorts out R_2, and the inductor's impedance ωL grows without limit, so the current, and with it the power, falls toward zero.

Reality check: Take one frequency, $\omega = 5.0 \times 10^5\,\text{rad/s}$. There $\omega C = 0.050\,\text{S} = 1/R_2$, so $R_2 // \tilde{Z}_C = 1/[(0.050 + 0.050i)\,\text{S}] = (10 - 10i)\,\Omega$, and $\omega L = 50\,\Omega$:

$$\tilde{Z} = (20 + 40i)\,\Omega, \qquad |\tilde{Z}|^2 = 2000\,\Omega^2, \qquad \langle P \rangle = \frac{(1.0\,\text{V})^2(20\,\Omega)}{2(2000\,\Omega^2)} = 5.0\,\text{mW},$$

with the power factor $\cos\phi = 20/\sqrt{2000} = 0.45$. All of this power goes into the two resistors: R_1 takes in $\frac{1}{2}|\tilde{I}|^2 R_1 = 2.5\,\text{mW}$, and R_2, across which the complex amplitude of the voltage is $\tilde{I}(10 - 10i)\,\Omega$, takes in $\frac{1}{2}|\tilde{I}|^2(200\,\Omega^2)/R_2 = 2.5\,\text{mW}$, which add up to 5.0 mW.

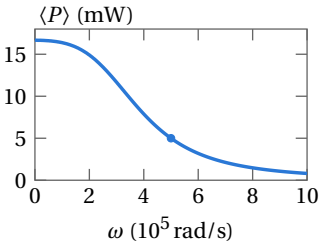


Figure 19.15 The average power that the source of Example 19.4 supplies, against the frequency. The dot marks $\omega = 5.0 \times 10^5\,\text{rad/s}$.

A real source of alternating voltage, such as a signal generator or the output of an amplifier, behaves like an ideal source in series with an internal impedance $\tilde{Z}_s$, and the average power it delivers to a load of impedance $\tilde{Z}_L$ depends on $\tilde{Z}_L$. It is largest when the load's impedance is the complex conjugate of the source's, $\tilde{Z}_L = \tilde{Z}_s^*$, so that the reactances cancel and the resistances are equal (Exercise P19.15). Choosing a load, or a network between the source and the load, to meet this condition is called **impedance matching**. A transmission line needs a match of a different kind, a load equal to the line's characteristic impedance, so that nothing is reflected (Section 23.8.3).

19.7 Transfer functions and filters

Circuits are often used to change a signal: to pass some frequencies and block others, or to shift a phase. A circuit of this kind has an input, where a voltage with the complex amplitude $\tilde{V}_{\text{in}}$ is applied, and an output, where a voltage $\tilde{V}_{\text{out}}$ appears. The ratio of the two is the circuit's **transfer function**,

$$\tilde{H}(\omega) = \frac{\tilde{V}_{\text{out}}}{\tilde{V}_{\text{in}}}. \tag{19.28}$$

For each ω it is a complex number. Its magnitude $|\tilde{H}|$ is the ratio of the amplitude of the output to that of the input, and its phase is the phase of the output minus the phase of the input.

19.7.1 The voltage divider

Most of the important filters are voltage dividers: an impedance $\tilde{Z}_1$ in series with the input, and an impedance $\tilde{Z}_2$ across the output (Figure 19.16). If nothing draws current from the output, the same current $\tilde{I} = \tilde{V}_{\text{in}}/(\tilde{Z}_1 + \tilde{Z}_2)$ flows through both impedances, and $\tilde{V}_{\text{out}} = \tilde{Z}_2\tilde{I}$, so

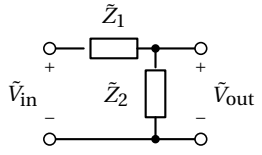


Figure 19.16 A voltage divider. The input is at the left and the output at the right; the bottom wire is common to both.

$$\tilde{H} = \frac{\tilde{Z}_2}{\tilde{Z}_1 + \tilde{Z}_2}. \tag{19.29}$$

With two resistors, the output is the fixed fraction $R_2/(R_1 + R_2)$ of the input at every frequency. The condition matters: whatever is connected to the output must have an impedance much larger than $|\tilde{Z}_2|$, so that the current it draws is negligible (Exercise P19.16 shows what happens otherwise).

19.7.2 High-pass and low-pass filters

Make $\tilde{Z}_1$ a capacitor and $\tilde{Z}_2$ a resistor, as in the following example.

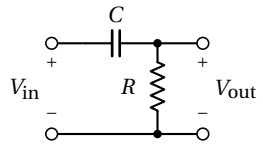


Figure 19.17 The circuit of Example 19.5.

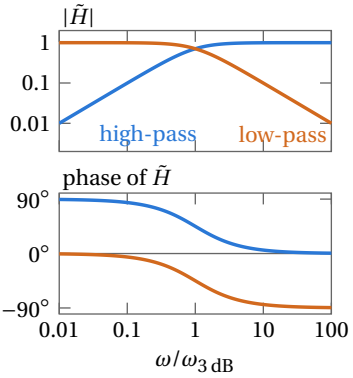
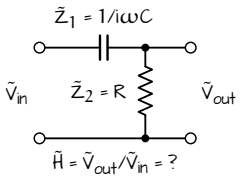


Figure 19.18 Bode plots of the high-pass filter of Example 19.5 (blue) and of the low-pass filter described below it (orange): the magnitude of $\tilde{H}$ on logarithmic axes (top), and its phase (bottom), against $\omega/\omega_{3\,dB}$.

Example 19.5

In the circuit of Figure 19.17, a capacitor C is in series with the input, and the output is taken across a resistor R. Find the transfer function. At what frequency is $|\tilde{H}| = 1/\sqrt{2}$, and what is the phase of $\tilde{H}$ there? For $R = 1.0\,\text{k}\Omega$ and $C = 0.16\,\mu\text{F}$, by what factor does the circuit reduce the amplitude of a signal at 60 Hz?

Solution:



The first step is a sketch of the circuit as a voltage divider: the capacitor is $\tilde{Z}_1$ and the resistor is $\tilde{Z}_2$.

By (19.29), with $\tilde{Z}_1 = 1/i\omega C$ and $\tilde{Z}_2 = R$,

$$\tilde{H} = \frac{R}{R + 1/i\omega C} = \frac{i\omega RC}{1 + i\omega RC}.$$

The product RC has the units of time, so $1/RC$ is a frequency. Call it $\omega_{3\,dB} = 1/RC$; then

$$\tilde{H} = \frac{i\omega/\omega_{3\,dB}}{1 + i\omega/\omega_{3\,dB}}.$$

At $\omega = \omega_{3\,dB}$,

$$\begin{aligned}\tilde{H} &= \frac{i}{1+i} \\ &= \frac{i(1-i)}{2} \\ &= \frac{1+i}{2} \\ &= \frac{1}{\sqrt{2}}\angle 45°,\end{aligned}$$

so $|\tilde{H}| = 1/\sqrt{2}$ at $\omega_{3\,dB}$, where the output leads the input by 45°. At low frequencies, $\omega \ll \omega_{3\,dB}$, $\tilde{H} \approx i\omega/\omega_{3\,dB}$: the output is small, and it leads the input by nearly 90°. At high frequencies $\tilde{H}$ approaches 1. The circuit passes high frequencies and blocks low ones: it is a high-pass filter (Figure 19.18, blue).

For $R = 1.0\,\text{k}\Omega$ and $C = 0.16\,\mu\text{F}$, $\omega_{3\,dB} = 6250\,\text{rad/s}$, or $\nu_{3\,dB} = \omega_{3\,dB}/2\pi = 995\,\text{Hz}$. A 60 Hz signal has $\omega/\omega_{3\,dB} = 60/995 = 0.0603$, and

$$|\tilde{H}| = \frac{0.0603}{\sqrt{1 + 0.0603^2}} = 0.060:$$

the circuit reduces the amplitude of the 60 Hz signal by a factor of 17, while signals above about 5 kHz pass with their amplitudes reduced by less than 2 percent.

Reality check: The limits of $\tilde{H}$ also follow from the circuit itself. As $\omega \to 0$, the capacitor's impedance $1/\omega C$ grows without limit, so no current flows and the output is zero.

As $\omega \to \infty$, its impedance goes to zero, the capacitor acts as a wire that joins the output to the input, and $\tilde{H} \to 1$.

Where does the name $\omega_{3\,\mathrm{dB}}$ come from? Power goes as the square of an amplitude, so at $\omega_{3\,\mathrm{dB}}$ the output delivers half of the power that it would deliver with $\tilde{H} = 1$. Ratios of powers are often given in **decibels**: the ratio P_2/P_1 is $10\log_{10}(P_2/P_1)$ decibels, written dB. Power goes as the square of a voltage, so a ratio of voltage amplitudes $|\tilde{H}|$ is $20\log_{10}|\tilde{H}|$ in decibels. Half of the power is $10\log_{10}(1/2) = -3.01\,\mathrm{dB}$, and so $\omega_{3\,\mathrm{dB}}$ is called the **3 dB point**. The response of a filter is usually plotted on logarithmic axes: $|\tilde{H}|$, or $20\log_{10}|\tilde{H}|$, and the phase of $\tilde{H}$, against $\log\omega$. This is a **Bode plot**[8] (Figure 19.18). On such axes the response of the high-pass filter is nearly two straight lines: $|\tilde{H}| \approx \omega/\omega_{3\,\mathrm{dB}}$ below the 3 dB point, which rises by a factor of 10, or 20 dB, for each factor of 10 in frequency, and $|\tilde{H}| \approx 1$ above it.

Interchange the two elements, with the resistor in series with the input and the capacitor across the output, and the circuit becomes a **low-pass filter**: it passes low frequencies and blocks high ones, with the same 3 dB point $\omega_{3\,\mathrm{dB}} = 1/RC$ (Exercise P19.16).

19.7.3 The band-pass filter

A filter can also pass a band of frequencies and block the rest. In the circuit of Figure 19.19, a resistor R is in series with the input, and the output is taken across an inductor L and a capacitor C in parallel. The impedance of the pair is

$$\tilde{Z}_L /\!/ \tilde{Z}_C = \frac{i\omega L}{1-\omega^2/\omega_0^2}, \qquad \omega_0 = \frac{1}{\sqrt{LC}}$$

(Exercise P19.17), and the voltage divider (19.29) gives

$$\tilde{H} = \frac{i\omega L}{R\left(1-\omega^2/\omega_0^2\right) + i\omega L}. \tag{19.30}$$

At $\omega = \omega_0$ the first term of the denominator vanishes, and $\tilde{H} = 1$: the output equals the input. Away from ω_0 the pair has a small impedance, since the inductor nearly shorts out the output at low frequencies and the capacitor does so at high ones, and the output is small. This is a **band-pass filter**. Figure 19.20 shows its response for $R = 100\,\mathrm{k\Omega}$, $C = 10\,\mathrm{nF}$ and $L = 10\,\mathrm{mH}$, for which $\omega_0 = 1.0 \times 10^5\,\mathrm{rad/s}$: $|\tilde{H}|$ rises to 1 in a narrow peak at ω_0, and the phase swings from $+90°$ to $-90°$ across it. How narrow the peak is depends on R (Exercise P19.17).

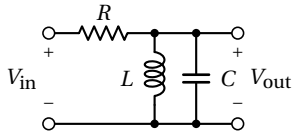


Figure 19.19 A band-pass filter: a resistor in series with the input, and an inductor and a capacitor in parallel across the output.

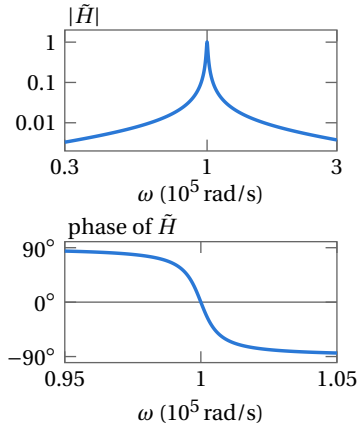


Figure 19.20 The response of the band-pass filter of Figure 19.19 with $R = 100\,\mathrm{k\Omega}$, $C = 10\,\mathrm{nF}$ and $L = 10\,\mathrm{mH}$: the magnitude of $\tilde{H}$ on logarithmic axes (top), and its phase near $\omega_0 = 1.0 \times 10^5\,\mathrm{rad/s}$ (bottom, on a linear axis).

19.8 Summary

Matter responds to fields: it polarizes and magnetizes, and a conductor carries a current. This chapter completed Maxwell's equations in matter for changing

[8] Named for Hendrik Bode, who devised it at Bell Telephone Laboratories in the 1930s.

fields, described conduction by Ohm's law and Drude's scattering model, treated inductance and the energy of the magnetic field, and solved circuits driven by sinusoidal sources with complex impedances.

Maxwell's equations in matter

$\mathbf{J}_p = \partial\mathbf{P}/\partial t$	the polarization current: bound charge in motion
$\nabla\cdot\mathbf{D} = \rho_f, \quad \nabla\times\mathbf{H} = \mathbf{J}_f + \partial\mathbf{D}/\partial t$	with $\nabla\cdot\mathbf{B} = 0$ and Faraday's law; $\partial\mathbf{D}/\partial t$ includes the polarization current
$\mathbf{D} = \varepsilon\mathbf{E}, \quad \mathbf{B} = \mu\mathbf{H}, \quad \mathbf{J}_f = \sigma\mathbf{E}$	constitutive relations of linear materials
the boundary conditions of Table 19.1	they hold at every instant when the fields change

Conduction

$\mathbf{J} = \sigma\mathbf{E}, \quad \rho = 1/\sigma$	Ohm's law, a constitutive relation that metals obey
$\mathbf{J} = -Ne\mathbf{v}_d$	the conduction electrons drift opposite to the field
$R = \rho L/A, \quad \Delta V = IR$	a uniform wire
$\mathbf{J}\cdot\mathbf{E} = \sigma E^2, \quad P = I\,\Delta V = I^2R$	ohmic heating
$\nabla^2 V = 0$ inside, $\quad \partial V/\partial n = 0$ on insulating surfaces	a steady current in a conductor of any shape; relaxation with mirror images at the edges
$m\,d\mathbf{v}/dt = -e\mathbf{E} - (m/\tau)\mathbf{v}$, $\sigma_0 = Ne^2\tau/m$	Drude's scattering model; τ the damping time

Inductance

$V_L = L\,dI/dt$	an inductor resists a change in its current
$I(t) = (\mathscr{E}/R)\left(1 - e^{-Rt/L}\right)$	the LR circuit; time constant L/R
$W = \frac{1}{2}LI^2; \quad U = \dfrac{1}{2\mu_0}\displaystyle\int B^2\,d\tau$	the work needed to set up the currents, done against the induced field
$U = \frac{1}{2}L_1I_1^2 + MI_1I_2 + \frac{1}{2}L_2I_2^2$, $M^2 \le L_1L_2$	two circuits; coupling coefficient $k = M/\sqrt{L_1L_2}$
$V_2/V_1 = N_2/N_1$	an ideal transformer

Circuits

$\sum I = 0$ at a junction, $\sum V = 0$ around a loop	Kirchhoff's laws; they hold for complex amplitudes too
$\tilde{V} = \tilde{Z}\tilde{I}$: $\tilde{Z}_R = R$, $\tilde{Z}_C = 1/i\omega C$, $\tilde{Z}_L = i\omega L$	impedances, in the phase-lead convention (an assumed $e^{i\omega t}$)
$\tilde{Z}_1 + \tilde{Z}_2$, $\tilde{Z}_1 // \tilde{Z}_2 = \left(1/\tilde{Z}_1 + 1/\tilde{Z}_2\right)^{-1}$	in series, in parallel
$I(t) = (V_0/\lvert\tilde{Z}\rvert)\cos(\omega t - \phi)$	the current lags the source if the phase ϕ of $\tilde{Z}$ is positive
$\omega_0 = 1/\sqrt{LC}$	resonance of a series RLC circuit: $\tilde{Z} = R$
$\langle P\rangle = \frac{1}{2}\mathrm{Re}\left[\tilde{V}\tilde{I}^*\right] = \frac{1}{2}V_0 I_0 \cos\phi$	average power; power factor $\cos\phi$; only the resistance takes in power
$\tilde{H} = \tilde{V}_{\text{out}}/\tilde{V}_{\text{in}} = \tilde{Z}_2/(\tilde{Z}_1 + \tilde{Z}_2)$	transfer function of a voltage divider whose output draws no current
$\omega_{3\,\text{dB}} = 1/RC$, $20\log_{10}\lvert\tilde{H}\rvert$	RC filters; half the power at the 3 dB point

Exercises

Exercises for 19.1 The Ampère–Maxwell law in matter

Warm-up questions Read Section 19.1, then answer before class

W19.1 The field in the block of Figure 19.1 is switched on, held steady, and later switched off, so that the polarization returns to zero. Sketch the polarization current density J_p against the time for the whole sequence, as in Figure 19.2b. Which way does the polarization current flow while the polarization shrinks?

W19.2 For an oscillating field $E = E_0 \cos \omega t$, the amplitude of the polarization current grows in proportion to ω. Does the polarization current then carry more charge across each square meter of the block in each half cycle when ω is larger? (Take χ_e to be the same at both frequencies, and reason as the section does for a field that is switched on.)

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P19.1 Model a dielectric as N molecules per unit volume, each a charge $+q$ and a charge $-q$ separated by the vector $\mathbf{d}$, which points from $-q$ to $+q$. The midpoint of each molecule stays fixed.

(a) Show that the polarization is $\mathbf{P} = Nq\mathbf{d}$.

(b) When $\mathbf{d}$ changes, by stretching or by turning, each charge moves. Find the velocity of each charge in terms of $d\mathbf{d}/dt$, and show that together the charges carry the current density $\partial \mathbf{P}/\partial t$, (19.1).

(c) A molecule that turns without stretching has charges that circulate. Show that its two charges together have no magnetic dipole moment, so that turning molecules add nothing to the magnetization. (A point charge q at $\mathbf{r}$ moving with velocity $\mathbf{v}$ has the magnetic dipole moment $\frac{1}{2} q\, \mathbf{r} \times \mathbf{v}$.)

P19.2 Free charge and bound charge are each conserved separately.

(a) Take the divergence of the last of Maxwell's equations in matter, (19.2), and use the first to show that $\partial \rho_f / \partial t + \nabla \cdot \mathbf{J}_f = 0$.

(b) Show that the bound charge density $\rho_b = -\nabla \cdot \mathbf{P}$ and the bound current density $\nabla \times \mathbf{M} + \partial \mathbf{P}/\partial t$ obey a continuity equation of their own. Why does the magnetization current play no part in it?

P19.3 Two linear, isotropic materials meet at a flat interface (Figure 19.21): material 1 on the left, with ε_{r1} and μ_{r1}, and material 2 on the right, with $\varepsilon_{r2} = 2\varepsilon_{r1}$ and $\mu_{r2} = 2\mu_{r1}$. There is no free charge or free current on the interface, though there may be bound charge and bound current. At one instant the fields just to the left of the interface are $\mathbf{E}_1$ and $\mathbf{B}_1$, as shown.

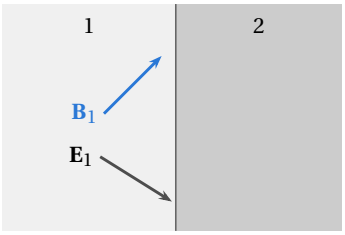


Figure 19.21 The interface of Exercise P19.3, with the fields just to the left of it at one instant.

(a) Which boundary conditions connect the components of $\mathbf{B}_1$ and $\mathbf{B}_2$, the field just to the right of the interface? Write them out, and say which of Maxwell's equations each comes from.

(b) Sketch $\mathbf{B}_2$ at that instant, using $\mathbf{B}_1$ for the scale.

(c) Do the same for $\mathbf{E}_2$.

P19.4 Suppose that near a surface at rest there is no free charge or current, and that the fields oscillate as $e^{-i\omega t}$, with $\omega \neq 0$ (Section 18.2.4).

(a) Use Faraday's law to show that if $\mathbf{E}^{\parallel}$ is continuous everywhere on the surface, then so is $B^{\perp}$. (The component of $\nabla \times \mathbf{E}$ along $\hat{\mathbf{n}}$ involves only derivatives of the parallel components of $\mathbf{E}$ in directions along the surface.)

(b) In the same way, show that if $\mathbf{H}^{\parallel}$ is continuous, so is $D^{\perp}$.

(c) So for such fields only the two conditions on the parallel components are independent. What is different for static fields?

Exercises for 19.2 Ohm's law and conductivity

Warm-up questions Read Section 19.2, then answer before class

W19.3 Section 19.2.1 shows that a steady current leaves no charge inside a wire of uniform conductivity. Now a copper wire is joined end to end to an aluminum wire of the same cross section, and a steady current flows from the copper into the aluminum. Is there charge at the junction? Of which sign? (Is J the same on both sides? Is E?)

W19.4 By (19.10) the heating per unit volume is both σE^2, which grows with σ, and J^2/σ, which falls. A copper wire and an aluminum wire of the same length and cross section are connected to one battery (a) in series, (b) in parallel. In each case, which wire heats more? Which form of (19.10) settles each case, and why?

W19.5 In Figure 19.5, at what angle do the equipotentials (orange) meet the strip's insulating edges? At what angle do the current lines (blue) meet the electrodes? Explain each with the boundary conditions of Section 19.2.4.

Also: without looking back, summarize this section in a few sentences, and say what

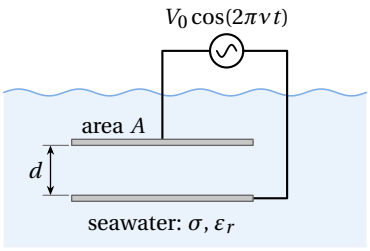


Figure 19.22 The capacitor of Exercise P19.5, submerged in seawater.

you found hardest to understand.

P19.5 Conduction current and displacement current. A parallel-plate capacitor, with plates of area A a distance d apart, operates submerged in seawater (Figure 19.22). The seawater between the plates has the conductivity σ and the relative permittivity ε_r, and the potential difference between the plates is $V_0\cos(2\pi\nu t)$. Ignore the fringing of the field at the edges of the plates.

(a) Find the current that the seawater conducts from one plate to the other, the potential difference divided by the resistance, as a function of time.

(b) Find the displacement current between the plates as a function of time: not just the displacement current density (16.3), but its integral over a plane between the plates, in amperes.

(c) Find the ratio of the amplitude of the conduction current to that of the displacement current.

(d) At 15°C, seawater with a salinity of 35 parts per thousand has a relative permittivity of about 71 at 1 GHz, and its conductivity is about 4 S/m.[9] Find the ratio at $\nu = 1$ GHz. What does your answer say about when the displacement current matters in seawater?

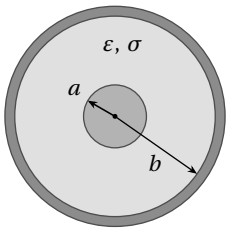


Figure 19.23 The coaxial cable of Exercise P19.6, in cross section.

P19.6 A long coaxial cable has an inner conductor of radius a and a thin outer conductor of radius b, separated by insulation with permittivity ε and a small conductivity σ (Figure 19.23). A potential difference ΔV is held between the conductors.

(a) Find the current that leaks from the inner conductor to the outer one through a length ℓ of the cable, and from it the leakage resistance R of that length.

(b) Find the capacitance C of the same length, and show that $RC = \varepsilon/\sigma$.

(c) Show that $RC = \varepsilon/\sigma$ for any two electrodes, of any shapes, surrounded by a uniform material with permittivity ε and conductivity σ. (Gauss's law, (9.25) with $\mathbf{D} = \varepsilon\mathbf{E}$, gives the charge on one electrode, and Ohm's law gives the current that leaves it, both from the same surface integral of $\mathbf{E}$.)

(d) The cable is charged and then disconnected from everything. How does the charge on it change with time?

[9] The permittivity (its real part; Section 22.4 explains why such a material also has an imaginary part) is read from Fig. 1(a) of "Frequency, temperature and salinity variation of the permittivity of seawater," R. Somaraju and J. Trumpf, IEEE Trans. Antennas Propag. 54, 3441–3448 (2006), which also takes the conductivity of seawater to be 4 S/m. https://doi.org/10.1109/TAP.2006.884290

P19.7 Section 19.2.4 stated that the potential in a conductor that carries a steady current is unique when V is given on the electrodes and $\partial V/\partial n = 0$ on the rest of the conductor's surface. Suppose that V_1 and V_2 both satisfy Laplace's equation in the conductor and both of these boundary conditions, and let $V_3 = V_1 - V_2$.

(a) What boundary conditions does V_3 satisfy?

(b) Integrate $\nabla\cdot(V_3\nabla V_3)$ over the conductor, as in the proof of the second uniqueness theorem (Section 7.1.4), and show that $\int |\nabla V_3|^2\, d\tau = 0$.

(c) Conclude that $V_1 = V_2$. Where does the proof need at least one electrode?

P19.8 A strip of resistive film with sheet resistance R_s narrows abruptly from the width $2w$ to the width w halfway along its length (Figure 19.24). Each part is $3w$ long, the two parts are centered on the same line, and electrodes cover the two ends.

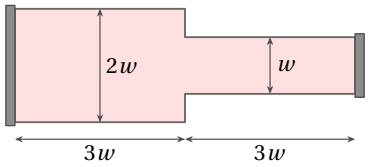


Figure 19.24 The strip of resistive film of Exercise P19.8, with electrodes across its ends.

(a) Estimate the resistance between the electrodes by treating each part as a separate straight strip, with the current spread uniformly across its width.

(b) Find the resistance with the relaxation method of Section 19.2.4, making the grid finer until the result changes by less than 0.1 percent. Plot the equipotentials, and describe where the current crowds together.

(c) Your answer to (b) should be larger than your estimate in (a). Explain why the estimate must be too low. (Imagine thin sheets of a perfect conductor inserted across the strip at closely spaced intervals. What do they do to the equipotentials? Can adding conductors ever increase a resistance?)

Exercises for 19.3 Microscopic view of conduction

Warm-up questions Read Section 19.3, then answer before class

W19.6 Sections 19.2 and 19.3 compare the conduction electrons of a metal with a skydiver in free fall. In what ways are the two alike? In what ways are they different?

W19.7 Silver has the largest conductivity in Table 19.2, and a classmate concludes that silver must have the most conduction electrons per unit volume. But aluminum has about three times as many per unit volume as silver (Example 19.1 gives aluminum's N). What does (19.15) say must be different about the conduction electrons of silver?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P19.9 The Hall effect in Drude's scattering model. (Exercise P10.10 found the Hall voltage from a balance of forces.) With a magnetic field present, (19.13) becomes

$$m\frac{d\mathbf{v}}{dt} = -e(\mathbf{E}+\mathbf{v}\times\mathbf{B}) - \frac{m}{\tau}\mathbf{v}.$$

A long strip of metal carries a current along its length, the x direction, so that $\mathbf{J}$ has only an x component, in a uniform field $\mathbf{B} = B\hat{\mathbf{z}}$, as in Figure 10.23.

(a) Show that in the steady state $\mathbf{E}$ has both an x and a y component, each proportional to J, and find them in terms of the DC conductivity σ_0, (19.15), and the cyclotron frequency $\omega_c = eB/m$.

(b) So $\mathbf{E}$ is no longer parallel to $\mathbf{J}$. Find the angle between them, and show that it is small when $\omega_c\tau \ll 1$.

(c) Show that E_y/JB does not depend on τ, so that it does not depend on how often the electrons collide. What does it depend on?

Exercises for 19.4 Inductance

Warm-up questions Read Section 19.4, then answer before class

W19.8 While the current in Figure 19.6a grows, the end of the inductor where it enters is at the higher potential. Which end is at the higher potential while the current, still flowing the same way, decreases? Explain with (19.16).

W19.9 Section 19.4.2 asks you to compare the energy $\frac{1}{2}LI^2$ of an inductor with the energy $\frac{1}{2}CV^2$ of a capacitor. In what ways are an inductor and a capacitor alike? In what ways are they different? (Think of what each one resists, and of where each one stores its energy.)

W19.10 Section 19.4.3 builds up the current I_1 first and then I_2, and notes that building up I_2 first gives the same total energy only because $M_{12} = M_{21}$. Why must the two orders give the same energy? (Could you build the currents up in one order and bring them down in the other?)

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P19.10 Energy in the circuit of Figure 19.6.

(a) Using (19.17), find the energy that the battery has supplied and the energy that has become heat in the resistor between $t = 0$ and the time t. Show that their difference is $\frac{1}{2}LI^2$ at every instant.

(b) At what time is energy going into the inductor fastest? Give it in units of L/R.

(c) Long after the switch is closed, the battery is replaced by a wire, quickly enough that the current keeps flowing through the resistor and the inductor alone. Find $I(t)$ from then on, and the total energy that becomes heat in the resistor. Where did that energy come from?

P19.11 Two circular loops of radius a share the z axis: loop 1 lies in the plane $z = 0$ and loop 2 in the plane $z = h$. Both carry the same current I in the same sense. In this exercise you follow, numerically, where the energy goes when loop 2 is brought in from far away. Use the exact off-axis field of a loop given in Exercise P10.22, and give your answers in units of $\mu_0 a I^2$ for energies, $\mu_0 a$ for M, and $\mu_0 I^2$ for forces.

(a) Find the mutual inductance M for $h = a$ by integrating the field of loop 1 over the flat disk bounded by loop 2. Check your answer against the Neumann formula (12.31).

(b) The energy stored in the field is $(1/2\mu_0)\int B^2\,d\tau$, with $\mathbf{B} = \mathbf{B}_1 + \mathbf{B}_2$. Show that it is the sum of each loop's own field energy and the cross term

$$W_{12} = \frac{1}{\mu_0}\int \mathbf{B}_1 \cdot \mathbf{B}_2\,d\tau.$$

Each loop's own energy is infinite for a wire of zero thickness, but it does not change as the loops move, so only W_{12} matters here.

(c) Evaluate W_{12} numerically for $h = a$, and compare it with MI^2. Use cylindrical coordinates. The integrand is singular at each wire, but the singularity is integrable: split the integration at $s = a$, $z = 0$ and $z = h$, and let an adaptive integrator do the rest.

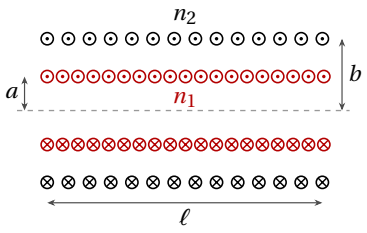


Figure 19.25 The two coaxial solenoids of Exercise P19.12, cut through their axis.

(d) Find the force on loop 2 from (10.14) as a function of h. Is it attractive or repulsive? Then find the work an outside agent does to bring loop 2 slowly from very far away to $h = a$, with both currents held fixed.

(e) Your answers to (c) and (d) have the same size and opposite signs. Use them to find the energy the two current sources supply while loop 2 is brought in. Explain, using Faraday's law, why the sources must supply energy at all, and why they supply equal shares.

P19.12 Two long solenoids of the same length ℓ share an axis. The inner one has radius a and n_1 turns per unit length, and the outer one has radius $b > a$ and n_2 turns per unit length (Figure 19.25). Treat both as long solenoids, ignoring the fields near their ends.

(a) Find L_1, L_2, and M. Find M from the flux of the outer solenoid's field through the turns of the inner one, and check it with the flux of the inner solenoid's field through the turns of the outer one.

(b) Show that the coupling coefficient is $k = a/b$, and explain why it is less than 1 even though all of the inner solenoid's flux passes through every turn of the outer one.

Exercises for 19.5 Complex impedance

Warm-up questions Read Section 19.5, then answer before class

W19.11 The source of Example 19.2 is switched on at $t = 0$, when its voltage is largest and the capacitor has no charge. What is the current just after it is switched on? Is that the value that the example's $I(t)$ gives at $t = 0$? If not, what makes up the difference?

W19.12 In the series circuit of Example 19.3, which element has the largest impedance at very low frequencies, and which at very high frequencies? In each limit, does the current lead or lag the source's voltage?

W19.13 At resonance the impedance of the series circuit of Example 19.3 is R, as if the inductor and the capacitor were not there. Do they still store energy? At the instant when the current is largest, where is the stored energy? Where is it a quarter period later? (Use Figure 19.9.)

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

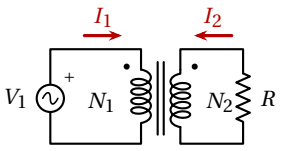


Figure 19.26 The transformer of Exercise P19.13. The currents are counted positive as drawn, into the ends of the coils marked with dots.

P19.13 A transformer with a load. A source drives the primary coil of a transformer, of N_1 turns, with the voltage $V_1 = V_0 \cos \omega t$, and the secondary coil, of N_2 turns, is connected to a resistor R (Figure 19.26). The coils have the self-inductances L_1 and L_2, the mutual inductance M, and no

resistance. The currents I_1 and I_2 are counted positive as drawn, so that positive currents in both coils make flux in the same direction in the core, and $M > 0$.

(a) Explain why the complex amplitudes obey $\tilde{V}_1 = i\omega L_1 \tilde{I}_1 + i\omega M \tilde{I}_2$ and $0 = i\omega M \tilde{I}_1 + (i\omega L_2 + R)\tilde{I}_2$.

(b) Find the impedance $\tilde{Z}_{\text{in}} = \tilde{V}_1 / \tilde{I}_1$ that the source sees.

(c) Section 19.4.4 showed that for an ideal transformer $M^2 = L_1 L_2$ and $L_2/L_1 = (N_2/N_1)^2$. Show that then $\tilde{Z}_{\text{in}} = i\omega L_1 R/(R + i\omega L_2)$, and that when $\omega L_2 \gg R$ the source sees a resistance $(N_1/N_2)^2 R$.

(d) In the same limit, find $|\tilde{I}_2/\tilde{I}_1|$, and the ratio of the amplitude of the voltage across the resistor to V_0. Then show, without taking the limit, that the average power that the source delivers equals the average power that the resistor takes in.

(e) What current does the primary draw when the resistor is disconnected? What average power does it take in then?

Exercises for 19.6 Power in AC circuits

Warm-up questions Read Section 19.6, then answer before class

W19.14 A circuit made only of resistors, capacitors and inductors can never deliver energy, on average, to the source that drives it. Use (19.27) to explain why the real part of its impedance can never be negative. In which quadrants of the complex plane can its impedance lie?

W19.15 At $\omega = 5.0 \times 10^5$ rad/s, the two resistors of Example 19.4 take in equal shares of the average power. As $\omega \to 0$, which resistor takes in the larger share, and by what factor? Which one takes in nearly all of the power at high frequencies?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P19.14 The resonance of the series circuit of Example 19.3 (Figure 19.11).

(a) Show that the average power delivered to the circuit is

$$\langle P \rangle = \frac{V_0^2}{2R} \frac{R^2}{R^2 + (\omega L - 1/\omega C)^2},$$

and that it is largest at $\omega_0 = 1/\sqrt{LC}$.

(b) Show that $\langle P \rangle$ falls to half of its largest value at two frequencies whose difference is $\Delta\omega = R/L$. Compare with Exercise P18.9, using the correspondence at the end of Example 19.3.

(c) The ratio $Q = \omega_0/\Delta\omega$ is the circuit's quality factor (Section 23.7.2 defines it for a cavity in another way). Show that $Q = \omega_0 L/R$, and that at resonance the voltage across the inductor has the amplitude QV_0. For $R = 10\,\Omega$, $L = 1.0\,\text{mH}$, $C = 1.0\,\text{nF}$ and $V_0 = 1.0\,\text{V}$, find ω_0, Q, and that amplitude.

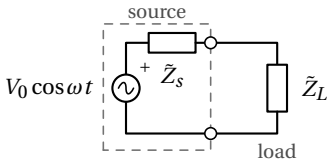


Figure 19.27 The source and the load of Exercise P19.15.

P19.15 Impedance matching. A source of alternating voltage behaves like an ideal source, whose voltage is $V_0 \cos\omega t$, in series with an internal impedance $\tilde{Z}_s = R_s + iX_s$. It drives a load of impedance $\tilde{Z}_L = R_L + iX_L$ (Figure 19.27).

(a) Take both impedances to be resistances, $\tilde{Z}_s = R_s$ and $\tilde{Z}_L = R_L$. Find the average power that the load takes in, and show that it is largest when $R_L = R_s$. What fraction of the average power that the ideal source supplies reaches the load then?

(b) Now let both be complex. Show that the load takes in the average power

$$\langle P_L \rangle = \frac{V_0^2 R_L}{2\left[(R_s + R_L)^2 + (X_s + X_L)^2\right]},$$

and that it is largest when $\tilde{Z}_L = \tilde{Z}_s^*$, not when $\tilde{Z}_L = \tilde{Z}_s$. (Choose X_L first, then R_L.)

(c) Check (b) numerically. For $\tilde{Z}_s = (50 + 30i)\,\Omega$ and $V_0 = 1.0\,\text{V}$, make a contour plot of $\langle P_L \rangle$ against R_L and X_L, and find its largest value and where it occurs.

(d) Power companies do not match the impedances of the loads on their lines to the impedance of the source, while radio transmitters are matched to their antennas with care. Why the difference?

Exercises for 19.7 Transfer functions and filters

Warm-up questions Read Section 19.7, then answer before class

W19.16 At low frequencies the output of the high-pass filter of Example 19.5 leads the input by nearly 90°. Explain why in terms of the current through the circuit, without using the formula for $\tilde{H}$.

W19.17 Suppose the capacitor were removed from the band-pass filter of Figure 19.19, leaving the resistor in series with the input and the inductor alone across the output. Would the circuit pass high frequencies or low ones? Explain with the voltage divider (19.29), without computing $\tilde{H}$.

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P19.16 The low-pass filter of Section 19.7.2 has a resistor R in series with the input and a capacitor C across the output (Figure 19.28).

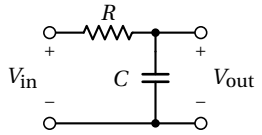


Figure 19.28 The low-pass filter of Exercise P19.16.

(a) Find its transfer function, and show that it can be written

$$\tilde{H} = \frac{1}{1 + i\omega/\omega_{3\,\text{dB}}}, \qquad \omega_{3\,\text{dB}} = \frac{1}{RC}.$$

(b) Find $|\tilde{H}|$ and the phase of $\tilde{H}$ at $\omega = \omega_{3\,\text{dB}}$, and show that well above $\omega_{3\,\text{dB}}$ the output falls by 20 dB for each factor of 10 in frequency. Compare with Figure 19.18.

(c) Now connect a resistor R_L across the output. Find the new transfer function. How large must R_L be for the result of (a) to hold within 1 percent at low frequencies?

P19.17 The band-pass filter of Figure 19.19.

(a) Show that the impedance of the inductor and the capacitor in parallel is $\tilde{Z}_L // \tilde{Z}_C = i\omega L/(1 - \omega^2/\omega_0^2)$, with $\omega_0 = 1/\sqrt{LC}$, and derive (19.30).

(b) What is the impedance of the pair at ω_0? If the output has the amplitude V_0 there, what are the complex amplitudes of the currents through the inductor and through the capacitor, taking the output's phase to be zero? What current flows through the resistor?

(c) Show that $|\tilde{H}| = 1/\sqrt{2}$ at two frequencies whose difference is $\Delta\omega = 1/RC$. For the values of Figure 19.20, find $\Delta\omega$ and the ratio $\omega_0/\Delta\omega$.

Exercises for 19.8 Summary

P19.18 True or false? Explain your reasoning.

(a) Ohm's law, $\mathbf{J} = \sigma\mathbf{E}$, holds for every material.

(b) Inside a wire that carries a steady current, the electric field is zero, as it is inside any conductor.

(c) The conduction electrons in a wire that carries a current drift along it at nearly the speed of light.

(d) The energy stored in an inductor that carries the current I is $\frac{1}{2}LI^2$, however quickly the current was built up.

(e) Two coils can have a mutual inductance larger than either of their self-inductances.

(f) The current through an inductor leads the voltage across it by 90°.

(g) In a series RLC circuit at its resonant frequency, the current is in phase with the source's voltage.

(h) The amplitudes of the voltages across two elements in series add up to the amplitude of the voltage across both.

(i) The average power delivered to an ideal capacitor is zero.

(j) The time average of the power is $\frac{1}{2}\mathrm{Re}[\tilde{V}\tilde{I}]$.

(k) At the 3 dB point of a filter, the amplitude of the output is half of the amplitude of the input.

P19.19 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as the relaxation method for the current in a conductor of any shape, the drift of electrons in Drude's model with collisions at random times, the coupling coefficient of two coils from the Neumann formula, the transient of a series RLC circuit when its source is switched on, the power factor of a load and how a capacitor can raise it, or the response of a filter made of several stages.

Chapter 20

Forces and Conservation Laws

20.1 Conservation of charge: the continuity equation

Electric charge is conserved: the total charge of an isolated system never changes. Charges can move around, and positive and negative charges can appear or disappear in pairs, as when a gamma ray passing near a nucleus creates an electron and a positron, but the net charge stays the same. No experiment has ever found an exception.

That statement is about the system as a whole. By itself it would allow charge to disappear from one place and reappear at the same instant somewhere far away, and charge does not do that. It is conserved *locally*: the charge inside any region can change only by flowing in or out through the region's boundary. Relativity in fact requires the local version (Section 25.3.1). Two events that are simultaneous for one observer are not simultaneous for another, so a charge that vanished here and reappeared there at the same instant would, according to some observers, be missing for a while.

To turn the local statement into an equation, take a fixed volume $\mathcal{V}$ whose boundary is the closed surface $\mathcal{S}$, with $d\mathbf{a}$ pointing outward as in the divergence theorem. The charge inside is

$$Q_{\text{enc}}(t) = \int_{\mathcal{V}} \rho \, d\tau. \tag{20.1}$$

The flow of charge is described by the current density $\mathbf{J}$ (Chapter 10): $\mathbf{J} \cdot d\mathbf{a}$ is the charge per unit time that crosses the area element $d\mathbf{a}$ in the direction of $d\mathbf{a}$. The rate at which charge leaves $\mathcal{V}$ is therefore the flux of $\mathbf{J}$ out through $\mathcal{S}$, and charge that leaves is charge the volume no longer has:

$$\frac{dQ_{\text{enc}}}{dt} = -\oint_{\mathcal{S}} \mathbf{J} \cdot d\mathbf{a}. \qquad \text{(continuity equation, integral form)} \tag{20.2}$$

With words, it says that the charge inside a closed surface decreases at exactly the rate at which current flows out through the surface. A capacitor plate being charged through a wire is the simplest example (Figure 20.1). Draw a closed

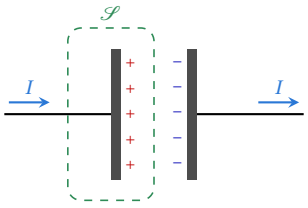


Figure 20.1 A closed surface (dashed) around one plate of a capacitor that is being charged. The current I enters through the wire, and no current crosses the surface anywhere else, so the charge inside grows at the rate $dQ_{\text{enc}}/dt = I$.

surface around one plate. The wire carries a current I in through the surface, and no current crosses the surface anywhere else, because none flows in the gap between the plates. The flux of $\mathbf{J}$ out of the surface is therefore $-I$, and the charge on the plate grows at the rate

$$\frac{dQ_{\text{enc}}}{dt} = I. \tag{20.3}$$

To get the differential form, write Q_{enc} as the integral of ρ. Because the volume does not move, the time derivative can be taken inside the integral, where it becomes a partial derivative. The divergence theorem, (1.36), turns the flux into a volume integral of $\nabla \cdot \mathbf{J}$, so (20.2) becomes

$$\int_{\mathcal{V}} \frac{\partial \rho}{\partial t}\, d\tau = -\int_{\mathcal{V}} \nabla \cdot \mathbf{J}\, d\tau. \tag{20.4}$$

This holds for every volume, however small, which is possible only if the two integrands are equal at every point:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J} = 0. \qquad \text{(continuity equation)} \tag{20.5}$$

This is the **continuity equation** for electric charge. (Older books, and Maxwell himself, call it the *equation of continuity*.) With words: wherever current spreads out from a point, so that $\nabla \cdot \mathbf{J} > 0$, the charge density there decreases, and wherever current converges, charge piles up.

When the charge density does not change with time, $\partial\rho/\partial t = 0$, and the continuity equation becomes

$$\nabla \cdot \mathbf{J} = 0. \qquad \text{(steady currents)} \tag{20.6}$$

As much current flows into any region as flows out of it. This is the condition on the currents of magnetostatics (Chapter 10), and it is the reason for Kirchhoff's junction rule: the currents flowing into a junction of wires add up to the currents flowing out.

Charge conservation does not have to be added to Maxwell's equations as a separate law; it is built into them. Take the divergence of the Ampère–Maxwell law, (2.4). The divergence of a curl is zero, (1.57), and Gauss's law, (2.1), replaces $\nabla \cdot \mathbf{E}$ with ρ/ε_0:

$$\begin{aligned} 0 &= \nabla \cdot (\nabla \times \mathbf{B}) \\ &= \mu_0 \nabla \cdot \mathbf{J} + \mu_0 \varepsilon_0 \frac{\partial}{\partial t}(\nabla \cdot \mathbf{E}) \\ &= \mu_0 \left(\nabla \cdot \mathbf{J} + \frac{\partial \rho}{\partial t} \right). \end{aligned} \tag{20.7}$$

This is (20.5) again (Exercise P2.1 asked you to show it). Without Maxwell's term $\mu_0\varepsilon_0\, \partial\mathbf{E}/\partial t$, the same steps would give $\nabla \cdot \mathbf{J} = 0$ even where charge is piling up, as it does on the capacitor plate. That contradiction is the reason the term must be there (Section 16.1).

One continuity equation of many. Any quantity that is conserved locally obeys an equation of the same form: the rate of change of its density, plus the divergence of its flux density, is zero,

$$\frac{\partial}{\partial t}(\text{density}) + \nabla \cdot (\text{flux density}) = 0, \tag{20.8}$$

where the flux density is the amount of the quantity that crosses a unit area per unit time. For electric charge they are ρ and $\mathbf{J}$. The mass density ρ_m of a fluid flowing with velocity $\mathbf{v}$ obeys

$$\frac{\partial \rho_m}{\partial t} + \nabla \cdot (\rho_m \mathbf{v}) = 0, \tag{20.9}$$

and in quantum mechanics the probability density $|\Psi|^2$ obeys a continuity equation whose flux density is called the probability current. If the quantity can be exchanged with something else, the right-hand side is not zero; it is the rate, per unit volume, at which the quantity is gained. The rest of this chapter derives two such equations for the electromagnetic field. In the one for energy (Section 20.2), the energy density of the field takes the place of ρ, the Poynting vector takes the place of $\mathbf{J}$, and the right-hand side is the negative of the rate, per unit volume, at which the field does work on charges, because energy the field gives to the charges is energy the field no longer has. The other is for momentum (Section 20.3). Charge is the simplest case: there is nothing for it to be exchanged with, so the right-hand side of its continuity equation is zero.

20.2 Energy conservation: Poynting's theorem

The fields store energy. In electrostatics the energy density is $\frac{1}{2}\varepsilon_0 E^2$, (5.24), and in magnetostatics it is $B^2/2\mu_0$ (Section 12.7). When the fields change, the energy also moves: a wave carries it across empty space (Section 16.5), and wherever the fields do work on charges, energy leaves the fields. This section finds the continuity equation for energy that Section 20.1 promised, and with it the rate at which the fields carry energy from place to place (Figure 20.2).

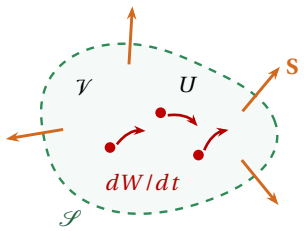


Figure 20.2 The energy bookkeeping of this section, for a fixed volume $\mathcal{V}$ inside the closed surface $\mathcal{S}$: the fields inside store the energy U, they do work at the rate dW/dt on the charges inside as they move them (red arrows), and they carry energy out through the surface, where the Poynting vector $\mathbf{S}$ (orange) crosses it.

20.2.1 The work done on charges

A charge q that moves with velocity $\mathbf{v}$ through the fields $\mathbf{E}$ and $\mathbf{B}$ feels the force $q(\mathbf{E}+\mathbf{v}\times\mathbf{B})$, (10.2). In a time dt it moves $d\boldsymbol{\ell} = \mathbf{v}\,dt$, and the force does the work

$$dW = q\,(\mathbf{E}+\mathbf{v}\times\mathbf{B})\cdot\mathbf{v}\,dt = q\,\mathbf{E}\cdot\mathbf{v}\,dt,$$

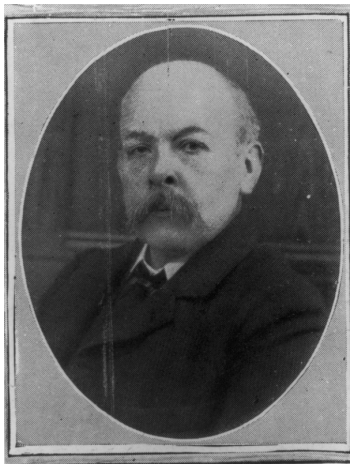
because $\mathbf{v}\times\mathbf{B}$ is perpendicular to $\mathbf{v}$: magnetic forces do no work (Section 10.1.3). For charge spread out with density ρ, let $\mathbf{v}$ stand for the velocity of the charges at each point, a vector field. The fields do work on the charge $\rho\,d\tau$ in a volume

element at the rate $\rho\,\mathbf{E}\cdot\mathbf{v}\,d\tau$, and $\rho\mathbf{v}=\mathbf{J}$, so the rate at which they do work on all the charges in a volume $\mathcal{V}$ is

$$\frac{dW}{dt}=\int_{\mathcal{V}}\mathbf{E}\cdot\mathbf{J}\,d\tau. \tag{20.10}$$

$\mathbf{J}$ here is a vector field describing all of the currents from all of the moving charges. If there are many charges, each with their own velocity, their contributions to $\mathbf{J}$ add together. With words, the fields give energy to moving charges at the rate $\mathbf{E}\cdot\mathbf{J}$ per unit volume. The energy may become kinetic energy of the charges or, in a resistor, heat. Where $\mathbf{E}\cdot\mathbf{J}$ is negative, as in a battery, where chemical reactions push the current against the electric field, the charges give energy to the fields.

20.2.2 Poynting's theorem



John Henry Poynting (1852–1914, English) Poynting was born in Monton, near Manchester. He studied at Owens College in Manchester and at Trinity College, Cambridge, where he graduated third wrangler in mathematics in 1876, and he did research under James Clerk Maxwell at the Cavendish Laboratory. In 1880 he became the first professor of physics at the new Mason Science College in Birmingham, which later became the University of Birmingham, and he stayed there for the rest of his life. Besides the theorem named for him, he measured the gravitational constant with a beam balance, studied the pressure of light, and in 1903 realized that sunlight slowly draws small particles orbiting the Sun inward, the Poynting–Robertson effect. (Wikipedia)

Where does that energy come from? Use the Ampère–Maxwell law, (2.4), to write the current density in terms of the fields, $\mathbf{J}=\nabla\times\mathbf{B}/\mu_0-\varepsilon_0\,\partial\mathbf{E}/\partial t$:

$$\mathbf{E}\cdot\mathbf{J}=\frac{1}{\mu_0}\mathbf{E}\cdot(\nabla\times\mathbf{B})-\varepsilon_0\,\mathbf{E}\cdot\frac{\partial\mathbf{E}}{\partial t}.$$

The product rule for the divergence of a cross product, (1.53), gives $\mathbf{E}\cdot(\nabla\times\mathbf{B})=\mathbf{B}\cdot(\nabla\times\mathbf{E})-\nabla\cdot(\mathbf{E}\times\mathbf{B})$, and Faraday's law, (2.3), replaces $\nabla\times\mathbf{E}$ with $-\partial\mathbf{B}/\partial t$. Since $\mathbf{B}\cdot\partial\mathbf{B}/\partial t=\frac{1}{2}\,\partial(B^2)/\partial t$, and likewise for $\mathbf{E}$,

$$\mathbf{E}\cdot\mathbf{J}=-\frac{\partial}{\partial t}\left(\frac{\varepsilon_0}{2}E^2+\frac{1}{2\mu_0}B^2\right)-\frac{1}{\mu_0}\nabla\cdot(\mathbf{E}\times\mathbf{B}).$$

The quantity in parentheses is the sum of the energy densities of the electric and magnetic fields,

$$u=\frac{\varepsilon_0}{2}E^2+\frac{1}{2\mu_0}B^2, \tag{20.11}$$

and the vector in the divergence is the **Poynting vector**,

$$\mathbf{S}=\frac{1}{\mu_0}\mathbf{E}\times\mathbf{B}. \tag{20.12}$$

With them the result is

$$\boxed{\frac{\partial u}{\partial t}+\nabla\cdot\mathbf{S}=-\mathbf{E}\cdot\mathbf{J}}. \tag{20.13}$$

This is **Poynting's theorem**. It has the form of the continuity equations of Section 20.1, with the energy density u in the place of the density and $\mathbf{S}$ in the place of the flux density. The right side is the negative of the rate, per unit volume, at which the fields give energy to the charges, since energy given to the charges is energy the fields no longer have.

For a fixed volume $\mathcal{V}$ bounded by the closed surface $\mathcal{S}$, integrate (20.13) over the volume, and turn the integral of $\nabla\cdot\mathbf{S}$ into a surface integral with the divergence theorem. With $U=\int_{\mathcal{V}}u\,d\tau$, the energy stored in the fields inside, and (20.10),

$$\frac{dW}{dt}=-\frac{dU}{dt}-\oint_{\mathcal{S}}\mathbf{S}\cdot d\mathbf{a}. \tag{20.14}$$

This is a statement of the conservation of energy. If the energy stored in the fields inside a volume decreases, that energy either goes into the work done on the charges inside, or else it is carried out of the volume through its surface. The last term is the energy that leaves through the surface per unit time, so **S** is the energy that the fields carry per unit time across a unit area, and it points in the direction in which the energy flows. Its units are watts per square meter. Nothing in the derivation assumed waves, or any particular time dependence: the theorem holds for any fields.

Only the flux of **S** through closed surfaces enters (20.14). Static fields can have $\mathbf{S} \neq 0$: a charge held near a magnet is surrounded by an electric and a magnetic field that are not parallel. But where nothing changes and no current flows, (20.13) gives $\nabla \cdot \mathbf{S} = 0$, so the energy that **S** describes only circulates (Figure 20.3), and as much flows into any region as flows out of it (Exercise P20.5). Such fields do carry momentum, though (Section 20.3.4).

John Henry Poynting published the theorem in 1884.[1] A few months later Oliver Heaviside found the same result independently, and he published it in 1885.[2]

! **No waves needed.** Whenever the Poynting vector has a component perpendicular to a surface, energy is transported across that surface whether or not waves are involved.

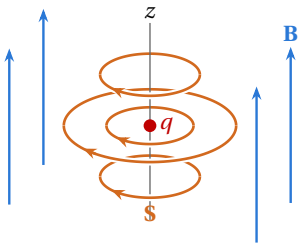


Figure 20.3 The Poynting vector (orange) around a positive charge at rest in a uniform magnetic field (blue) along z. It circles the z axis, clockwise as seen from above; it is strongest near the charge in the plane $z = 0$, and zero on the z axis.

20.2.3 Energy flow into a wire

The Poynting vector says where the energy goes in an ordinary circuit, and the answer is surprising at first.

Example 20.1

A straight wire of radius a and length ℓ carries a steady current I, spread uniformly over its cross section, and the potential difference between its ends is V_0. Find the Poynting vector at the surface of the wire and the rate at which energy flows into the wire through its surface, and compare that rate with the power $V_0 I$ that the wire turns into heat.

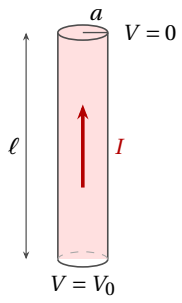


Figure 20.4 The wire of Example 20.1.

[1] "On the Transfer of Energy in the Electromagnetic Field," J. H. Poynting, Phil. Trans. R. Soc. Lond. 175, 343–361 (1884). https://doi.org/10.1098/rstl.1884.0016.

[2] "On the Transmission of Energy through Wires by the Electric Current," O. Heaviside, Electrician 14, 178 (1885), the second section of his series "Electromagnetic Induction and Its Propagation"; reprinted in O. Heaviside, *Electrical Papers*, Vol. 1 (Macmillan, London, 1892), pp. 434–441. https://archive.org/details/electricalpapers01heavuoft.

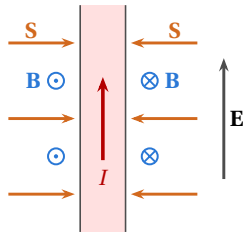
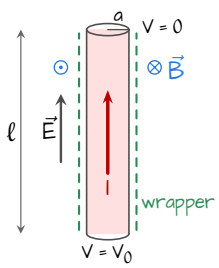


Figure 20.5 The Poynting vector at the surface of the wire of Example 20.1 points straight in, everywhere: the energy that the wire turns into heat enters through its sides. Animation of the energy flowing in from the space around the wire and turning into heat throughout it; none of it travels along the inside of the wire.

Solution:



The first step is a sketch of what is given and what is asked: the wire, with the current I flowing from the end at V_0 toward the end at zero, and $\mathbf{E}$ pointing the same way. $\mathbf{B}$ circles the wire. The surface through which the energy flows in is the wrapper of the wire (dashed).

The current is steady and uniform, so the electric field inside the wire is uniform and points along it, from the end at V_0 to the end at zero:

$$\mathbf{E} = \frac{V_0}{\ell}\hat{\mathbf{z}},$$

with $\hat{\mathbf{z}}$ along the current. The parallel component of $\mathbf{E}$ is continuous at the surface (Section 5.2), so just outside the wire the component of $\mathbf{E}$ along the wire is also V_0/ℓ.[3] The magnetic field at the surface follows from Ampère's law around a circle of radius a (Example 11.1): $\mathbf{B} = (\mu_0 I/2\pi a)\,\hat{\boldsymbol{\phi}}$. At the surface, then,

$$\begin{aligned}\mathbf{S} &= \frac{1}{\mu_0}\mathbf{E}\times\mathbf{B}\\ &= \frac{1}{\mu_0}\frac{V_0}{\ell}\frac{\mu_0 I}{2\pi a}\left(\hat{\mathbf{z}}\times\hat{\boldsymbol{\phi}}\right)\\ &= -\frac{V_0 I}{2\pi a\ell}\hat{\mathbf{s}},\end{aligned}$$

since $\hat{\mathbf{z}}\times\hat{\boldsymbol{\phi}} = -\hat{\mathbf{s}}$. The Poynting vector points straight into the wire, everywhere on its surface (Figure 20.5). The energy that flows in through the wrapper per unit time is $|\mathbf{S}|$ times the wrapper's area, $2\pi a\ell$:

$$P_{\text{in}} = \frac{V_0 I}{2\pi a\ell}\,(2\pi a\ell) = V_0 I.$$

This is exactly the power $V_0 I$ that the wire turns into heat. No energy flows in through the ends of the wire. (Why does no energy flow in through the ends of the wire? Answer in footnote.[4]) Inside the wire nothing changes with time, so $dU/dt = 0$, and (20.14) says that the work done on the charges, at the rate $V_0 I$, is the energy that flows in through the wrapper.

Reality check: The units of $V_0 I/2\pi a\ell$ are watts per square meter, and the energy flows in, as it must for a wire that turns it into heat.

[3] Just outside the wire, $\mathbf{E}$ also has a part perpendicular to the wire, from the charges on the wire's surface. That part tilts $\mathbf{S}$, so that some energy also flows parallel to the wire, but it adds nothing to the flux through the wrapper.

[4] Inside the wire $\mathbf{E}$ points along the wire, and $\mathbf{S} = \mathbf{E}\times\mathbf{B}/\mu_0$ is perpendicular to $\mathbf{E}$, so on the ends $\mathbf{S}$ is parallel to the surface and has no component along its normal.

The energy that heats the wire does not travel along the inside of the wire from the battery. Inside the battery, where $\mathbf{E}\cdot\mathbf{J}<0$, the same reasoning gives a Poynting vector that points out through the battery's sides. The energy leaves the battery through its sides, travels through the space around the circuit, and enters the wire through its sides.

20.2.4 Time averages

When the fields oscillate at a single frequency, $\mathbf{S}$ oscillates at twice that frequency, and what usually matters is its average over a cycle, which sets how much energy is delivered over many cycles. Write the fields with complex amplitudes, $\mathbf{E}(\mathbf{r},t)=\mathrm{Re}[\tilde{\mathbf{E}}(\mathbf{r})e^{-i\omega t}]$ and $\mathbf{B}(\mathbf{r},t)=\mathrm{Re}[\tilde{\mathbf{B}}(\mathbf{r})e^{-i\omega t}]$ (Section 18.2.4). Applying (18.16) to each product of components in $\mathbf{E}\times\mathbf{B}$ gives

$$\langle\mathbf{S}\rangle=\frac{1}{2\mu_0}\,\mathrm{Re}\left[\tilde{\mathbf{E}}\times\tilde{\mathbf{B}}^*\right]. \tag{20.15}$$

The size of $\langle\mathbf{S}\rangle$ is the **intensity** I, the average power per unit area that the fields carry. For the wave of Section 16.5, with $\mathbf{E}=E_0\cos(kz-\omega t)\,\hat{\mathbf{x}}$ and a magnetic field along $\hat{\mathbf{y}}$ with $B=E/c$, the complex amplitudes are $\tilde{\mathbf{E}}=E_0e^{ikz}\hat{\mathbf{x}}$ and $\tilde{\mathbf{B}}=(E_0/c)\,e^{ikz}\hat{\mathbf{y}}$, and

$$\langle\mathbf{S}\rangle=\frac{1}{2\mu_0}\,\frac{E_0^2}{c}\,\hat{\mathbf{z}}=\frac{1}{2}\,c\varepsilon_0E_0^2\,\hat{\mathbf{z}},$$

using $1/\mu_0=c^2\varepsilon_0$. Sunlight above Earth's atmosphere has an intensity of about $1361\,\mathrm{W/m^2}$.[5] A single wave of this kind with that intensity would have the amplitudes $E_0=\sqrt{2I/c\varepsilon_0}\approx 1.0\,\mathrm{kV/m}$ and $E_0/c\approx 3.4\,\mu\mathrm{T}$.

20.2.5 In linear materials

Inside matter the same steps can start from Maxwell's equations in matter, (19.2). The work that counts is now the work done on the free charges, at the rate $\mathbf{E}\cdot\mathbf{J}_f$ per unit volume; the work done in polarizing and magnetizing the material becomes part of the stored energy. With $\nabla\times\mathbf{H}=\mathbf{J}_f+\partial\mathbf{D}/\partial t$, and the product rule in the form $\mathbf{E}\cdot(\nabla\times\mathbf{H})=\mathbf{H}\cdot(\nabla\times\mathbf{E})-\nabla\cdot(\mathbf{E}\times\mathbf{H})$, the steps give

$$\mathbf{E}\cdot\mathbf{J}_f=-\mathbf{E}\cdot\frac{\partial\mathbf{D}}{\partial t}-\mathbf{H}\cdot\frac{\partial\mathbf{B}}{\partial t}-\nabla\cdot(\mathbf{E}\times\mathbf{H}).$$

In a linear material whose ε and μ do not change with time, $\mathbf{E}\cdot\partial\mathbf{D}/\partial t=\varepsilon\,\mathbf{E}\cdot\partial\mathbf{E}/\partial t=\partial(\frac{1}{2}\mathbf{E}\cdot\mathbf{D})/\partial t$, and in the same way $\mathbf{H}\cdot\partial\mathbf{B}/\partial t=\partial(\frac{1}{2}\mathbf{B}\cdot\mathbf{H})/\partial t$. So

$$\frac{\partial u}{\partial t}+\nabla\cdot\mathbf{S}=-\mathbf{E}\cdot\mathbf{J}_f,\qquad u=\frac{1}{2}\left(\mathbf{E}\cdot\mathbf{D}+\mathbf{B}\cdot\mathbf{H}\right),\qquad \mathbf{S}=\mathbf{E}\times\mathbf{H}. \tag{20.16}$$

[5] "A new, lower value of total solar irradiance: Evidence and climate significance," G. Kopp and J. L. Lean, Geophys. Res. Lett. 38, L01706 (2011). https://doi.org/10.1029/2010GL045777. Their value is $1360.8\pm0.5\,\mathrm{W/m^2}$.

In empty space these are (20.11) and (20.12) again, and in a material the energy density is the sum of the energy densities of a linear dielectric and a linear magnetic material, (9.39) and (13.35). For fields that oscillate at one frequency, the average is $\langle \mathbf{S} \rangle = \frac{1}{2}\text{Re}[\tilde{\mathbf{E}} \times \tilde{\mathbf{H}}^*]$, as in (20.15). The derivation assumed that ε and μ are constants, so that the material responds to the fields at once. For fields that oscillate quickly they depend on the frequency, and Section 22.4 takes up what changes.

20.3 Momentum conservation: the Maxwell stress tensor

Section 20.2 found the continuity equation for the energy of the fields. This section finds the one for their momentum, and with it a second way to calculate electromagnetic forces. The need for it first appears as a failure of Newton's third law.

20.3.1 Newton's third law and moving charges

Newton's third law says that two objects exert equal and opposite forces on each other. It is the reason that momentum is conserved in mechanics: the forces that the parts of a system exert on each other cancel in pairs, so only outside forces can change the system's total momentum.

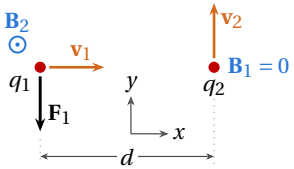


Figure 20.6 Two positive charges moving at right angles (velocities in orange). Straight ahead of q_1, where q_2 is, q_1 produces no magnetic field, so q_2 feels no magnetic force. The magnetic field of q_2 at q_1 points out of the page (blue), and the magnetic force on q_1 points along $-y$ (black).

Moving charges do not obey the law. Take two positive charges that move slowly at right angles to each other: q_1 is at the origin and moves along $+x$ with speed v_1, and q_2 is on the x axis at $x = d$, straight ahead of q_1, and moves along $+y$ with speed v_2 (Figure 20.6). The electric forces between them act along the line that joins them, the x axis, so neither has a y component.[6] The magnetic field of a slowly moving charge is approximately $(\mu_0/4\pi)\, q\mathbf{v} \times \hat{\boldsymbol{\imath}}/\imath^2$ (Section 10.5). At q_2, straight ahead of q_1, the vectors $\mathbf{v}_1$ and $\hat{\boldsymbol{\imath}}$ point the same way, so q_1 produces no magnetic field there, and q_2 feels no magnetic force at all. At q_1, though, $\hat{\boldsymbol{\imath}} = -\hat{\mathbf{x}}$, and the field of q_2 is

$$\mathbf{B}_2 = \frac{\mu_0}{4\pi}\frac{q_2 v_2 \hat{\mathbf{y}} \times (-\hat{\mathbf{x}})}{d^2} = \frac{\mu_0}{4\pi}\frac{q_2 v_2}{d^2}\hat{\mathbf{z}},$$

so q_1 feels the magnetic force

$$q_1 v_1 \hat{\mathbf{x}} \times \mathbf{B}_2 = -\frac{\mu_0}{4\pi}\frac{q_1 q_2 v_1 v_2}{d^2}\hat{\mathbf{y}}.$$

The forces that the two charges exert on each other do not add to zero: their y components add to $-(\mu_0/4\pi)\, q_1 q_2 v_1 v_2/d^2$. Left to themselves, the two charges would change their total momentum, with no outside force acting on them.

Momentum is conserved nevertheless, because the electric and magnetic fields carry momentum too. As the momentum of the charges changes, the

[6] For slowly moving charges this is Coulomb's law. It holds at any speed for charges that move at constant velocity, whose electric fields point straight away from where they are at that instant (Section 26.7).

momentum stored in their fields changes by the same amount in the opposite direction. The rest of this section finds how much momentum the fields hold and how it moves from place to place.

20.3.2 The force on charges, in terms of the fields

The momentum that charges gain comes from the force that the fields exert on them, just as the energy that they gain comes from the work that the fields do on them (Section 20.2.1). Poynting's theorem came from writing that work in terms of the fields alone; the same can be done for the force.

A charge q with velocity $\mathbf{v}$ feels the force $q(\mathbf{E}+\mathbf{v}\times\mathbf{B})$, (10.2). For charge spread out with density ρ, the charge in a volume element is $\rho\,d\tau$, and $\rho\mathbf{v}=\mathbf{J}$, so the force per unit volume is

$$\mathbf{f}=\rho\mathbf{E}+\mathbf{J}\times\mathbf{B}, \tag{20.17}$$

and the total force on the charges in a volume $\mathcal{V}$ is the integral of $\mathbf{f}$ over $\mathcal{V}$. The goal is a continuity equation for momentum with the form of those of Section 20.1,

$$\frac{\partial}{\partial t}\big(\text{momentum density of the fields}\big)+\nabla\cdot\big(\text{momentum flux density}\big)=-\mathbf{f},$$

whose right side is the negative of the rate, per unit volume, at which the fields give momentum to the charges. The units say what kind of quantity the flux density must be. A momentum density has units of $(\mathrm{kg\,m/s})/\mathrm{m}^3$, so its flux density has units of $(\mathrm{kg\,m/s})/(\mathrm{m}^2\,\mathrm{s})=\mathrm{N/m}^2$: a force per unit area, the units of a pressure. And since momentum is a vector, its flux density needs two directions: the direction of the momentum, and the orientation of the surface that it crosses. Three components of momentum, crossing surfaces of three orientations, make nine numbers: a rank 2 tensor (Section 8.2).

! **Units of flux densities.** Charge, $\mathbf{J}$: $\mathrm{C}/(\mathrm{m}^2\,\mathrm{s})$. Energy, $\mathbf{S}$: $\mathrm{J}/(\mathrm{m}^2\,\mathrm{s})=\mathrm{W/m}^2$. Momentum: $(\mathrm{kg\,m/s})/(\mathrm{m}^2\,\mathrm{s})=\mathrm{N/m}^2$, the units of a pressure.

The force density in terms of the fields

Gauss's law, (2.1), and the Ampère–Maxwell law, (2.4), give $\rho=\varepsilon_0\nabla\cdot\mathbf{E}$ and $\mathbf{J}=\nabla\times\mathbf{B}/\mu_0-\varepsilon_0\,\partial\mathbf{E}/\partial t$. Put them into (20.17):

$$\mathbf{f}=\varepsilon_0(\nabla\cdot\mathbf{E})\mathbf{E}+\frac{1}{\mu_0}(\nabla\times\mathbf{B})\times\mathbf{B}-\varepsilon_0\frac{\partial\mathbf{E}}{\partial t}\times\mathbf{B}.$$

To make the magnetic terms look like the electric one, add $(1/\mu_0)(\nabla\cdot\mathbf{B})\mathbf{B}$, which is zero by Gauss's law for magnetism, (2.2), and reverse the order of the cross product in the second term:

$$\mathbf{f}=\varepsilon_0(\nabla\cdot\mathbf{E})\mathbf{E}+\frac{1}{\mu_0}(\nabla\cdot\mathbf{B})\mathbf{B}-\frac{1}{\mu_0}\mathbf{B}\times(\nabla\times\mathbf{B})-\varepsilon_0\frac{\partial\mathbf{E}}{\partial t}\times\mathbf{B}.$$

The last term is part of the time derivative of the Poynting vector, $\mathbf{S}=\mathbf{E}\times\mathbf{B}/\mu_0$. The product rule and Faraday's law, (2.3), give

$$\mu_0\frac{\partial\mathbf{S}}{\partial t}=\frac{\partial\mathbf{E}}{\partial t}\times\mathbf{B}+\mathbf{E}\times\frac{\partial\mathbf{B}}{\partial t}=\frac{\partial\mathbf{E}}{\partial t}\times\mathbf{B}-\mathbf{E}\times(\nabla\times\mathbf{E}),$$

so the last term of $\mathbf{f}$ is $-\mu_0\varepsilon_0\,\partial\mathbf{S}/\partial t - \varepsilon_0\mathbf{E}\times(\nabla\times\mathbf{E})$. Now each field appears in the combination $\mathbf{A}\times(\nabla\times\mathbf{A})$. The product rule for the gradient of a dot product (Section 1.7), with both vectors equal to $\mathbf{A}$, gives $\nabla(A^2) = 2\mathbf{A}\times(\nabla\times\mathbf{A}) + 2(\mathbf{A}\cdot\nabla)\mathbf{A}$, so that $\mathbf{A}\times(\nabla\times\mathbf{A}) = \frac{1}{2}\nabla(A^2) - (\mathbf{A}\cdot\nabla)\mathbf{A}$. Using this for $\mathbf{E}$ and for $\mathbf{B}$ gives

$$\begin{aligned}\mathbf{f} = \varepsilon_0\left[(\nabla\cdot\mathbf{E})\mathbf{E} + (\mathbf{E}\cdot\nabla)\mathbf{E} - \tfrac{1}{2}\nabla(E^2)\right] \\ + \frac{1}{\mu_0}\left[(\nabla\cdot\mathbf{B})\mathbf{B} + (\mathbf{B}\cdot\nabla)\mathbf{B} - \tfrac{1}{2}\nabla(B^2)\right] - \mu_0\varepsilon_0\frac{\partial\mathbf{S}}{\partial t}.\end{aligned} \tag{20.18}$$

20.3.3 The Maxwell stress tensor

The brackets in (20.18) are sums of derivatives of products of field components, and they can be collected into a single tensor. Define the nine quantities

$$T_{ij} = \varepsilon_0\left(E_iE_j - \tfrac{1}{2}\delta_{ij}E^2\right) + \frac{1}{\mu_0}\left(B_iB_j - \tfrac{1}{2}\delta_{ij}B^2\right), \tag{20.19}$$

where i and j each stand for x, y, or z, and δ_{ij} is the Kronecker delta (Section 8.2.1). Together they make the **Maxwell stress tensor**, a rank 2 tensor that can be written as a 3×3 matrix, $\overleftrightarrow{\mathbf{T}}$. It is symmetric, $T_{ji} = T_{ij}$. Two of its elements, written out, are

$$\begin{aligned}T_{xx} &= \frac{\varepsilon_0}{2}\left(E_x^2 - E_y^2 - E_z^2\right) + \frac{1}{2\mu_0}\left(B_x^2 - B_y^2 - B_z^2\right), \\ T_{xy} &= \varepsilon_0E_xE_y + \frac{1}{\mu_0}B_xB_y.\end{aligned}$$

With them, the x component of (20.18) is

$$f_x = \frac{\partial T_{xx}}{\partial x} + \frac{\partial T_{xy}}{\partial y} + \frac{\partial T_{xz}}{\partial z} - \mu_0\varepsilon_0\frac{\partial S_x}{\partial t}, \tag{20.20}$$

as you can check by writing out both sides (Exercise P20.6), and the y and z components have the same form, with the second and third rows of $\overleftrightarrow{\mathbf{T}}$ in place of the first. The three derivatives in (20.20) are the divergence of the vector (T_{xx}, T_{xy}, T_{xz}), the first row of the matrix. Write $\nabla\cdot\overleftrightarrow{\mathbf{T}}$ for the vector whose components are the divergences of the three rows. Then

$$\boxed{\mathbf{f} = \nabla\cdot\overleftrightarrow{\mathbf{T}} - \mu_0\varepsilon_0\frac{\partial\mathbf{S}}{\partial t}}. \tag{20.21}$$

☞ *Caution*: Some references write the Maxwell stress tensor as σ_{ij}, with the same signs as here. Mind the sign of the momentum flux density (Section 20.3.4): it is $-T_{ij}$, not T_{ij}.

I call this result the **MST theorem** (for Maxwell stress tensor), by analogy with Poynting's theorem, although I have not seen the name used elsewhere. Like Poynting's theorem, it holds for any fields, static or changing.

To get the force on the charges in a volume $\mathcal{V}$ bounded by the closed surface $\mathcal{S}$, integrate (20.21) over the volume. The divergence theorem, (1.36), applied to each row of $\overleftrightarrow{\mathbf{T}}$, turns the integral of $\nabla\cdot\overleftrightarrow{\mathbf{T}}$ into an integral over the surface:

$$\mathbf{F} = \oint_{\mathcal{S}}\overleftrightarrow{\mathbf{T}}\cdot d\mathbf{a} - \mu_0\varepsilon_0\frac{d}{dt}\int_{\mathcal{V}}\mathbf{S}\,d\tau. \qquad \text{(MST theorem, integral form)} \tag{20.22}$$

Here $\overleftrightarrow{\mathbf{T}} \cdot d\mathbf{a}$ is the vector whose components are $\sum_j T_{ij}\, da_j$: the matrix $\overleftrightarrow{\mathbf{T}}$ times the column of components of $d\mathbf{a}$. With $d\mathbf{a} = \hat{\mathbf{n}}\, da$, where $\hat{\mathbf{n}}$ is the outward unit normal, the integral form written out is

$$\begin{pmatrix} F_x \\ F_y \\ F_z \end{pmatrix} = \oint_{\mathscr{S}} \begin{pmatrix} T_{xx} & T_{xy} & T_{xz} \\ T_{yx} & T_{yy} & T_{yz} \\ T_{zx} & T_{zy} & T_{zz} \end{pmatrix} \begin{pmatrix} n_x \\ n_y \\ n_z \end{pmatrix} da - \mu_0 \varepsilon_0 \frac{d}{dt} \int_{\mathscr{V}} \mathbf{S}\, d\tau,$$

which is the most useful form for calculations.

For static fields the last term is zero, and (20.22) says something remarkable: the force on all of the charges inside a closed surface depends only on the fields *on* the surface. To use it, draw a closed surface around the charges whose force you want, as you would draw a Gaussian surface around the charges whose field you want. We will call it a **Maxwellian surface**. Then work out $\overleftrightarrow{\mathbf{T}} \cdot \hat{\mathbf{n}}$ on each part of the surface, with $\hat{\mathbf{n}}$ pointing outward there, and integrate. Section 20.4 does this for several arrangements of charges.

The integral form also says what $\overleftrightarrow{\mathbf{T}}$ means. The force on the charges inside is a sum of contributions $\overleftrightarrow{\mathbf{T}} \cdot \hat{\mathbf{n}}\, da$ from the patches of the surface, as though the fields on the outside of each patch pulled or pushed on whatever is inside, the way the material on one side of an imaginary surface inside a solid pulls or pushes on the material on the other side. Forces per unit area of that kind are called *stresses*, which is the origin of the tensor's name, and its units are those of a stress, $\mathrm{N/m^2}$. On a patch whose outward normal points along x, the force per unit area is (T_{xx}, T_{yx}, T_{zx}). Its component along the normal, T_{xx}, is a stress that pulls outward if it is positive, or pushes inward if it is negative; its components along the patch, T_{yx} and T_{zx}, are shear stresses. So each diagonal element of $\overleftrightarrow{\mathbf{T}}$ gives a stress that pulls outward or pushes inward, according to its sign, and the off-diagonal elements give shear stresses.

Take, for example, a uniform electric field E along x, with no magnetic field. Then $T_{xx} = \frac{1}{2}\varepsilon_0 E^2$, $T_{yy} = T_{zz} = -\frac{1}{2}\varepsilon_0 E^2$, and the off-diagonal elements are zero. On the surface of an imaginary box in this field, the field pulls outward on the two faces that the field lines cross, and pushes inward on the other four, in each case with the force per unit area $\frac{1}{2}\varepsilon_0 E^2$ (Figure 20.7). Each field line pulls along its length, like a stretched elastic band, and pushes sideways on its neighbors, in each case with the force per unit area $\frac{1}{2}\varepsilon_0 E^2$. Magnetic field lines do the same, with $B^2/2\mu_0$ in place of $\frac{1}{2}\varepsilon_0 E^2$. This was Michael Faraday's picture of lines of force, which tend to shorten themselves and push each other apart sideways. Maxwell gave it mathematical form, first in his paper "On Physical Lines of Force" (1861) and then in his *Treatise on Electricity and Magnetism* (1873), where he described the stress in the field as a pull along the lines of force and an equal push in all directions at right angles to them.

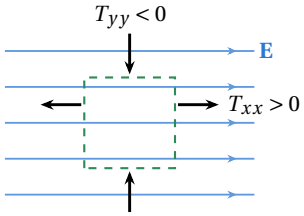


Figure 20.7 The stresses of a uniform electric field (blue) on the faces of an imaginary box (black arrows): the stress pulls outward on the two faces that the field lines cross, where $T_{xx} > 0$, and pushes inward on the other four, where T_{yy} and T_{zz} are negative. The forces add to zero, since no charge is inside.

20.3.4 Momentum in the fields

In a region with no charges, $\mathbf{f} = 0$, and the MST theorem, (20.21), can be written

$$\frac{\partial}{\partial t}\left(\mu_0\varepsilon_0\mathbf{S}\right) + \nabla\cdot\left(-\overleftrightarrow{\mathbf{T}}\right) = 0.$$

This has the form of the continuity equations of Section 20.1, one for each component of a vector: $\mu_0\varepsilon_0\mathbf{S}$ is the density, and $-\overleftrightarrow{\mathbf{T}}$ is its flux density. The units of $\mu_0\varepsilon_0\mathbf{S}$ are those of momentum per unit volume, and it is the momentum stored in the fields per unit volume, the **momentum density** of the fields:

$$\mathbf{g} = \mu_0\varepsilon_0\mathbf{S} = \frac{1}{c^2}\,\mathbf{S}. \tag{20.23}$$

The negative of the stress tensor is the momentum flux density: $-T_{ij}$ is the amount of the i component of momentum that crosses a unit area perpendicular to the j axis per unit time, counted as positive when it crosses toward $+j$, the way that $\mathbf{J}$ and $\mathbf{S}$ count charge and energy. The minus sign follows from Newton's second law. With $\hat{\mathbf{n}}$ pointing outward, $\overleftrightarrow{\mathbf{T}}\cdot\hat{\mathbf{n}}$ is the force per unit area that the fields outside a surface exert on whatever is inside it, and a force is the rate at which momentum flows *in*. So a positive T_{ij} means that the i component of momentum flows toward $-j$, in through a face whose outward normal points along $+j$; a negative T_{ij} means that it flows toward $+j$, out through that face. A stretched rope shows why. It pulls the hands at its two ends toward each other: the hand at the left end gains momentum that points to the right, and the hand at the right end loses it, so momentum that points to the right flows through the rope toward the left, opposite to the direction in which it points. A rod that pushes the hands apart carries that momentum toward the right.

Where there are charges, $\mathbf{f}$ is the rate, per unit volume, at which the fields change the charges' momentum, so the integral of $\mathbf{f}$ over a volume is the rate of change of the mechanical momentum $\mathbf{p}_{\text{mech}}$ of the charges inside. With $\mathbf{p}_{\text{fields}} = \int_{\mathcal{V}}\mathbf{g}\,d\tau$, the momentum stored in the fields inside, the integral form of the MST theorem, (20.22), becomes

$$\frac{d}{dt}\left(\mathbf{p}_{\text{mech}} + \mathbf{p}_{\text{fields}}\right) = \oint_{\mathcal{S}}\overleftrightarrow{\mathbf{T}}\cdot d\mathbf{a}. \tag{20.24}$$

This is a statement of the conservation of momentum. The total momentum inside the surface, that of the charges plus that of the fields, changes only as momentum flows out through the surface, at the rate $\oint(-\overleftrightarrow{\mathbf{T}})\cdot d\mathbf{a}$, or in. If the surface is so far away that the fields on it are negligible, the total does not change. For the two charges of Section 20.3.1, the momentum that their forces on each other give them is matched by the momentum that their fields lose. For a plane wave in vacuum, $\mathbf{S}$ is c times the energy density along the direction of travel, so $\mathbf{g}$ is the energy density divided by c; Section 21.5.1 uses this to find the pressure that light exerts.

Static fields can hold momentum too, wherever $\mathbf{E}$ and $\mathbf{B}$ are both nonzero and not parallel. When such fields change, their momentum passes to the charges, as the next example shows.

Example 20.2

A parallel-plate capacitor, with plates of area A a distance d apart, carries the charges Q and $-Q$, and sits in a uniform magnetic field $\mathbf{B}$ parallel to its plates, produced by magnets far away. Ignore fringing. (a) Find the momentum stored in the fields. (b) The capacitor is then discharged through a wire that runs straight across the gap from one plate to the other. Find the impulse, the integral of the force over time, that the magnetic field gives the wire while the current flows, and compare it with your answer to (a).

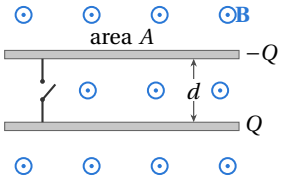
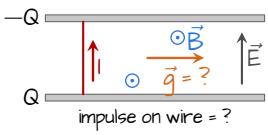


Figure 20.8 The capacitor of Example 20.2, edge on, in the magnetic field (blue), with the wire across the gap that will discharge it.

Solution:



The first step is a sketch of what is given and what is asked: the plates, with $\mathbf{E}$ from Q to $-Q$ and $\mathbf{B}$ out of the page; the momentum density $\mathbf{g}$, along $\mathbf{E}\times\mathbf{B}$; and the wire, whose current I flows along $\mathbf{E}$ while the capacitor discharges.

(a) Take z along $\mathbf{E}$, from the positive plate to the negative one, and x along $\mathbf{B}$. Between the plates $\mathbf{E} = (\sigma/\varepsilon_0)\,\hat{\mathbf{z}} = (Q/\varepsilon_0 A)\,\hat{\mathbf{z}}$, and outside them $\mathbf{E} = 0$, so the fields store momentum only between the plates, where

$$\begin{aligned}\mathbf{g} &= \mu_0\varepsilon_0\mathbf{S}\\ &= \varepsilon_0\mathbf{E}\times\mathbf{B}\\ &= \varepsilon_0\frac{Q}{\varepsilon_0 A}B\,(\hat{\mathbf{z}}\times\hat{\mathbf{x}})\\ &= \frac{QB}{A}\hat{\mathbf{y}}.\end{aligned}$$

This is uniform, so the momentum in the fields is $\mathbf{g}$ times the volume Ad between the plates:

$$\mathbf{p}_{\text{fields}} = QBd\,\hat{\mathbf{y}}.$$

(b) While the capacitor discharges, a current $I(t)$ flows through the wire from the positive plate to the negative one, along $+\hat{\mathbf{z}}$. The magnetic force on the wire, of length d, is $I\,d\,\hat{\mathbf{z}}\times B\,\hat{\mathbf{x}} = IBd\,\hat{\mathbf{y}}$. Over the whole discharge $\int I\,dt = Q$, the charge that flows, so the impulse on the wire is

$$\int \mathbf{F}\,dt = Bd\,\hat{\mathbf{y}}\int I\,dt = QBd\,\hat{\mathbf{y}}.$$

This is exactly the momentum that the fields held before the discharge, and it points the same way. When the electric field disappears, so does the momentum of the fields, and the wire, with the capacitor attached to it, takes it up. The impulse does not depend on how quickly the capacitor discharges.

Reality check: QBd has units of $\mathrm{C\cdot T\cdot m = C\cdot(kg/Cs)\cdot m = kg\,m/s}$, those of momentum.

20.4 Forces from the stress tensor

For static fields, the integral form of the MST theorem, (20.22), gives the force on the charges inside a Maxwellian surface from the fields on the surface alone:

$$\mathbf{F} = \oint_{\mathscr{S}} \overleftrightarrow{\mathbf{T}} \cdot d\mathbf{a}. \qquad \text{(static fields)} \quad (20.25)$$

This section uses it in two ways: to find the force per unit area on a surface, and to find the force between groups of charges.

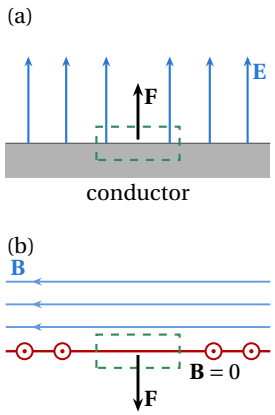


Figure 20.9 Maxwellian surfaces (dashed green) around small patches of (a) the surface of a charged conductor, where the field lines (blue) end on the surface and pull it outward, and (b) a sheet of surface current, where the magnetic field lines (blue) run along the sheet on one side and push it toward the other side (black arrows).

20.4.1 Pressure on a surface

Section 4.5.2 gave the force per unit area on the surface of a conductor, $\frac{1}{2}\varepsilon_0 E^2$ pointing outward, (4.7), as the charge on the surface times the average of the fields on its two sides. The stress tensor gives the same result without asking which field acts on the surface charge. Draw a Maxwellian surface around a small patch of the conductor's surface, of area A: a thin box with one face just outside the conductor and one just inside it (Figure 20.9a). Take z along the outward normal. Outside, $\mathbf{E} = E\hat{\mathbf{z}}$, so

$$T_{zz} = \varepsilon_0\left(E^2 - \tfrac{1}{2}E^2\right) = \tfrac{1}{2}\varepsilon_0 E^2, \qquad T_{xz} = T_{yz} = 0;$$

inside, $\mathbf{E} = 0$, and every element of $\overleftrightarrow{\mathbf{T}}$ is zero; and the sides of the box are too small to matter. Only the outer face contributes, with $\hat{\mathbf{n}} = \hat{\mathbf{z}}$, so the force on the patch is $T_{zz}A\hat{\mathbf{z}} = \frac{1}{2}\varepsilon_0 E^2 A\hat{\mathbf{z}}$, and the force per unit area is that of (4.7). With words, the field lines that end on the surface pull on it, like stretched elastic bands (Section 20.3.3).

A sheet of surface current, with a magnetic field on one side of it and none on the other, feels a force in the same way (Figure 20.9b). There the field $\mathbf{B}$ is parallel to the sheet, so on the face of the box on the field side, with n the direction of its outward normal, $T_{nn} = (1/\mu_0)(0 - \frac{1}{2}B^2) = -B^2/2\mu_0$: a pressure. The field lines push sideways on the sheet, and the sheet is pushed toward the side with no field, with the force per unit area $B^2/2\mu_0$. The windings of a solenoid are a sheet of this kind, and they are pushed outward (Exercise P20.13).

20.4.2 Forces between charges

The force that one group of charges exerts on another can always be found from Coulomb's law, by adding up the forces between pairs of charges. (20.25) offers another route: surround one group with a Maxwellian surface, and integrate the stress tensor of the total field over it. The route is worth taking when the fields on some convenient surface are simple, even if the charges are not. The best surfaces are usually made of planes of symmetry, where the field is simple, closed by pieces so far away that the field there is negligible. The first example has a known answer, which makes it a good test of the method.

Example 20.3

Two point charges, each equal to q, sit on the z axis at $z = d$ and $z = -d$. Use the Maxwell stress tensor to find the force on the upper charge.

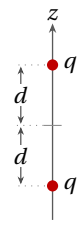
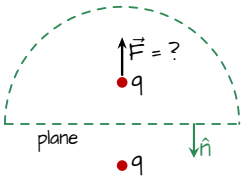


Figure 20.10 The two charges of Example 20.3.

Solution:



The first step is a sketch of what is given and what is asked: the two charges, and the force on the upper one, which points along z by symmetry. The Maxwellian surface (dashed) is the xy plane, closed by a hemisphere far above, on the plane the outward normal points down.

The fields are static, so (20.25) applies. As the hemisphere grows, its contribution goes to zero, and only the plane is left. (Why does the hemisphere's contribution go to zero as the hemisphere grows? Answer in footnote.[7]) On the plane $\hat{\mathbf{n}} = -\hat{\mathbf{z}}$, so the z component of the force is

$$F_z = \int_{\text{plane}} \left(T_{zx}\cdot 0 + T_{zy}\cdot 0 + T_{zz}\cdot(-1)\right) da = -\int_{\text{plane}} T_{zz}\, da.$$

At a point of the plane a distance s from the axis, each charge is a distance $\sqrt{s^2+d^2}$ away. The z components of their fields cancel, and the components along $\hat{\mathbf{s}}$ add:

$$\mathbf{E} = 2\,\frac{1}{4\pi\varepsilon_0}\,\frac{q}{s^2+d^2}\,\frac{s}{\sqrt{s^2+d^2}}\,\hat{\mathbf{s}} = \frac{1}{4\pi\varepsilon_0}\,\frac{2qs}{(s^2+d^2)^{3/2}}\,\hat{\mathbf{s}}.$$

Since $E_z = 0$ and $\mathbf{B} = 0$, $T_{zz} = \varepsilon_0\left(E_z^2 - \frac{1}{2}E^2\right) = -\frac{1}{2}\varepsilon_0 E^2$. With $da = 2\pi s\, ds$,

$$F_z = \frac{\varepsilon_0}{2}\int_0^\infty \left(\frac{1}{4\pi\varepsilon_0}\right)^2 \frac{4q^2s^2}{(s^2+d^2)^3}\,2\pi s\, ds = \frac{q^2}{4\pi\varepsilon_0}\int_0^\infty \frac{s^3\, ds}{(s^2+d^2)^3}.$$

The integral is $1/4d^2$ (substitute $u = s^2 + d^2$), so

$$\mathbf{F} = \frac{1}{4\pi\varepsilon_0}\,\frac{q^2}{4d^2}\,\hat{\mathbf{z}} = \frac{1}{4\pi\varepsilon_0}\,\frac{q^2}{(2d)^2}\,\hat{\mathbf{z}}.$$

Reality check: This is Coulomb's law for two charges a distance $2d$ apart, and the force points up, away from the lower charge, as it should for two charges of the same sign.

With words, the field lines from the two charges meet at the plane and push against each other there, with the pressure $\frac{1}{2}\varepsilon_0 E^2$, and this pressure pushes the charges apart.

The stress tensor shows its value when the charges are spread out, as in the next example. Coulomb's law would need a sixfold integral, over the points of both halves.

[7] Far away the field falls off as $1/r^2$, so the stress tensor falls off as $1/r^4$, while the area of the hemisphere grows only as r^2.

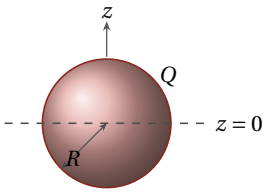
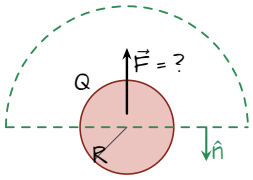


Figure 20.11 The charged ball of Example 20.4, and the plane $z = 0$ that divides it into two halves.

Example 20.4

An insulating ball of radius R carries the charge Q, spread uniformly through its volume. Imagine it cut into a northern and a southern half by the plane $z = 0$. The two halves repel each other. How large a force is needed to hold them together?

Solution:



The first step is a sketch of what is given and what is asked: the ball, and the force on its northern half, which points along z by symmetry. The Maxwellian surface (dashed) encloses the northern half: the xy plane, closed by a hemisphere far above.

The fields are static, so (20.25) applies, and as in Example 20.3 the hemisphere far away contributes nothing. The plane runs through the middle of the ball, but that does not matter. (Why may the Maxwellian surface pass through the charged ball? Answer in footnote.[8]) On the plane $\mathbf{E}$ points along $\hat{\mathbf{s}}$, by symmetry, so again $T_{zz} = -\frac{1}{2}\varepsilon_0 E^2$, and

$$F_z = -\int_{\text{plane}} T_{zz}\, da = \frac{\varepsilon_0}{2}\int_0^\infty E^2\, 2\pi s\, ds.$$

The field of the ball (Example 4.5) points away from its center, with the size $Qr/4\pi\varepsilon_0 R^3$ inside and $Q/4\pi\varepsilon_0 r^2$ outside, and on the plane $r = s$. So

$$\begin{aligned} F_z &= \frac{\varepsilon_0}{2}\left(\frac{Q}{4\pi\varepsilon_0}\right)^2 2\pi\left[\int_0^R \frac{s^3}{R^6}\, ds + \int_R^\infty \frac{ds}{s^3}\right] \\ &= \frac{Q^2}{16\pi\varepsilon_0}\left[\frac{1}{4R^2} + \frac{1}{2R^2}\right] \\ &= \frac{3Q^2}{64\pi\varepsilon_0 R^2}. \end{aligned}$$

The northern half is pushed up with this force, and the southern half down with a force of the same size, so holding the halves together takes a force $3Q^2/64\pi\varepsilon_0 R^2$ on each, toward the other.

Reality check: The result has the form $Q^2/\varepsilon_0 R^2$ times a number, the units of a force. For comparison, two point charges $Q/2$ at the centers of charge of the halves, a distance $3R/4$ apart, would repel each other with $Q^2/9\pi\varepsilon_0 R^2$, more than twice as much: the halves are too close together to be treated as points.

Tutorial 20.1 uses the stress tensor to find the force on a charge and on a wire in an applied field, and the force between two parallel wires.

[8] The plane holds no charge, since it has no thickness and the charge density is finite. So the closed surface encloses all of the northern half's charge and none of the southern half's, and the integral gives the force on the northern half.

Tutorial 20.1 Forces from the Maxwell stress tensor Printable version

Background

For static fields, the force on the charges and currents inside a Maxwellian surface is the integral of the stress tensor over the surface, (20.25), and the field that goes into the stress tensor is the total field on the surface. In this tutorial the total field is the sum of two fields: the field of the object whose force we want, and a field applied from outside. Each element of $\overleftrightarrow{\mathbf{T}}$ is quadratic in the fields, so splitting the field in two splits the stress tensor into three parts: an *applied part* $\overleftrightarrow{\mathbf{T}}^{\text{app}}$, made of the applied field alone; a *self part* $\overleftrightarrow{\mathbf{T}}^{\text{self}}$, made of the object's own field alone; and a *cross part* $\overleftrightarrow{\mathbf{T}}^{\text{cross}}$, each of whose terms contains one of each. You will find out which of the three parts push or pull on the object.

The force per unit area on a patch of the surface is the vector $\overleftrightarrow{\mathbf{T}} \cdot \hat{\mathbf{n}}$, whose components are $\sum_j T_{ij} n_j$, with $\hat{\mathbf{n}}$ the outward unit normal (Section 20.3.3).

Part 1 — A point charge in a uniform field

A point charge $q > 0$ sits at the origin in the uniform field $\mathbf{E}_0 = E_0 \hat{\mathbf{z}}$, with $E_0 > 0$, produced by charges far away. For the Maxwellian surface, take a sphere of radius a centered on the charge.

(a) Sketch the field lines of the total field near the charge. Where is the field stronger than E_0, and where is it weaker? Is there a point where it is zero? From the picture of field lines as stretched elastic bands (Section 20.3.3), predict the direction of the force on the charge.

(b) Write the total field as $\mathbf{E} = \mathbf{E}_0 + \mathbf{E}_q$, where $\mathbf{E}_q$ is the field of the charge. Show that the stress tensor (20.19) splits into the three parts of the background, with the cross part

$$T_{ij}^{\text{cross}} = \varepsilon_0 \left(E_{0i} E_{qj} + E_{qi} E_{0j} - \delta_{ij} \mathbf{E}_0 \cdot \mathbf{E}_q \right).$$

(c) The applied part is the stress tensor of a uniform field, the same at every point. Explain why its integral over any closed surface is zero.

(d) On the sphere, $\mathbf{E}_q = (q/4\pi\varepsilon_0 a^2)\,\hat{\mathbf{n}}$. Find $\overleftrightarrow{\mathbf{T}}^{\text{self}} \cdot \hat{\mathbf{n}}$ there, and integrate it over the sphere. What does your result say about whether a charge can push or pull on itself?

(e) Find $\overleftrightarrow{\mathbf{T}}^{\text{cross}} \cdot \hat{\mathbf{n}}$ on the sphere. You should find that it is the same vector at every point of the sphere, even though $\hat{\mathbf{n}}$ points a different way at each point. Integrate it over the sphere to find the force on the charge. Does your result depend on a? Should it?

Part 2 — A wire in a uniform field

A long straight wire of radius R lies along the z axis and carries the current $I > 0$ in the $+z$ direction, in the uniform field $\mathbf{B}_0 = B_0\hat{\mathbf{x}}$, with $B_0 > 0$, produced by magnets far away. For the Maxwellian surface, take a cylinder of radius $a > R$ and length ℓ, centered on the wire and closed by two flat end caps.

(a) Sketch the field lines of the total field in the xy plane. On which side of the wire is the field stronger? Magnetic field lines pull along their length and push sideways on their neighbors, with the pressure $B^2/2\mu_0$. Predict the direction of the force on the wire.

(b) The stress tensor splits into applied, self, and cross parts again, with $1/\mu_0$ in place of ε_0 and the magnetic fields in place of the electric ones. Explain why the applied part gives no net force, and why the two end caps together give no net force from any of the three parts.

(c) On the curved surface, the wire's own field is $\mathbf{B}_I = (\mu_0 I/2\pi a)\,\hat{\boldsymbol{\phi}}$, and the outward normal is $\hat{\mathbf{n}} = \hat{\mathbf{s}}$, where

$$\hat{\boldsymbol{\phi}} = -\sin\phi\,\hat{\mathbf{x}} + \cos\phi\,\hat{\mathbf{y}}, \qquad \hat{\mathbf{s}} = \cos\phi\,\hat{\mathbf{x}} + \sin\phi\,\hat{\mathbf{y}}.$$

Show that the self part gives no net force.

(d) Find $\overleftrightarrow{\mathbf{T}}^{\text{cross}} \cdot \hat{\mathbf{n}}$ on the curved surface, and integrate it to find the force on the length ℓ of the wire. Does it depend on a? Write your result as a vector equation that does not depend on which way the x and y axes point.

Part 3 — Two parallel wires

Two long parallel wires, much thinner than the distance d between them, carry the currents I_1 and I_2 in the same direction. Wire 1 lies along the z axis, and wire 2 along the line $x = d$, $y = 0$. Surround a length ℓ of wire 1 with a cylinder of radius a, larger than wire 1 but smaller than d.

(a) Sketch the field lines of the total field in the xy plane. Is the field between the wires stronger or weaker than on their far sides? Predict whether the wires attract or repel.

(b) The total field is $\mathbf{B}_1 + \mathbf{B}_2$, and Part 2 showed that wire 1's self part gives no net force. The field $\mathbf{B}_2$ is not uniform across the cylinder, so the argument that you used for the applied part in Part 2 does not work here. Use the MST theorem, (20.21), to explain why the part made of $\mathbf{B}_2$ alone still gives no net force on the cylinder. (Apply the theorem to the field $\mathbf{B}_2$ alone, whose only current is in wire 2.)

(c) Explain why the force on wire 1 that you find from the cylinder cannot depend on a, as long as a stays between the radius of wire 1 and d.

(d) So make a small enough that $\mathbf{B}_2$ is nearly uniform across the cylinder. Use your result from 2(d) to find the force per unit length on wire 1, and its direction.

Part 4 — Wrap-up

(a) Compare your results with the forces you already know: (3.6) for a charge, (10.3) for a wire, and Example 10.2 for two parallel wires. Did your sketches

predict the right directions? In all three parts only the cross part gave a force. In your own words, why?

(b) The negative of the stress tensor is the momentum flux density (Section 20.3.4). In Part 2, does momentum flow into the cylinder or out of it, and at what rate? If the wire is free to move, where does that momentum go, and where do you think it comes from?

20.5 Angular momentum in the fields

Since the fields carry momentum, they carry angular momentum too. A particle with momentum $\mathbf{p}$ at the position $\mathbf{r}$ has the angular momentum $\mathbf{r}\times\mathbf{p}$ about the origin, so the fields have the angular momentum $\mathbf{r}\times\mathbf{g}$ per unit volume, and the angular momentum stored in the fields in a volume $\mathcal{V}$ is

$$\mathbf{L}_{\text{fields}} = \int_{\mathcal{V}} \mathbf{r}\times\mathbf{g}\,d\tau = \varepsilon_0 \int_{\mathcal{V}} \mathbf{r}\times(\mathbf{E}\times\mathbf{B})\,d\tau. \tag{20.26}$$

Static fields can store angular momentum, and when they change, it can pass to the charges.

Example 20.5

Two long, coaxial, insulating cylindrical shells of length ℓ carry opposite charges: the inner one, of radius a, carries Q, and the outer one, of radius b, carries $-Q$, each spread evenly over its surface. They sit in a uniform magnetic field $\mathbf{B} = B_0\hat{\mathbf{z}}$ along their axis, produced by a solenoid around both, and they are free to turn about the axis. (a) Find the angular momentum stored in the fields. (b) The current in the solenoid is then turned down steadily, so that the field falls linearly from B_0 to zero in a time t_0. Find the torque on each shell while the field falls and the angular momentum that each shell gains, and compare the total with your answer to (a). Ignore the ends of the shells, and the magnetic fields of the shells once they turn.

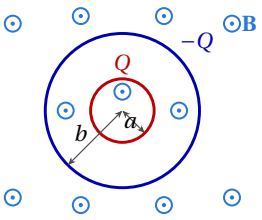
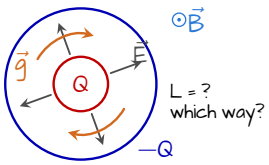


Figure 20.12 The charged shells of Example 20.5, seen along their axis, in the uniform magnetic field (blue).

Solution:



The first step is a sketch of what is given and what is asked: the two shells seen along the axis, with $\mathbf{E}$ outward between them and $\mathbf{B}$ out of the page. The momentum density $\varepsilon_0\mathbf{E}\times\mathbf{B}$ circles the axis. Which way each shell turns is to be found.

(a) By Gauss's law, the field between the shells is $\mathbf{E} = (Q/2\pi\varepsilon_0\ell s)\,\hat{\mathbf{s}}$, and elsewhere $\mathbf{E} = 0$: inside the inner shell there is no charge, and outside the outer shell the total charge is

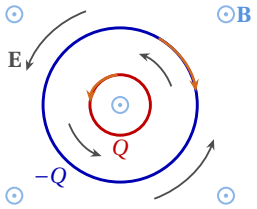


Figure 20.13 The shells of Example 20.5, seen from above while the magnetic field (blue) dies away. The induced electric field (black) circles the axis counterclockwise, and it turns the inner shell, with the charge Q, counterclockwise and the outer one, with $-Q$, clockwise (orange).

zero. Between the shells the momentum density is

$$\begin{aligned}\mathbf{g} &= \varepsilon_0 \mathbf{E}\times\mathbf{B}\\ &= \varepsilon_0 \frac{Q}{2\pi\varepsilon_0 \ell s} B_0\,(\hat{\mathbf{s}}\times\hat{\mathbf{z}})\\ &= -\frac{QB_0}{2\pi\ell s}\,\hat{\boldsymbol{\phi}},\end{aligned}$$

which circles the axis clockwise as seen from above. With $\mathbf{r} = s\hat{\mathbf{s}} + z\hat{\mathbf{z}}$, the z component of $\mathbf{r}\times\mathbf{g}$ is $s\,g_\phi = -QB_0/2\pi\ell$, the same everywhere between the shells. (The other part of $\mathbf{r}\times\mathbf{g}$, along $\hat{\mathbf{s}}$, adds up to zero.) Multiplying by the volume between the shells, $\pi(b^2-a^2)\ell$, gives

$$\mathbf{L}_{\text{fields}} = -\frac{QB_0}{2}\left(b^2-a^2\right)\hat{\mathbf{z}}.$$

(b) As B falls, Faraday's law induces an electric field that circles the axis. Around a circle of radius s, $E_\phi\, 2\pi s = -d\Phi_B/dt = -\pi s^2\, dB/dt$, so $E_\phi = -\frac{1}{2}s\,dB/dt$. On the ramp $dB/dt = -B_0/t_0$, so $E_\phi = sB_0/2t_0$, positive and constant while the field falls. All of the inner shell's charge is at $s = a$, so the torque on it about the axis is

$$\tau_{\text{inner}} = aQE_\phi(a) = \frac{Qa^2B_0}{2t_0},$$

the same throughout the ramp, and in the time t_0 the inner shell gains the angular momentum

$$L_{\text{inner}} = \tau_{\text{inner}}\, t_0 = \frac{Qa^2B_0}{2}.$$

The outer shell, with the charge $-Q$ at $s = b$, feels the torque $-Qb^2B_0/2t_0$ and gains $L_{\text{outer}} = -Qb^2B_0/2$. The inner shell turns counterclockwise as seen from above and the outer one clockwise (Figure 20.13). Together they gain

$$L_{\text{inner}} + L_{\text{outer}} = -\frac{QB_0}{2}\left(b^2-a^2\right),$$

the angular momentum that the fields held before. Once $B = 0$ the fields hold none, so the angular momentum has passed from the fields to the shells, and the total is conserved.

The linear ramp did not matter. For any way of switching the field off, the torque on the inner shell at each instant is $\tau_{\text{inner}} = aQE_\phi(a) = -\frac{1}{2}Qa^2\,dB/dt$, so

$$\begin{aligned}L_{\text{inner}} &= \int \tau_{\text{inner}}\,dt\\ &= -\frac{Qa^2}{2}\int_{B_0}^{0} dB\\ &= \frac{Qa^2B_0}{2},\end{aligned}$$

which depends only on the field at the start and at the end, not on how quickly or how steadily it falls in between. The same holds for the outer shell.

Reality check: QB_0b^2 has units of $\mathrm{C\cdot T\cdot m^2} = \mathrm{kg\,m^2/s}$, those of angular momentum.

Exercise P15.9 described a puzzle that Richard Feynman discussed in his lectures: a disk with charges fixed around its rim, threaded by a long solenoid. When the solenoid's current is switched off, the induced electric field turns the disk, which ends up with angular momentum, although nothing was turning before. The resolution is the same as for the shells: before the switch, the fields held that angular momentum. Where they held it is less obvious. Inside the solenoid **B** is strong, and the electric field of the charges reaches in, yet for a long solenoid the angular momentum stored there is small. Inside an ideal solenoid, with $\mathbf{B} = B_0\hat{\mathbf{z}}$, the z component of $\mathbf{r}\times\mathbf{g}$ is $-\varepsilon_0 B_0 s E_s$, so the angular momentum inside depends only on the flux of **E** through the curved sides of cylinders around the axis; by Gauss's law, since no charge is inside, each such flux is the negative of the flux through the cylinder's ends, which is small for a long cylinder. Nearly all of the angular momentum is stored *outside* the solenoid, where the electric field of the charges overlaps the weak magnetic field that returns from one end of the solenoid to the other, over a region larger than the solenoid itself (Figure 20.14). A numerical integration over all of space, for a solenoid ten times as long as its radius, with the charges at twice its radius, finds only 1.5 percent of the angular momentum inside the solenoid and half of it more than 7 radii from the center of the disk, and a total equal to the angular momentum that the disk ends up with.

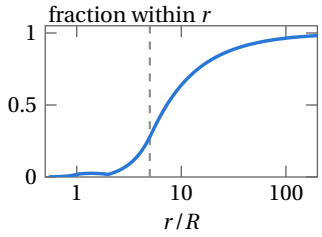


Figure 20.14 Where the fields of Feynman's disk store their angular momentum, for a solenoid of radius R and length $10R$, with the charges at radius $2R$ (computed numerically): the fraction of the angular momentum that lies within a distance r of the disk's center. The dashed line marks the distance from the center to the ends of the solenoid.

20.6 Summary

Charge, energy, momentum, and angular momentum are all conserved, and conserved locally: the amount of each in a region changes only as it flows through the region's boundary. Energy, momentum, and angular momentum can also pass between the charges and the fields. This chapter wrote the continuity equations for charge, energy, and momentum, found the densities and flux densities of energy and momentum in the fields, and showed that even static fields can store momentum and angular momentum.

Charge

$\partial\rho/\partial t + \nabla\cdot\mathbf{J} = 0$	the continuity equation
$dQ_{\text{enc}}/dt = -\oint \mathbf{J}\cdot d\mathbf{a}$	its integral form

Energy

$u = \frac{1}{2}\varepsilon_0 E^2 + B^2/2\mu_0, \quad \mathbf{S} = \mathbf{E}\times\mathbf{B}/\mu_0$	energy density; Poynting vector
$\partial u/\partial t + \nabla\cdot\mathbf{S} = -\mathbf{E}\cdot\mathbf{J}$	Poynting's theorem
$dW/dt = -dU/dt - \oint \mathbf{S}\cdot d\mathbf{a}$	its integral form
$\langle\mathbf{S}\rangle = \text{Re}[\tilde{\mathbf{E}}\times\tilde{\mathbf{B}}^*]/2\mu_0$	time average; its size is the intensity
$u = \frac{1}{2}(\mathbf{E}\cdot\mathbf{D} + \mathbf{B}\cdot\mathbf{H}), \quad \mathbf{S} = \mathbf{E}\times\mathbf{H}$	in linear materials

Momentum and angular momentum

$\mathbf{f} = \rho\mathbf{E} + \mathbf{J} \times \mathbf{B}$	force per unit volume on charges
$T_{ij} = \varepsilon_0 \left(E_i E_j - \frac{1}{2}\delta_{ij}E^2\right) + \frac{1}{\mu_0}\left(B_i B_j - \frac{1}{2}\delta_{ij}B^2\right)$	the Maxwell stress tensor: field lines pull along their length and push sideways
$\mathbf{f} = \nabla \cdot \overleftrightarrow{\mathbf{T}} - \mu_0\varepsilon_0 \partial\mathbf{S}/\partial t$	the MST theorem
$\mathbf{F} = \oint \overleftrightarrow{\mathbf{T}} \cdot d\mathbf{a} - \mu_0\varepsilon_0 (d/dt) \int \mathbf{S}\, d\tau$	force on the charges enclosed
$\mathbf{g} = \mu_0\varepsilon_0\mathbf{S} = \mathbf{S}/c^2$	momentum density; $-T_{ij}$ is its flux density
$d(\mathbf{p}_{\text{mech}} + \mathbf{p}_{\text{fields}})/dt = \oint \overleftrightarrow{\mathbf{T}} \cdot d\mathbf{a}$	conservation of momentum
$\frac{1}{2}\varepsilon_0 E^2$ outward; $B^2/2\mu_0$	forces per unit area on a conductor's surface and on a current sheet
$\mathbf{L}_{\text{fields}} = \varepsilon_0 \int \mathbf{r} \times (\mathbf{E} \times \mathbf{B})\, d\tau$	angular momentum stored in the fields, even static ones

Exercises

Exercises for 20.1 Conservation of charge: the continuity equation

Warm-up questions Read Section 20.1, then answer before class

W20.1 Suppose the closed surface of Figure 20.1 were enlarged to surround both plates of the capacitor, so that the wire to each plate crosses it. What does (20.2) then say about the charge inside? Is that what you expect for the two plates?

W20.2 In the pair creation described at the start of this section, an electron and a positron appear inside a small closed surface, and then the electron alone crosses the surface on its way out. While it crosses, is $\oint \mathbf{J} \cdot d\mathbf{a}$ positive or negative? Does (20.2) then give the change in the charge inside that you expect?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

Exercises for 20.2 Energy conservation: Poynting's theorem

Warm-up questions Read Section 20.2, then answer before class

W20.3 In Example 20.1 the energy that heats the wire flows in through its sides, carried by $\mathbf{S} = \mathbf{E} \times \mathbf{B}/\mu_0$, which contains the magnetic field. A classmate concludes that the magnetic field does part of the work of heating the wire. Do you agree? Which field does the work on the charges, and why is the other one needed?

W20.4 Suppose the current in the wire of Example 20.1 alternates slowly, so that at each instant the fields are those of the example for the current at that instant. Does energy ever flow out of the wire through its sides, for instance while the current flows the other way? Explain.

W20.5 At a fixed point in the wave of Section 20.2.4, sketch S_z against the time for one period of the fields. Does $\mathbf{S}$ ever point toward $-z$? Why does it oscillate at twice the frequency of the fields?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P20.1 A long coaxial cable has a solid inner conductor of radius a and a thin outer conductor of radius b. A battery at one end keeps the inner conductor at the potential V relative to the outer one, and a resistor at the other end closes the circuit, so that a current I flows along the inner conductor and back along the outer one (Figure 20.15). Take the conductors to have

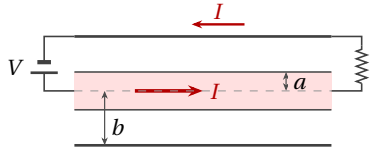


Figure 20.15 The cable of Exercise P20.1, cut through its axis, with the battery at one end and the resistor at the other.

no resistance.

(a) Find **E** and **B** between the conductors.

(b) Find the Poynting vector between the conductors.

(c) Integrate it over the space between the conductors in a plane perpendicular to the axis, and show that the power carried along the cable is VI.

(d) Which way does the energy flow, and where does it go? What does $\mathbf{E} = 0$ inside the conductors say about the energy that flows inside them?

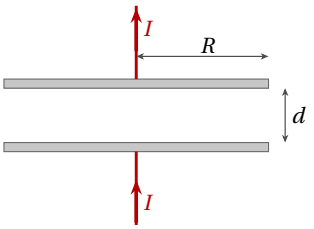


Figure 20.16 The capacitor of Exercise P20.2, edge on, while it is being charged.

P20.2 A capacitor with circular plates of radius R, a distance d apart, is charged by a constant current I, as in Example 16.1 (Figure 20.16). Ignore fringing.

(a) What are the directions of **E** and **B** between the plates, and of the Poynting vector at the edge of the plates? What does the direction of **S** tell you?

(b) When the charge on the plates is Q, find **S** at the edge, $s = R$, and the rate at which energy flows in through the cylindrical surface that joins the edges of the plates.

(c) Show that this rate equals the rate at which the energy stored in the field between the plates increases.

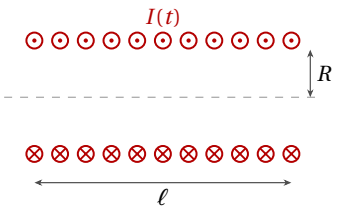


Figure 20.17 The solenoid of Exercise P20.3, cut through its axis.

P20.3 A long solenoid of radius R and length ℓ, with n turns per unit length, carries a current $I(t)$ that increases with time (Figure 20.17).

(a) Find the magnetic field inside, and the electric field inside, using the induced field of Example 15.4.

(b) Find the Poynting vector inside, and show that it points toward the axis.

(c) Find the rate at which energy flows in through a cylinder of radius R and length ℓ just inside the windings, and show that it equals $d(\frac{1}{2}LI^2)/dt$, where $L = \mu_0 n^2 \pi R^2 \ell$ is the inductance of the solenoid.

(d) Where does this energy come from? Consider the sign of $\mathbf{E} \cdot \mathbf{J}$ in the windings.

P20.4 A typical red laser pointer puts out 5 mW in a beam about 1 mm across.

(a) Find the intensity of the beam, taking the power to be spread evenly over a circle 1.0 mm in diameter.

(b) Treat the beam as a wave like the one of Section 20.2.4, and find the amplitudes of its electric and magnetic fields.

(c) Find the average energy density in the beam, and show that the intensity is c times it.

P20.5 A point charge q sits at rest at the origin in a uniform magnetic field $\mathbf{B} = B_0\hat{\mathbf{z}}$ (Figure 20.18), produced by coils far away. Nothing changes with time.

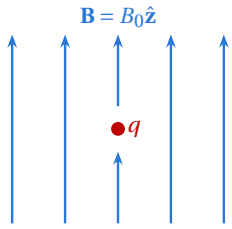


Figure 20.18 The charge and field of Exercise P20.5.

(a) Find the Poynting vector everywhere, in spherical coordinates, and show that it makes the pattern of Figure 20.3. Then make a plot like that figure: in Mathematica with `VectorPlot3D` (or `StreamPlot3D` for field lines), or in Python with the `quiver` function of `matplotlib` on three-dimensional axes.

(b) Show that $\nabla \cdot \mathbf{S} = 0$, and explain with Poynting's theorem why it must be.

(c) Does energy pile up anywhere? Does it go anywhere?

Exercises for 20.3 Momentum conservation: the Maxwell stress tensor

Warm-up questions Read Section 20.3, then answer before class

W20.6 Suppose the magnetic field in Example 20.2 pointed perpendicular to the plates, along **E**, instead of parallel to them. How much momentum would the fields store? What impulse would the wire receive while the capacitor discharges? Do the two answers agree?

W20.7 Section 20.3.4 writes the MST theorem, for a region with no charges, as a continuity equation for momentum, as Poynting's theorem is one for energy. In what ways are the two equations similar? Different? (Compare their densities, their flux densities, and how many equations each one stands for.)

W20.8 A uniform electric field **E** points at 45° between the $+x$ and $+y$ axes, with no magnetic field. Use (20.19) to decide whether T_{xx} is positive, negative, or zero. Which way does the force per unit area point on a face of an imaginary box whose outward normal points along $+x$? (Section 20.3.3 gives that force in terms of the elements of $\overleftrightarrow{\mathbf{T}}$.)

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P20.6 Show that (20.20) is the x component of (20.18). Write out the x components of the electric terms of (20.18); then work out $\partial T_{xx}/\partial x + \partial T_{xy}/\partial y + \partial T_{xz}/\partial z$ for the electric part of (20.19), and compare. (The magnetic terms work the same way.)

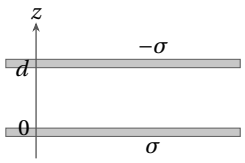


Figure 20.19 The capacitor of Exercise P20.7, edge on.

P20.7 A large parallel-plate capacitor has its plates in the planes $z = 0$ and $z = d$, with the surface charge density σ on the lower plate and $-\sigma$ on the upper one (Figure 20.19). Ignore fringing.

(a) Find all nine elements of the Maxwell stress tensor between the plates, and write them as a 3×3 matrix.

(b) Use the integral form of the MST theorem, with a Maxwellian surface around a patch of the upper plate, to find the force per unit area on that plate. Compare with (4.7).

(c) Find the momentum that crosses a unit area of the plane $z = d/2$ per unit time. Which component of momentum crosses, and which way does it flow?

(d) Nothing between the plates changes with time, so the momentum that crosses the gap must be taken up by the plates. Use this to find the force per unit area on the upper plate again, and compare with (b).

P20.8 A plane wave travels along $+z$ in vacuum, with $\mathbf{E} = E_0 \cos(kz - \omega t)\,\hat{\mathbf{x}}$ and $\mathbf{B} = (E_0/c)\cos(kz - \omega t)\,\hat{\mathbf{y}}$.

(a) Show that every element of the Maxwell stress tensor is zero except T_{zz}, and that $T_{zz} = -u$, where u is the energy density of the wave.

(b) Find the momentum flux density of the wave. Which component of momentum flows, and which way? Show that the flux is c times the momentum density, as it should be for momentum that travels at the speed c.

(c) The wave strikes a surface in the plane $z = 0$ that absorbs it completely. Use the time average of (b) to find the force per unit area on the surface, in terms of the wave's intensity.

P20.9 Estimate how much momentum the sunlight in a sunny office holds at any instant, in units of kgm/s. Sunlight above Earth's atmosphere has an intensity of $1361\,\mathrm{W/m^2}$ (Section 20.2.4), but the light inside an office is much dimmer than the light outside. State every estimate and approximation that you make, and add up the sizes of the momenta as though all of the light traveled in the same direction.

Exercises for 20.4 Forces from the stress tensor

Warm-up questions Read Section 20.4, then answer before class

W20.9 A classmate reads (20.25) and concludes that charges outside a Maxwellian surface cannot affect the force on the charges inside it, since only the fields on the surface enter the integral. What is wrong with this reasoning? How do the charges outside enter?

W20.10 Suppose the charges of Example 20.3 were unequal: q at $z = d$ and $2q$ at $z = -d$. The plane $z = 0$, closed by a hemisphere far away, still encloses only the upper charge. What would change in the field on the plane, and in T_{zz} there? Would the integral be as simple?

W20.11 Suppose the ball of Example 20.4 were a conductor carrying the same charge Q. Where on the plane $z = 0$ would the field change? Would the force between the two halves be larger or smaller?

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P20.10 Repeat Example 20.3 with the charge q at $z = d$ and the charge $-q$ at $z = -d$.

(a) Find $\mathbf{E}$ on the plane $z = 0$, and the elements T_{zx}, T_{zy}, and T_{zz} there.

(b) Find the force on the upper charge, and check it with Coulomb's law.

(c) In the example, the field lines met at the plane and pushed against each other. What do they do here?

P20.11 Find the force between the halves of the charged ball of Example 20.4 again, with a different Maxwellian surface: the surface of the northern half itself, made of a hemispherical bowl of radius R and the flat disk of radius R in the plane $z = 0$ that closes it.

(a) On the bowl, $\mathbf{E} = (Q/4\pi\varepsilon_0 R^2)\,\hat{\mathbf{r}}$. Find the z component of $\overleftrightarrow{\mathbf{T}} \cdot \hat{\mathbf{n}}$ there, in spherical coordinates, and integrate it over the bowl.

(b) Find the disk's contribution to F_z, using the field inside the ball.

(c) Show that the total agrees with the example. Which route was faster?

P20.12 Two electric dipoles, each with the dipole moment $\mathbf{p} = p\hat{\mathbf{z}}$, sit on the z axis at $z = d$ and $z = -d$, a distance $2d$ apart (Figure 20.20). Use the Maxwell stress tensor to find the force between them, as a function of p and d, with the xy plane, closed by a hemisphere far away, as your Maxwellian surface. You will need the field on the plane as a function of s: start from the field of a dipole, (8.15), and remember that the dipoles are not at the origin. Is the force attractive or repulsive?

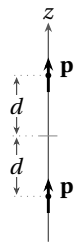


Figure 20.20 The two dipoles of Exercise P20.12.

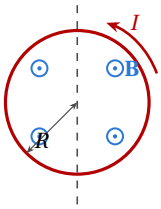


Figure 20.21 The solenoid of Exercise P20.13, seen along its axis, and the plane through the axis that divides it into two halves.

P20.13 A long solenoid of radius R has the magnetic field B inside and, ideally, none outside. Its windings are pushed outward with the force per unit area $B^2/2\mu_0$ (Section 20.4.1).

(a) Imagine the solenoid divided into two halves by a plane through its axis (Figure 20.21). Use that plane, closed far away, as a Maxwellian surface to find the force, per unit length of the solenoid, that pushes one half away from the other.

(b) Show that the same force follows from adding up the components of the outward pressure on the half's windings perpendicular to the plane.

(c) The magnet of a typical whole-body MRI scanner is a solenoid with $B = 3.0\,\mathrm{T}$ and a radius of about 0.5 m. Find the pressure on its windings, in atmospheres, and the force per meter of length from (a).

Exercises for 20.5 Angular momentum in the fields

Warm-up questions Read Section 20.5, then answer before class

W20.12 Suppose the shells of Example 20.5 were fastened to the floor instead of being free to turn. When the field is switched off, would the fields still lose their angular momentum? If so, where would it go?

W20.13 The angular momentum (20.26) is measured about the origin. Would the angular momentum that the fields of Example 20.5 store change if it were measured about a point off the axis? (What is the total momentum of these fields?)

Also: without looking back, summarize this section in a few sentences, and say what you found hardest to understand.

P20.14 Take the shells of Example 20.5, in the same field $B_0\hat{\mathbf{z}}$, but leave the field on. Instead, connect the shells with a thin wire that runs straight out from the inner shell to the outer one, fixed to both, so that the shells discharge through it.

(a) While the current I flows outward along the wire, find the magnetic force on the wire and its torque about the axis.

(b) Find the angular momentum that the shells and the wire gain while the shells discharge, and compare with the angular momentum stored in the fields, found in the example. Does it depend on how quickly the shells discharge?

(c) Which way do the shells and the wire turn, as seen from above? Compare with the example, in which the two shells turned in opposite directions.

Exercises for 20.6 Summary

P20.15 True or false? Explain your reasoning.

(a) Charge could disappear in one place and reappear at the same instant somewhere else, as long as the total charge stays the same.

(b) Wherever nothing changes with time, the Poynting vector is zero.

(c) The energy that heats a wire carrying a steady current flows into the wire through its sides.

(d) Two moving charges always exert equal and opposite forces on each other.

(e) The elements of the Maxwell stress tensor have units of force per unit area.

(f) The stress of a field pushes inward on a surface that its field lines cross, and pulls outward on a surface parallel to them.

(g) For static fields, the force on the charges inside a closed surface can be found from the fields on the surface alone.

(h) Static fields cannot store momentum.

(i) The momentum density of the fields is the Poynting vector divided by c.

(j) The surface of a charged conductor is pushed inward by its field.

(k) Fields can store angular momentum even when no charge is moving.

P20.16 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as the flow of energy around a circuit made of a battery, wires, and a resistor, the Poynting vector of two oscillating dipoles, the force between two charged objects found by integrating the stress tensor numerically over a Maxwellian surface, the stresses on the field lines of two charges drawn along with the lines, the momentum stored in the fields of a charge near a magnet, or where the fields of Feynman's disk store their angular momentum for solenoids of different lengths.

Part VI

Electromagnetic Waves

The agreement of the results seems to show that light and magnetism are affections of the same substance, and that light is an electromagnetic disturbance propagated through the field according to electromagnetic laws.

— James Clerk Maxwell,
"A Dynamical Theory of the Electromagnetic Field" (1865)*

*"A dynamical theory of the electromagnetic field," J. C. Maxwell, Phil. Trans. R. Soc. Lond. 155, 459–512 (1865), Part VI, paragraph 97, p. 498. https://doi.org/10.1098/rstl.1865.0008.

Chapter 21

Real Waves

Section 16.5 found that Maxwell's equations allow electric and magnetic fields to travel through empty space as waves, at the speed $c = 1/\sqrt{\mu_0\varepsilon_0}$, and that light is such a wave. This chapter takes up these waves in detail. It starts with the wave equation inside matter and with the vocabulary of sinusoidal waves, using waves on a string wherever the string makes an idea easier to see. Then come polarization, the relations between **E** and **B** that Maxwell's equations require of a plane wave, and the energy and momentum that waves carry. The chapter ends with what happens when a wave reaches the boundary between two materials: how much of it is reflected, how much is transmitted, and in which directions the two waves go. That is much of the physics of optics.

Throughout this chapter the materials are linear and isotropic, they do not conduct, and their permittivity and permeability are real constants. Chapter 22 relaxes these assumptions, for conductors and for materials whose response depends on the frequency, and Chapter 23 confines waves between conducting walls.

21.1 The wave equation in vacuum and in matter

In empty space every Cartesian component of **E** and of **B** obeys the wave equation (16.13),

$$\nabla^2 f = \frac{1}{v^2}\frac{\partial^2 f}{\partial t^2},$$

with $v = 1/\sqrt{\mu_0\varepsilon_0} = c$ (Section 16.5). In one dimension any function of $z - vt$ solves it: a disturbance of any shape that travels toward $+z$ at the speed v without changing shape. A function of $z + vt$ travels toward $-z$, and the general solution in one dimension is a sum of the two.

Inside matter, Maxwell's equations take the form (19.2). Take a material that is linear and isotropic, so that $\mathbf{D} = \varepsilon\mathbf{E}$ and $\mathbf{B} = \mu\mathbf{H}$ (19.3); that is uniform, so that ε and μ are the same everywhere; and that has no free charges or currents. Then

$\nabla\cdot\mathbf{D} = \varepsilon\nabla\cdot\mathbf{E}$ and $\nabla\times\mathbf{H} = (1/\mu)\nabla\times\mathbf{B}$, and the four equations become

$$\nabla\cdot\mathbf{E} = 0, \qquad \nabla\cdot\mathbf{B} = 0, \qquad \nabla\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t}, \qquad \nabla\times\mathbf{B} = \mu\varepsilon\frac{\partial\mathbf{E}}{\partial t}. \tag{21.1}$$

These are the equations of empty space, (16.11), with $\mu_0\varepsilon_0$ replaced by $\mu\varepsilon$. The steps of Section 16.5 therefore go through unchanged, and give

$$\nabla^2\mathbf{E} = \mu\varepsilon\frac{\partial^2\mathbf{E}}{\partial t^2}, \qquad \nabla^2\mathbf{B} = \mu\varepsilon\frac{\partial^2\mathbf{B}}{\partial t^2}. \tag{21.2}$$

(The uniformity matters. Where ε changes from place to place, $\nabla\cdot\mathbf{E}$ is not zero even when there is no free charge, and the equation for $\mathbf{E}$ gains another term; Exercise P21.1.) Waves in the material travel at the speed

$$v = \frac{1}{\sqrt{\mu\varepsilon}} = \frac{c}{n}, \qquad n \equiv \sqrt{\frac{\mu\varepsilon}{\mu_0\varepsilon_0}} = \sqrt{\mu_r\varepsilon_r}, \tag{21.3}$$

where n is the material's **index of refraction**. For most materials that are not ferromagnetic, μ_r differs from 1 by less than about 10^{-4} (Table 13.1), so to an excellent approximation

$$n \approx \sqrt{\varepsilon_r}.$$

This is remarkable! The dielectric constant, which you likely first learned about in the context of capacitors, and the index of refraction, which you likely first learned about in the context of Snell's law, have this simple relationship with each other! Sadly, it seems like most introductory physics books do not highlight this connection. (Exercise P16.15 found it for a dielectric that is not magnetic.) Do be careful, though, because frequency is critical, and the "dielectric constant" is not "constant" — it varies strongly with frequency, as we will discuss more in Section 22.5, and typically we use capacitors at frequencies below about 10^9 Hz, whereas the index of refraction normally applies to optical frequencies (about 10^{15} Hz). The ε and μ in these equations are the values that the material has at the frequency of the wave, which need not be the static values of the earlier chapters: a material does not respond to fields that oscillate quickly in the same way as to steady ones (Section 19.1.2).

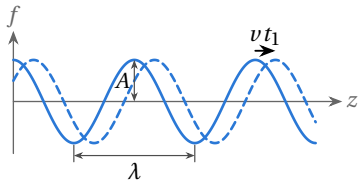


Figure 21.1 The sinusoidal wave (21.4), with $\phi = 0$, at $t = 0$ (solid) and at a later time t_1 (dashed): every crest has moved a distance vt_1 toward $+z$.

21.2 Sinusoidal waves

The simplest solutions of the wave equation, and the most useful, are sinusoidal. In one dimension, a sinusoidal wave that travels toward $+z$ is

$$f(z,t) = A\cos(kz - \omega t + \phi) = A\cos\big(k(z - vt) + \phi\big), \qquad v = \frac{\omega}{k}. \tag{21.4}$$

It is a function of $z - vt$, so it moves toward $+z$ at the speed v without changing shape (Figure 21.1). Its parts have names. A, the largest value of f, is the **amplitude**. k is the **wave number**: at a fixed time the wave repeats each time kz grows

by 2π, so the **wavelength**, the distance from one crest to the next, is $\lambda = 2\pi/k$. ω is the **angular frequency**: at a fixed place the wave repeats each time ωt grows by 2π, so the **period**, the time from one crest to the next, is $T = 2\pi/\omega$, and the **frequency**, the number of crests that pass each second, is $\nu = 1/T = \omega/2\pi$, in hertz. (Many books write f for the frequency; this book writes ν, and this chapter uses f for the wave itself.) ϕ is the **phase constant**, which shifts the whole wave toward $-z$ by the distance ϕ/k, the fraction $\phi/2\pi$ of a wavelength; at $z = 0$ the wave is $A\cos(\omega t - \phi)$, in the time-delay convention that this book uses for all waves (Section 18.2.1). And $v = \omega/k = \lambda\nu$, the speed at which the crests move, is the **phase velocity**.

Inside a material, $v = c/n$ by (21.3), so

$$\frac{\omega}{k} = \frac{c}{n}. \tag{21.5}$$

This is my favorite equation in the entire book! It's simple in the sense that we are just saying "velocity equals velocity," where the first velocity is the phase velocity of plane waves, ω/k, and the second velocity, c/n, is the definition of the index of refraction. But it is far more useful than you currently expect. Right now it says, for example, that $k = n\omega/c$: in a material the wave number is n times its value in vacuum for the same frequency, and the wavelength is λ_0/n, where $\lambda_0 = 2\pi c/\omega$ is the **vacuum wavelength** (Figure 21.2). Those are important points, but in the future it will become even more important, when we consider cases where we describe both the index of refraction and the wave number as complex quantities (Section 22.3): it will give us the relationship between $\tilde{k}$ and $\tilde{n}$, as well as, separately, the relationships between their real parts and between their imaginary parts.

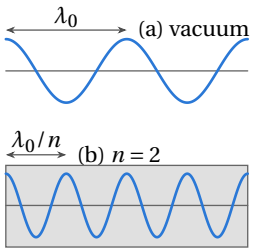


Figure 21.2 Snapshots of two waves of the same frequency, (a) in vacuum and (b) in a material with $n = 2$. By (21.5) the wave in the material has n times the wave number, so its wavelength is λ_0/n.

A wave of the same kind that travels toward $-z$ is a function of $z + vt$. This book writes it as $A\cos(-kz - \omega t + \phi)$, with the same $-\omega t$ in every wave and the direction of travel shown by the sign in front of kz (Section 18.2.4). In complex form, a sinusoidal wave is the real part of $\tilde{f} = \tilde{A}e^{i(kz-\omega t)}$, where the complex amplitude $\tilde{A} = Ae^{i\phi}$ holds both the amplitude and the phase constant (18.13), and the complex form is the one used for calculations from here on.

Sinusoidal waves matter so much because every wave can be built out of them. The wave equation is linear, so a sum of solutions is a solution, and every wave is a sum of sinusoidal waves: a wave that repeats is a Fourier series of them (Section 6.1), and a pulse that does not repeat is a continuous sum of them, a Fourier transform. So once you know what happens to every sinusoidal wave, you know what happens to every wave, as long as the medium treats all frequencies alike. Matter usually does not. Since n depends on the frequency, sinusoidal waves of different frequencies travel at different speeds, and a pulse made of them changes shape as it goes, which is called **dispersion** (Section 22.5.2).

The sinusoidal waves of this chapter are also infinite. They fill all of space and have always existed, as if they came from a source far away that has been oscillating at one frequency forever. No real wave is like that, but every real wave is a sum of such waves, so what this chapter finds applies to real waves as well.

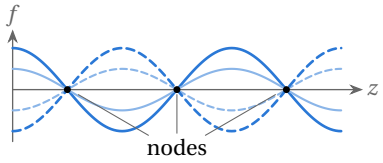


Figure 21.3 The standing wave $2A\cos kz\cos\omega t$ at $\omega t = 0$ (solid), $\pi/3$, $2\pi/3$ (light), and π (dashed). The nodes never move.

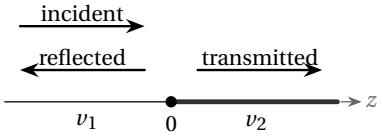


Figure 21.4 Two strings joined at $z = 0$, a light one with wave speed v_1 and a heavier one with v_2. A wave coming from the left (incident) is partly reflected and partly transmitted.

Two sinusoidal waves with the same amplitude and frequency that travel in opposite directions add up to

$$A\cos(kz-\omega t) + A\cos(-kz-\omega t) = 2A\cos kz\cos\omega t.$$

This is a **standing wave**: every point oscillates in step with every other, with an amplitude $2A|\cos kz|$ that depends on where it is, and the points where $\cos kz = 0$, half a wavelength apart, do not move at all (Figure 21.3). Nothing travels. Section 21.6.1 finds a standing wave in front of a mirror.

21.2.1 Reflection and transmission on a string

A string under tension F, with mass μ per unit length, carries transverse waves at the speed $v = \sqrt{F/\mu}$, and its sideways displacement $f(z,t)$ obeys the one-dimensional wave equation. (Here μ is a mass per unit length, not a permeability.) Tie two strings together, a light one on the left ($z < 0$) and a heavier one on the right ($z > 0$), and pull them taut. They have the same tension, but different wave speeds v_1 and v_2 (Figure 21.4). A sinusoidal wave sent along the first string reaches the knot. Some of it goes on into the second string, and some comes back. How much of each? This is the question that the rest of this chapter answers for light at the boundary between two materials, and the string shows the method with less algebra.

To the left of the knot there are two waves, the incident wave and the reflected wave, and to the right there is one, the transmitted wave:

$$\tilde{f}(z,t) = \begin{cases} \tilde{A}_I e^{i(k_1 z-\omega t)} + \tilde{A}_R e^{i(-k_1 z-\omega t)}, & z<0, \\ \tilde{A}_T e^{i(k_2 z-\omega t)}, & z>0. \end{cases}$$

All three have the frequency of whatever shakes the string far to the left. Their wave numbers are $k_1 = \omega/v_1$ and $k_2 = \omega/v_2$; the reflected wave travels in the same string as the incident wave, so it has the same wavelength. Whatever phase shifts the reflection and the transmission produce are carried by the complex amplitudes $\tilde{A}_R$ and $\tilde{A}_T$.

Two boundary conditions at the knot relate the amplitudes. First, the string does not break, so the displacement is the same just to the left and just to the right of the knot. Second, the slope $\partial f/\partial z$ is the same on the two sides. The tension pulls on the knot from both sides, and if the two slopes differed, the sideways components of the two pulls would not cancel, and a knot of negligible mass would have an infinite acceleration. At $z = 0$ the factors $e^{-i\omega t}$ cancel, and the two conditions become

$$\tilde{A}_I + \tilde{A}_R = \tilde{A}_T, \qquad k_1\left(\tilde{A}_I - \tilde{A}_R\right) = k_2\tilde{A}_T.$$

(The conditions apply to the real displacement, the real part of $\tilde{f}$. A complex $\tilde{f}$ that satisfies them at every instant has a real part that does too.) Solving for the

ratios, and using $k_1/k_2 = v_2/v_1$,

$$r \equiv \frac{\tilde{A}_R}{\tilde{A}_I} = \frac{v_2 - v_1}{v_2 + v_1}, \qquad t \equiv \frac{\tilde{A}_T}{\tilde{A}_I} = \frac{2v_2}{v_2 + v_1}. \tag{21.6}$$

These ratios are the **reflection coefficient** r and the **transmission coefficient** t (not to be confused with the time).

Here both are real. The transmitted wave is always upright, in phase with the incident wave at the knot. The reflected wave is upright when the second string carries waves faster, $v_2 > v_1$, and inverted when it carries them more slowly, $v_2 < v_1$: then r is negative, and since $-1 = e^{i\pi}$, the reflection shifts the phase by 180°. An infinitely heavy second string, which is the same as tying the first string to a wall, has $v_2 = 0$, so $r = -1$ and $t = 0$: the wave is reflected completely, upside down. For a second string with $v_2 = 2v_1$, $r = 1/3$ and $t = 4/3$; for one with $v_2 = v_1/2$, $r = -1/3$ and $t = 2/3$. Since r and t do not depend on the frequency, every sinusoidal wave in a pulse is reflected and transmitted in the same proportions, so the reflected and transmitted pulses have the shape of the incident one (Figure 21.5).

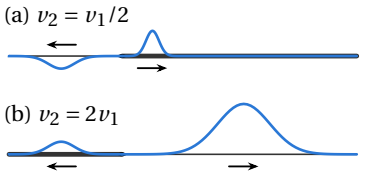


Figure 21.5 A pulse of height 1, sent from the left, after it has reached the knot. (a) A heavier second string: the reflected pulse is inverted, with height 1/3, and the transmitted pulse has height 2/3 and half the length. (b) A lighter second string: heights 1/3 and 4/3, and the transmitted pulse is twice as long.

Notice that $r + t$ is not 1, and that the transmitted wave can be larger than the incident wave. Amplitudes are not conserved; energy is. The average power that a sinusoidal wave carries along a string is

$$\langle P \rangle = \frac{1}{2}\mu\omega^2 A^2 v = \frac{F\omega^2 A^2}{2v}$$

(Exercise P21.2), where the second form uses $\mu = F/v^2$. The first form makes the power look proportional to v, but the tension and the frequency are the same on both sides of the knot, and what changes from one string to the other is v, so the second form is the useful one: the power is proportional to A^2/v. The power that arrives at the knot is the power that leaves it,

$$\frac{|\tilde{A}_I|^2}{v_1} = \frac{|\tilde{A}_R|^2}{v_1} + \frac{|\tilde{A}_T|^2}{v_2},$$

so the fractions of the incident power that are reflected and transmitted,

$$R \equiv |r|^2, \qquad T \equiv \frac{v_1}{v_2}|t|^2, \tag{21.7}$$

add up to 1. They are called the **reflectance** and the **transmittance** (also the reflectivity and the transmissivity). With (21.6),

$$R + T = \frac{(v_2 - v_1)^2 + 4v_1 v_2}{(v_2 + v_1)^2} = 1.$$

For the two examples above, $R + T = \frac{1}{9} + \frac{1}{2}\cdot\frac{16}{9} = 1$ and $R + T = \frac{1}{9} + 2\cdot\frac{4}{9} = 1$. The factor v_1/v_2 in T is easy to forget, and it has a counterpart for light, the index of refraction in the intensity (21.12).

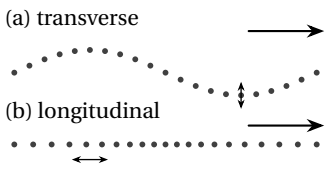


Figure 21.6 Bits of a medium as a wave passes, at one instant: (a) in a transverse wave they move sideways; (b) in a longitudinal wave they move back and forth along the direction of travel, crowding together in some places.

☞ **Caution: two meanings of "polarization."** The polarization of a wave describes the direction in which its field oscillates. This use of the word is not related to the polarization **P** of a material (Chapter 9), which is its dipole moment per unit volume.

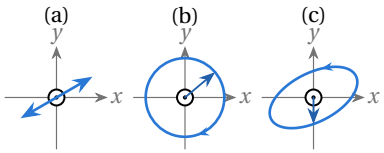


Figure 21.7 The path of the tip of **f** (or of **E**) at one point, as seen by someone the wave is coming toward, with the wave traveling out of the page: (a) linear polarization; (b) right-handed circular polarization, $\tilde{A}_y = -i\tilde{A}_x$; (c) elliptical polarization, $\tilde{A}_y = 0.7\,e^{i\pi/3}\tilde{A}_x$. The arrows on the paths show which way the tip goes around. Animation.

21.3 Polarization

A wave on a string is **transverse**: the string moves perpendicular to its length, which is the direction in which the wave travels. (In a **longitudinal** wave, such as a sound wave in air, the medium moves back and forth along the direction of travel; Figure 21.6.) The string can be displaced in any direction in the plane perpendicular to its length, so its displacement is really a vector. For a wave that travels along z,

$$\tilde{\mathbf{f}}(z,t) = \tilde{A}e^{i(kz-\omega t)}\,\hat{\mathbf{n}}, \qquad \hat{\mathbf{n}}\cdot\hat{\mathbf{z}} = 0, \tag{21.8}$$

where the unit vector $\hat{\mathbf{n}}$, perpendicular to the direction of travel, is the **polarization vector**. Section 21.4 shows that electromagnetic waves are transverse too, with the electric field in the role of **f**.

With $\hat{\mathbf{n}} = \cos\theta\,\hat{\mathbf{x}} + \sin\theta\,\hat{\mathbf{y}}$, the wave (21.8) is the sum of a wave polarized along $\hat{\mathbf{x}}$, with complex amplitude $\tilde{A}\cos\theta$, and a wave polarized along $\hat{\mathbf{y}}$, with complex amplitude $\tilde{A}\sin\theta$. The two have the same phase, and every point of the string moves back and forth along a line at the angle θ from the x axis (Figure 21.7a). Such a wave is **linearly polarized**. In general the two parts can have any amplitudes and phases,

$$\tilde{\mathbf{f}}(z,t) = \left(\tilde{A}_x\hat{\mathbf{x}} + \tilde{A}_y\hat{\mathbf{y}}\right)e^{i(kz-\omega t)},$$

so every transverse wave is a combination of two independent polarizations. What does a point of the string do when the two parts are out of phase? Take $\tilde{A}_x = A$ and $\tilde{A}_y = -iA = Ae^{-i\pi/2}$. At $z = 0$ the real displacement is

$$\mathbf{f}(0,t) = \mathrm{Re}\left[A\left(\hat{\mathbf{x}} - i\hat{\mathbf{y}}\right)e^{-i\omega t}\right] = A\left(\cos\omega t\,\hat{\mathbf{x}} - \sin\omega t\,\hat{\mathbf{y}}\right),$$

and the point goes around a circle of radius A, as if the end of the string were being swung in a circle (Figure 21.7b). This is **circular polarization**.

The displacement turns from $+x$ toward $-y$: clockwise as seen by someone the wave is coming toward. With $\tilde{A}_y = +iA$ it turns the other way, counterclockwise. The two senses are called right-handed and left-handed, but two conventions for which is which are in common use. This book follows many optics books and calls the polarization **right-handed** when, at a fixed point, the field turns clockwise as seen by someone the wave is coming toward, as in Figure 21.7b, and **left-handed** when it turns counterclockwise. The name also fits the whole wave at one instant: drawn at points along the z axis, the field vectors of a right-handed wave frozen in time have their tips on a helix with the twist of an ordinary right-handed screw (Figure 21.8). Engineers and radio astronomers often use the opposite names: for them, a wave that turns clockwise as seen by someone it is coming toward is left-handed. When you encounter circular polarization in other contexts, be sure to find out which convention is used.

When the two parts have different sizes, or a phase difference other than 0, 180°, or ±90°, the point moves around an ellipse, and the wave has **elliptical polarization** (Figure 21.7c). Linear and circular polarization are the two extreme cases of elliptical polarization. A linear polarization can also be built from two

circular ones that turn in opposite senses (Figure 21.8), with equal amplitudes (Exercise P21.6), so right- and left-handed circular polarization are just as good a pair of independent polarizations as $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$.

The light of the Sun and of an ordinary light bulb is **unpolarized**. It comes from an enormous number of atoms that radiate independently, so its polarization changes at random, far too quickly for any detector to follow, and on average no direction is preferred. Unpolarized light behaves like an equal mixture of two perpendicular linear polarizations whose relative phase wanders at random. The power that a surface reflects or transmits, for example, is the average of the powers that the two polarizations would give separately. That is true of powers, not of amplitudes: the amplitude of unpolarized light has no fixed phase to average.

You might be wondering how photons, the quantum-mechanical carriers of the energy of electromagnetic waves, fit into this description. For the most part this book, which describes classical electrodynamics, disregards photons (although Section 25.10 takes them up again), but what we have described here is the collective behavior of trillions of photons, or more. For individual photons, circular polarization is the fundamental description. A photon's spin angular momentum along its direction of travel, its **helicity**, is either $+\hbar$ or $-\hbar$, never 0 (a value that would go with a longitudinal wave). A photon of circularly polarized light has a definite helicity. In this book's convention a photon of right-handed light has negative helicity, i.e. angular momentum opposite to its direction of travel. This is often labeled σ^- (for $-\hbar$). Left-handed light has positive helicity, often labeled σ^+ (for $+\hbar$). Every observer agrees on a photon's helicity: no rotation and no change of reference frame turns a σ^+ photon into a σ^- photon.[1] That makes the helicity, like the (lack of) mass, a natural label for a photon.[2] The direction of a linear polarization, by contrast, is in general different for observers moving relative to each other, as is the direction of travel. Purely circularly polarized light is a collection of σ^+ photons or of σ^- photons. A single photon can also be in a superposition of the σ^- and σ^+ states; a superposition with equal amplitudes is a photon of linearly polarized light. The relative phase of the two σ states sets the direction of its polarization (Exercise P21.6 finds the same thing for classical waves). Unpolarized light is an equal mixture of σ^+ and σ^- photons with no phase relation between them, or, just as well, an equal mixture of horizontally and vertically polarized photons.

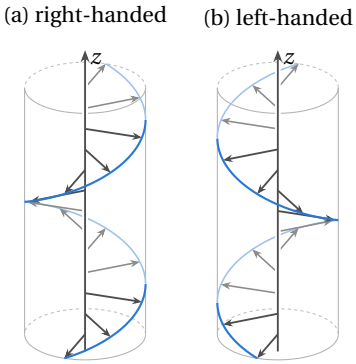


Figure 21.8 Circularly polarized waves frozen in time, traveling along $+z$: the electric field at points along the z axis, with its tips on a helix (blue). (a) A right-handed wave: a right-handed helix. (b) A left-handed wave: a left-handed helix. Animation.

☞ **Caution: two conventions for handedness.** Right- and left-handed circular polarization mean opposite things in different sources. In this book, as in many optics books, right-handed means clockwise as seen by someone the wave is coming toward; engineers and radio astronomers often call that left-handed. Find out which convention a source uses before you use its results.

[1] By contrast, an observer could overtake a spinning particle with mass, and then relative to the observer it would be moving the other direction with the opposite helicity; but nothing outruns a photon.

[2] Relativistic quantum theory classifies particles in this way, by quantities that every observer agrees on: "On Unitary Representations of the Inhomogeneous Lorentz Group," E. Wigner, Ann. Math. 40, 149–204 (1939). https://doi.org/10.2307/1968551. A photon can have either helicity because a mirror reflection, which leaves Maxwell's equations unchanged, turns one into the other.

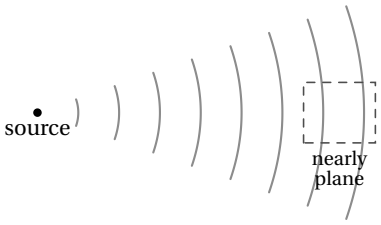


Figure 21.9 Wave fronts of a spherical wave from a small source, in a plane through the source. Far from the source a small patch of them (dashed box) is nearly flat: there the wave is nearly a plane wave.

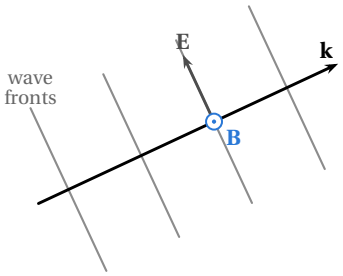


Figure 21.10 The crests of a plane wave, seen edge on, are planes perpendicular to the wave vector **k**, which points in the direction of travel. At every point the electric field (black) lies in the wave front, and the magnetic field (blue, out of the page here) is perpendicular to both **E** and **k**.

21.4 Plane electromagnetic waves

A small source, such as a light bulb seen from far away, sends out waves whose wave fronts, the surfaces on which the wave has a given phase, are spheres centered on the source. These are **spherical waves**. A small patch of a large sphere is nearly flat, so far from the source the wave is nearly a **plane wave**, a wave whose wave fronts are parallel planes (Figure 21.9). A laser beam is close to a plane wave too. Plane waves are the simplest wave solutions of Maxwell's equations, and the rest of this chapter is about them.

A sinusoidal plane wave that travels in the direction of the unit vector $\hat{\mathbf{k}}$ has the fields

$$\tilde{\mathbf{E}}(\mathbf{r},t) = \tilde{\mathbf{E}}_0\, e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}, \qquad \tilde{\mathbf{B}}(\mathbf{r},t) = \tilde{\mathbf{B}}_0\, e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}, \tag{21.9}$$

where the constant complex vectors $\tilde{\mathbf{E}}_0$ and $\tilde{\mathbf{B}}_0$ hold the amplitudes, the phases, and the directions of the fields, and the **wave vector**

$$\mathbf{k} = k\hat{\mathbf{k}} = k_x\hat{\mathbf{x}} + k_y\hat{\mathbf{y}} + k_z\hat{\mathbf{z}}$$

points in the direction of travel, with $|\mathbf{k}| = k = 2\pi/\lambda$. The real fields are the real parts of (21.9). The dot product $\mathbf{k}\cdot\mathbf{r}$ is k times the distance of the point $\mathbf{r}$ along $\hat{\mathbf{k}}$, so the phase $\mathbf{k}\cdot\mathbf{r}-\omega t$ is the same everywhere on a plane perpendicular to $\mathbf{k}$. Those planes are the wave fronts, and they move along $\hat{\mathbf{k}}$ at the speed ω/k (Figure 21.10). For $\mathbf{k} = k\hat{\mathbf{z}}$, (21.9) is the wave of Section 21.2 with $\hat{\mathbf{n}}$ in the direction of $\tilde{\mathbf{E}}_0$.

Each Cartesian component of the fields (21.9) satisfies the wave equation (21.2) as long as $\omega/k = v$, since the Laplacian multiplies the exponential by $-k^2$ and the second time derivative multiplies it by $-\omega^2$. But that is not enough. The wave equations follow from Maxwell's equations, not the other way around, and the fields must also satisfy the four equations (21.1) themselves. For the exponential in (21.9), the derivative with respect to x brings down a factor ik_x, and similarly for y and z, so

$$\nabla \to i\mathbf{k}, \qquad \frac{\partial}{\partial t} \to -i\omega,$$

the three-dimensional version of (18.14). The two divergence equations become $i\mathbf{k}\cdot\tilde{\mathbf{E}}_0 = 0$ and $i\mathbf{k}\cdot\tilde{\mathbf{B}}_0 = 0$, or

$$\mathbf{k}\cdot\tilde{\mathbf{E}}_0 = 0, \qquad \mathbf{k}\cdot\tilde{\mathbf{B}}_0 = 0.$$

With words, both fields are perpendicular to the direction of travel: electromagnetic waves are transverse. Faraday's law becomes $i\mathbf{k}\times\tilde{\mathbf{E}}_0 = i\omega\tilde{\mathbf{B}}_0$, so

$$\boxed{\tilde{\mathbf{B}}_0 = \frac{1}{\omega}\mathbf{k}\times\tilde{\mathbf{E}}_0 = \frac{1}{v}\hat{\mathbf{k}}\times\tilde{\mathbf{E}}_0}. \tag{21.10}$$

The Ampère–Maxwell law, $i\mathbf{k}\times\tilde{\mathbf{B}}_0 = -i\omega\mu\varepsilon\,\tilde{\mathbf{E}}_0$, adds nothing new. With (21.10) its left side is $(i/\omega)\,\mathbf{k}\times(\mathbf{k}\times\tilde{\mathbf{E}}_0) = -(ik^2/\omega)\,\tilde{\mathbf{E}}_0$, by the BAC-CAB rule (1.49) and $\mathbf{k}\cdot\tilde{\mathbf{E}}_0 = 0$, so it says only that $k^2 = \mu\varepsilon\omega^2$, which the wave equation already required.

So in a plane wave **B** is perpendicular to both **k** and **E**, and its size is $|\mathbf{E}|/v$, which is $|\mathbf{E}|/c$ in vacuum. Since k and ω are real, $\tilde{\mathbf{B}}_0$ has the same phase as $\tilde{\mathbf{E}}_0$: the two fields rise and fall together, reaching their largest values at the same places and the same times. (In a conductor **k** is complex, and **B** lags behind **E**; Section 22.2.3.) The cross product $\mathbf{E}\times\mathbf{B}$, and with it the Poynting vector, points along $\hat{\mathbf{k}}$, the direction in which the wave travels. Exercise P16.13 found all of this for one wave, with **E** along $\hat{\mathbf{x}}$ and **k** along $\hat{\mathbf{z}}$.

21.4.1 The impedance of a material

Inside the material $\mathbf{H} = \mathbf{B}/\mu$, so by (21.10) the ratio of the sizes of **E** and **H** in a plane wave depends only on the material:

$$\frac{|\tilde{\mathbf{E}}_0|}{|\tilde{\mathbf{H}}_0|} = \mu v = \sqrt{\frac{\mu}{\varepsilon}} \equiv Z. \tag{21.11}$$

Z is the material's **wave impedance**. Since **E** is measured in volts per meter and **H** in amperes per meter, Z is measured in ohms. In vacuum it is the **impedance of free space**,

$$Z_0 = \sqrt{\frac{\mu_0}{\varepsilon_0}} = \mu_0 c = 376.73\,\Omega,$$

and in a material that is not magnetic, $Z = Z_0/n$. The impedance comes back in the reflection of waves (Section 21.6), in transmission lines (Section 23.8), and in the radiation resistance of antennas (Section 28.2.3).

Example 21.1

A plane wave whose vacuum wavelength is 500 nm travels through glass with $n = 1.50$, which is not magnetic, in the direction of the unit vector $\hat{\mathbf{k}} = (3\hat{\mathbf{x}} + 4\hat{\mathbf{z}})/5$. Its electric field is along $\hat{\mathbf{y}}$, with an amplitude of 100 V/m. Find the wave's frequency, its wavelength in the glass, and its wave vector; write **E** and **B** as real functions of position and time, taking the phase constant to be zero; and find the amplitude of **H**.

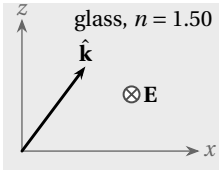
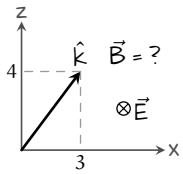


Figure 21.11 The direction of travel of the wave of Example 21.1, in the xz plane, and its electric field, along $\hat{\mathbf{y}}$ (into the page).

Solution:



The first step is a sketch of what is given and what is asked: the direction of travel $\hat{\mathbf{k}}$, in the xz plane, and **E** along $\hat{\mathbf{y}}$, into the page. **B** is still to be found, but it must be perpendicular to both of them.

The frequency is set by the source, and the wave keeps it when it enters the glass (Section 21.6 shows why). From the vacuum wavelength,

$$\nu = \frac{c}{\lambda_0} = \frac{2.998\times10^8\,\text{m/s}}{500\times10^{-9}\,\text{m}} = 6.00\times10^{14}\,\text{Hz}, \qquad \omega = 2\pi\nu = 3.77\times10^{15}\,\text{s}^{-1}.$$

In the glass the wave travels at $v = c/n = 2.00\times10^8\,\text{m/s}$, so by (21.5) its wavelength is $\lambda = \lambda_0/n = 333\,\text{nm}$, and

$$k = \frac{2\pi}{\lambda} = 1.885\times10^7\,\text{m}^{-1}, \qquad \mathbf{k} = k\,\frac{3\hat{\mathbf{x}}+4\hat{\mathbf{z}}}{5} = (1.131\hat{\mathbf{x}}+1.508\hat{\mathbf{z}})\times10^7\,\text{m}^{-1}.$$

$\mathbf{E}$ is perpendicular to $\mathbf{k}$, as it must be, since $\hat{\mathbf{y}}\cdot\hat{\mathbf{k}} = 0$. With the phase constant zero,

$$\mathbf{E}(\mathbf{r},t) = (100\,\text{V/m})\cos(\mathbf{k}\cdot\mathbf{r}-\omega t)\,\hat{\mathbf{y}}, \qquad \mathbf{k}\cdot\mathbf{r} = (1.131x+1.508z)\times10^7\,\text{m}^{-1}.$$

By (21.10), $\mathbf{B}$ has the size $E_0/v = (100\,\text{V/m})/(2.00\times10^8\,\text{m/s}) = 5.00\times10^{-7}\,\text{T}$ and the direction

$$\hat{\mathbf{k}}\times\hat{\mathbf{y}} = \frac{3\hat{\mathbf{x}}\times\hat{\mathbf{y}}+4\hat{\mathbf{z}}\times\hat{\mathbf{y}}}{5} = \frac{-4\hat{\mathbf{x}}+3\hat{\mathbf{z}}}{5},$$

so

$$\mathbf{B}(\mathbf{r},t) = (0.500\,\mu\text{T})\cos(\mathbf{k}\cdot\mathbf{r}-\omega t)\,(-0.8\hat{\mathbf{x}}+0.6\hat{\mathbf{z}}).$$

The wave impedance of the glass is $Z = Z_0/n = 251\,\Omega$, so $H_0 = E_0/Z = 0.398\,\text{A/m}$, which is also B_0/μ_0.

Checks: $\mathbf{B}$ is perpendicular to $\mathbf{k}$, since $3(-4)+4(3) = 0$, and to $\mathbf{E}$, since it has no y component. And $\mathbf{E}\times\mathbf{B}$ points along $\hat{\mathbf{y}}\times(-4\hat{\mathbf{x}}+3\hat{\mathbf{z}})/5 = (3\hat{\mathbf{x}}+4\hat{\mathbf{z}})/5 = \hat{\mathbf{k}}$, the direction of travel, as it must.

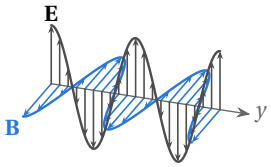


Figure 21.12 A linearly polarized plane wave that travels along y, at one instant: $\mathbf{E}$ (black) along z and $\mathbf{B}$ (blue) along x rise and fall together. The curves show the fields only at points on the y axis; at each instant the fields are the same at every point of a plane perpendicular to y.

21.5 Energy and momentum in waves

In empty space the energy density of the fields is $u = \frac{1}{2}\left(\varepsilon_0 E^2 + B^2/\mu_0\right)$, (20.11). In a plane wave $B = E/c$ at every point and every instant (Figure 21.12), and

$$\frac{B^2}{2\mu_0} = \frac{E^2}{2\mu_0 c^2} = \frac{1}{2}\varepsilon_0 E^2,$$

since $1/c^2 = \mu_0\varepsilon_0$: the electric and magnetic fields of the wave hold exactly the same amount of energy. For the wave $\mathbf{E} = E_0\cos(\mathbf{k}\cdot\mathbf{r}-\omega t+\phi)\,\hat{\mathbf{n}}$, then,

$$u = \varepsilon_0 E^2 = \varepsilon_0 E_0^2\cos^2(\mathbf{k}\cdot\mathbf{r}-\omega t+\phi), \qquad \langle u\rangle = \frac{1}{2}\varepsilon_0 E_0^2,$$

since the average of $\cos^2$ over a cycle is $\frac{1}{2}$. The Poynting vector (20.12) is

$$\mathbf{S} = \frac{1}{\mu_0}\mathbf{E}\times\mathbf{B} = \frac{E^2}{\mu_0 c}\hat{\mathbf{k}} = c\,u\,\hat{\mathbf{k}}.$$

With words, the energy of the wave moves along $\hat{\mathbf{k}}$ at the speed c. Its average is $\langle\mathbf{S}\rangle = c\langle u\rangle\hat{\mathbf{k}}$, and its size, the intensity, is $I = \frac{1}{2}c\varepsilon_0 E_0^2$, as Section 20.2.4 found from the complex amplitudes.

In a linear material the energy density and the Poynting vector are $u = \frac{1}{2}(\mathbf{E}\cdot\mathbf{D}+\mathbf{B}\cdot\mathbf{H}) = \frac{1}{2}(\varepsilon E^2 + B^2/\mu)$ and $\mathbf{S} = \mathbf{E}\times\mathbf{H}$, (20.16). With $B = E/v$ the two halves of

u are again equal, since $1/\mu v^2 = \varepsilon$, and $\mathbf{S} = (E^2/\mu v)\hat{\mathbf{k}} = v\,u\hat{\mathbf{k}}$: the energy travels at the speed v of the wave. The intensity is

$$I = v\langle u\rangle = \frac{1}{2}\varepsilon v E_0^2 = \frac{E_0^2}{2Z}. \tag{21.12}$$

For a material that is not magnetic, $\varepsilon v = \varepsilon_0 n^2 (c/n) = n\varepsilon_0 c$, so $I = \frac{1}{2} nc\varepsilon_0 E_0^2$. A wave with a given amplitude of $\mathbf{E}$ carries n times as much power in a material as in vacuum. This factor of n matters whenever the powers of waves in two different materials are compared (Section 21.6), just as the factor $1/v$ in the power of a wave on a string did in Section 21.2.1.

21.5.1 Momentum and radiation pressure

Fields carry momentum as well as energy. Section 20.3.4 shows that in empty space their momentum per unit volume, the momentum density, is

$$\mathbf{g} = \mu_0\varepsilon_0\mathbf{S} = \frac{1}{c^2}\mathbf{S}. \tag{21.13}$$

For a plane wave in vacuum, $\mathbf{S} = c\,u\hat{\mathbf{k}}$, so $\mathbf{g} = (u/c)\hat{\mathbf{k}}$: the momentum density is the energy density divided by c, pointing along the direction of travel. That is the relation $E = pc$ between the energy and the momentum of a particle with no mass (Section 25.10), which is what a photon is. The average momentum density is $\langle g\rangle = \langle u\rangle/c = I/c^2$.

When a wave strikes a surface that absorbs it, its momentum goes into the surface, which feels a force. Take a plane wave in vacuum that strikes a perfect absorber head on, and a patch of the surface with area A (Figure 21.13). In a time Δt the wave delivers to the patch all of the momentum in a column of cross section A and length $c\,\Delta t$, which is $\langle g\rangle Ac\,\Delta t$. The force on the patch is this momentum divided by Δt, and the force per unit area, the **radiation pressure**, is

$$P_{\text{rad}} = \frac{\langle g\rangle Ac\,\Delta t}{A\,\Delta t} = c\langle g\rangle = \frac{I}{c} \qquad \text{(perfect absorber)}. \tag{21.14}$$

A perfect mirror sends the wave straight back, reversing its momentum, so the change in the wave's momentum is twice as large, and the pressure is $2I/c$. On the absorber the pressure equals the average energy density of the wave, $I/c = \langle u\rangle$, and on the mirror it is twice that; a pressure and an energy density have the same units, $\text{N/m}^2 = \text{J/m}^3$.

Sunlight above Earth's atmosphere, with $I = 1361\,\text{W/m}^2$ (Section 20.2.4), pushes on a black surface with a pressure of $I/c = 4.54\times10^{-6}\,\text{Pa}$, less than 10^{-10} of atmospheric pressure. Usually that is negligible, but not always.

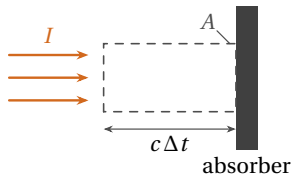


Figure 21.13 A plane wave strikes a perfect absorber head on. In a time Δt, the momentum in a column of length $c\,\Delta t$ (dashed) goes into the patch of area A at its end.

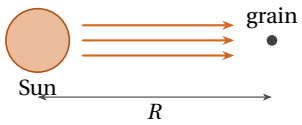


Figure 21.14 The dust grain of Example 21.2, in the sunlight at Earth's distance from the Sun (not to scale).

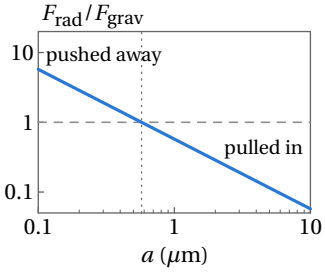
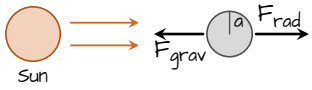


Figure 21.15 The ratio of the push of sunlight to the pull of the Sun's gravity on the grains of Example 21.2, against the radius a of a grain. Grains smaller than 0.57 μm are pushed away from the Sun.

Example 21.2

A small grain of dust in space is pushed away from the Sun by sunlight and pulled toward it by the Sun's gravity. Model the grain as a sphere of radius a and density $\rho_m = 1000\,\text{kg/m}^3$ (that of water) that absorbs all the sunlight that falls on it. Find the radius for which the two forces are equal, and say what happens to smaller grains. At Earth's distance from the Sun, $1.496 \times 10^{11}\,\text{m}$, sunlight has the intensity $1361\,\text{W/m}^2$; find the Sun's gravitational pull there from Earth's orbit, a nearly circular path of that radius that it completes once a year.[3]

Solution:



The first step is a sketch of what is given and what is asked: the grain, the sunlight that falls on it, and the two forces to compare, the push of the light, away from the Sun, and the pull of gravity, toward it.

The grain blocks the light that falls on a disk of area πa^2, its cross section, so by (21.14) the force of the light on it is

$$F_{rad} = \frac{I}{c}\pi a^2.$$

The Sun's gravity gives everything at Earth's distance the same acceleration $g_\odot$, the centripetal acceleration of Earth's orbit. With $R = 1.496 \times 10^{11}\,\text{m}$ and one year, $T = 3.156 \times 10^7\,\text{s}$,

$$g_\odot = \left(\frac{2\pi}{T}\right)^2 R = 5.93 \times 10^{-3}\,\text{m/s}^2,$$

and the grain, of mass $\frac{4}{3}\pi a^3 \rho_m$, is pulled toward the Sun with the force $F_{grav} = \frac{4}{3}\pi a^3 \rho_m g_\odot$. The two forces are equal when

$$\frac{I}{c}\pi a^2 = \frac{4}{3}\pi a^3 \rho_m g_\odot, \qquad \text{or} \qquad a = \frac{3I}{4c\rho_m g_\odot},$$

which is

$$a = \frac{3(1361\,\text{W/m}^2)}{4(2.998 \times 10^8\,\text{m/s})(1000\,\text{kg/m}^3)(5.93 \times 10^{-3}\,\text{m/s}^2)} = 0.57\,\mu\text{m}.$$

The ratio of the two forces, F_{rad}/F_{grav}, is $(0.57\,\mu\text{m})/a$ (Figure 21.15). Both forces fall off as the inverse square of the distance from the Sun, the intensity because the Sun's power spreads over larger and larger spheres, and gravity by Newton's law, so the ratio is the same at every distance. A grain smaller than about half a micrometer is pushed away from the Sun more strongly than it is pulled, wherever it is, and it leaves the solar system. Sunlight's push on small grains is why the dust tail of a comet points away from the Sun.

[3] The distance is one astronomical unit, defined to be exactly 149 597 870 700 m: "New IAU definition of the astronomical unit of length," International Astronomical Union (2012), https://syrte.obspm.fr/IAU_resolutions/IAUResol_2012_0.html.

Checks: the units of $I/c\rho_m g_\odot$ are $(\mathrm{N/m^2})/(\mathrm{N/m^3}) = \mathrm{m}$. A grain this small is about as large as the wavelength of sunlight, and for such a grain the light that it absorbs or scatters is not exactly the light that falls on its geometric cross section, so the answer is an estimate of the size, not an exact value.

21.6 Reflection and transmission at normal incidence

Now for the question that Section 21.2.1 answered for a string. A plane wave travels through one material and reaches a flat boundary with another: how much of it is reflected, and how much is transmitted? Let the boundary be the plane $z = 0$, with material 1 (permittivity ε_1, permeability μ_1, wave speed v_1) for $z < 0$ and material 2 for $z > 0$. The simplest case is a wave that strikes the boundary head on, traveling along $+z$. This is called **normal incidence**.

At normal incidence every direction perpendicular to z is like every other, so take the electric field of the incident wave to be along $\hat{\mathbf{x}}$. Take the electric fields of the reflected and transmitted waves along $\hat{\mathbf{x}}$ too; if one of them really points along $-\hat{\mathbf{x}}$, its amplitude will come out negative. (Parts along $\hat{\mathbf{y}}$ would obey the same two boundary conditions below with no incident wave to drive them, and they come out zero.) Each magnetic field then follows from (21.10), $\tilde{\mathbf{B}}_0 = (1/v)\,\hat{\mathbf{k}} \times \tilde{\mathbf{E}}_0$. For the reflected wave $\hat{\mathbf{k}} = -\hat{\mathbf{z}}$, so its magnetic field points along $-\hat{\mathbf{y}}$ (Figure 21.16):

$$\tilde{\mathbf{E}}_I = \tilde{E}_{0I}\, e^{i(k_1 z - \omega t)}\,\hat{\mathbf{x}}, \qquad \tilde{\mathbf{B}}_I = \frac{1}{v_1}\,\tilde{E}_{0I}\, e^{i(k_1 z - \omega t)}\,\hat{\mathbf{y}},$$

$$\tilde{\mathbf{E}}_R = \tilde{E}_{0R}\, e^{i(-k_1 z - \omega t)}\,\hat{\mathbf{x}}, \qquad \tilde{\mathbf{B}}_R = -\frac{1}{v_1}\,\tilde{E}_{0R}\, e^{i(-k_1 z - \omega t)}\,\hat{\mathbf{y}},$$

$$\tilde{\mathbf{E}}_T = \tilde{E}_{0T}\, e^{i(k_2 z - \omega t)}\,\hat{\mathbf{x}}, \qquad \tilde{\mathbf{B}}_T = \frac{1}{v_2}\,\tilde{E}_{0T}\, e^{i(k_2 z - \omega t)}\,\hat{\mathbf{y}},$$

with $k_1 = \omega/v_1$ and $k_2 = \omega/v_2$.

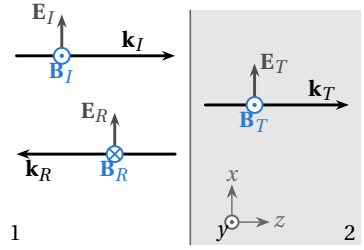


Figure 21.16 A wave at normal incidence on the boundary between material 1 and material 2 (gray). The electric fields of all three waves are taken along $+\hat{\mathbf{x}}$; each magnetic field then follows from $\hat{\mathbf{k}} \times \mathbf{E}$, so the reflected wave's points into the page.

All three waves have the same frequency. The boundary conditions require the fields just to the left of the boundary, the sums of the incident and reflected fields, to match the fields just to the right of it at every instant. That would be impossible if the waves had different frequencies, since a sum of oscillations at two frequencies cannot equal an oscillation at a third at every instant. So a wave keeps the frequency of its source when it crosses a boundary. Its wavelength changes, since $k = n\omega/c$ by (21.5).

With no free charge or current on the boundary, the conditions (19.4) say that $\mathbf{E}^{\parallel}$ and $\mathbf{B}^{\parallel}/\mu$ are the same on the two sides. (The conditions on $\varepsilon E^{\perp}$ and $B^{\perp}$ say nothing here, since none of the fields has a component along z.) At $z = 0$ every exponential is $e^{-i\omega t}$, which cancels, and the two conditions become

$$\tilde{E}_{0I} + \tilde{E}_{0R} = \tilde{E}_{0T}, \qquad \frac{1}{\mu_1 v_1}\left(\tilde{E}_{0I} - \tilde{E}_{0R}\right) = \frac{1}{\mu_2 v_2}\,\tilde{E}_{0T}.$$

It helps to name the ratio that appears in the second one,

$$\beta \equiv \frac{\mu_1 v_1}{\mu_2 v_2} = \frac{\mu_1 n_2}{\mu_2 n_1} = \frac{Z_1}{Z_2}, \tag{21.15}$$

where Z_1 and Z_2 are the wave impedances (21.11) of the two materials. Then the second condition is $\tilde{E}_{0I} - \tilde{E}_{0R} = \beta \tilde{E}_{0T}$. Adding the two conditions gives $2\tilde{E}_{0I} = (1+\beta)\tilde{E}_{0T}$, and putting $\tilde{E}_{0T}$ back into the first condition gives $\tilde{E}_{0R}$. The reflection and transmission coefficients are

$$\boxed{r \equiv \frac{\tilde{E}_{0R}}{\tilde{E}_{0I}} = \frac{1-\beta}{1+\beta}, \qquad t \equiv \frac{\tilde{E}_{0T}}{\tilde{E}_{0I}} = \frac{2}{1+\beta}}. \tag{21.16}$$

Both are real, so at the boundary the reflected and transmitted waves are either in phase with the incident wave or exactly out of phase with it. For most materials $\mu_1 \approx \mu_2 \approx \mu_0$, so that $\beta = v_1/v_2 = n_2/n_1$, and

$$r = \frac{n_1 - n_2}{n_1 + n_2} = \frac{v_2 - v_1}{v_2 + v_1}, \qquad t = \frac{2n_1}{n_1 + n_2} = \frac{2v_2}{v_2 + v_1},$$

exactly the coefficients (21.6) of the string. When light goes into a material with a larger index of refraction, $n_2 > n_1$, as from air into glass, r is negative: the reflected wave is inverted, shifted in phase by 180°, like the wave on a string that runs into a heavier string. When it goes into a material with a smaller index, the reflected wave is upright. The transmitted wave is always upright.

The intensity of each wave is $|\tilde{E}_0|^2/2Z$ (21.12), with the impedance of the material in which it travels. So the reflectance and the transmittance, the fractions of the incident power that are reflected and transmitted, are

$$R = |r|^2 = \left(\frac{1-\beta}{1+\beta}\right)^2, \qquad T = \frac{Z_1}{Z_2}|t|^2 = \frac{4\beta}{(1+\beta)^2}, \tag{21.17}$$

and $R + T = 1$, since $(1-\beta)^2 + 4\beta = (1+\beta)^2$: the power that reaches the boundary is the power that leaves it. For materials that are not magnetic,

$$R = \left(\frac{n_1 - n_2}{n_1 + n_2}\right)^2, \qquad T = \frac{4n_1 n_2}{(n_1 + n_2)^2}.$$

The factor $Z_1/Z_2 = \beta$ in T, which is n_2/n_1 for such materials, is the counterpart of the factor v_1/v_2 for the string (21.7): a wave with a given amplitude of **E** carries more power in a material with a larger index of refraction (Section 21.5). Notice also that R and T do not change when β is replaced by $1/\beta$ (Figure 21.17), so a boundary reflects the same fraction of the power whichever way a wave crosses it.

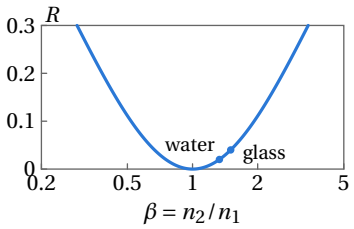


Figure 21.17 The reflectance at normal incidence, (21.17), for materials that are not magnetic, against $\beta = n_2/n_1$ on a logarithmic scale. It is the same for β and $1/\beta$. The dots mark light going from air into water and into glass.

Example 21.3

The beam of a 100 mW laser, traveling in air, strikes a flat glass plate head on. The glass has $n = 1.50$ and is not magnetic. Find the reflection and transmission coefficients and the powers of the reflected and transmitted beams. Then find the same quantities at the back surface of the plate, where the beam leaves the glass for air, ignoring multiple reflections inside the plate.

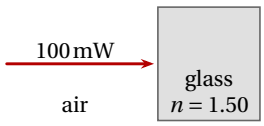
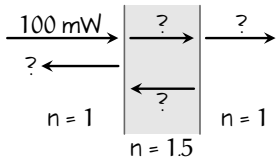


Figure 21.18 The laser beam and the glass plate of Example 21.3.

Solution:



The first step is a sketch of what is given and what is asked: the beam that arrives from the air with 100 mW, and the four beams whose powers we want, reflected and transmitted at the front surface and at the back surface. (They all lie on one line; the sketch draws them apart.)

At the front surface the light goes from air into glass, so $\beta = n_2/n_1 = 1.50$, and by (21.16)

$$r = \frac{1-1.50}{1+1.50} = -0.200, \qquad t = \frac{2}{1+1.50} = 0.800.$$

The reflected wave is inverted. By (21.17), $R = r^2 = 0.0400$ and $T = \beta t^2 = 1.50(0.800)^2 = 0.960$, so the reflected beam carries 4.00 mW and the transmitted beam 96.0 mW.

At the back surface the light goes from glass into air, so $\beta = 1/1.50 = 0.667$, and

$$r = \frac{1-0.667}{1+0.667} = +0.200, \qquad t = \frac{2}{1+0.667} = 1.200.$$

Now the reflected wave is upright, and the transmitted wave has a larger amplitude than the wave in the glass. Still $R = 0.0400$ and $T = 0.667(1.200)^2 = 0.960$: 3.84 mW is reflected back into the glass, and 92.2 mW leaves it.

So each surface between air and glass reflects 4% of the light that reaches it, whichever way the light crosses it, and a window loses about 8% of the light that falls on it. A camera lens made of several pieces of glass has ten or more such surfaces, and without the help of anti-reflection (AR) coatings, which create destructive interference between reflected waves, it would lose a third of the light: $0.96^{10} = 0.66$. AR coatings are a subject of optics.[4] And with a strong laser, a reflection of a few percent from a window or a lens can still harm an eye, so laser users must always be careful of stray reflections.

Checks: $R + T = 1$ at both surfaces. The transmission coefficient at the back surface is larger than 1, but the transmitted power is not larger than the power that arrives: the

[4] E.g., J. Peatross and M. Ware, Physics of Light and Optics, 2015 edition, Chapter 4, available at https://optics.byu.edu.

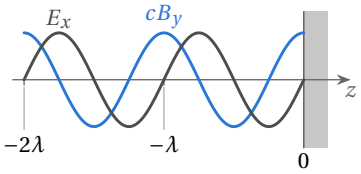


Figure 21.19 The standing wave (21.18) in front of a mirror (gray): E_x (black) at $\omega t = \pi/2$ and cB_y (blue) at $\omega t = 0$, the instants when each is largest. The electric field is always zero at the mirror's surface, and the magnetic field is largest there.

same amplitude carries less power in air than in glass, by the factor $1/1.50$ (21.12), so $T = \beta t^2$ is still less than 1, as on the string.

21.6.1 Reflection from a mirror

Good conductors act as mirrors.[5] A perfect conductor (Section 19.1.3) reflects a wave completely (Figure 21.19). The electric field inside it is zero, and $\mathbf{E}^{\parallel}$ is continuous, so at its surface $\tilde{E}_{0I} + \tilde{E}_{0R} = 0$:

$$r = -1, \qquad t = 0.$$

The condition on $\mathbf{B}^{\parallel}$ does not hold as it stands here, since the wave drives a current in the surface of the conductor, and that current makes $\mathbf{B}^{\parallel}$ jump (Table 19.1).

In front of the mirror the incident and reflected waves add up. For a wave in vacuum, with the real amplitude $\tilde{E}_{0I} = E_0$, $\tilde{E}_{0R} = -E_0$, and $k = \omega/c$, the electric field is

$$\tilde{\mathbf{E}} = E_0 \left(e^{ikz} - e^{-ikz} \right) e^{-i\omega t}\, \hat{\mathbf{x}} = 2iE_0 \sin kz\, e^{-i\omega t}\, \hat{\mathbf{x}},$$

and the magnetic fields of the two waves add in the same way, with a plus sign in place of the minus. The real fields are

$$\mathbf{E} = 2E_0 \sin kz \sin \omega t\, \hat{\mathbf{x}}, \qquad \mathbf{B} = \frac{2E_0}{c} \cos kz \cos \omega t\, \hat{\mathbf{y}} \qquad (z < 0). \tag{21.18}$$

This is a standing wave like the one of Section 21.2. The electric field is always zero at the surface of the mirror, as it must be, and at every half wavelength from it, while the magnetic field is largest at the mirror (Figure 21.19). The two fields are a quarter of a wavelength apart in space and a quarter of a period apart in time. The energy passes back and forth between the electric and the magnetic field, and the Poynting vector averages to zero: nothing travels. Two parallel mirrors make the cavity of a laser, and Exercise P21.11 finds which standing waves fit between them. The same exercise finds the radiation pressure $2I/c$ on a mirror (Section 21.5.1) a second way, as the force of the magnetic field on the current in the mirror's surface.

21.7 Oblique incidence: the laws of optics and the Fresnel equations

Most light does not strike a surface head on. Let a plane wave in material 1 meet the boundary $z = 0$ with its direction of travel at the angle θ_I from the normal. (The angles of reflection and refraction are measured from the normal too, by long custom.) Orient the x axis so that the incident wave vector lies in the xz plane,

$$\mathbf{k}_I = k_1 \left(\sin\theta_I\, \hat{\mathbf{x}} + \cos\theta_I\, \hat{\mathbf{z}} \right),$$

[5] The glass on your bathroom mirror is only there to support and protect the metal.

and for now let the reflected and transmitted waves have wave vectors $\mathbf{k}_R$ and $\mathbf{k}_T$ in any directions, and any polarizations (Figure 21.20).

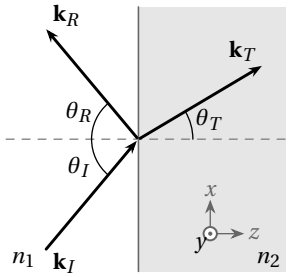


Figure 21.20 The wave vectors of the incident, reflected, and transmitted waves at a boundary (material 2 gray, with $n_2 > n_1$). All three lie in the plane of incidence, the plane of the page, and the angles are measured from the normal (dashed).

21.7.1 The laws of reflection and refraction

At $z = 0$ each boundary condition takes the form

$$\tilde{A}_1 e^{i(\mathbf{k}_I \cdot \mathbf{r} - \omega t)} + \tilde{A}_2 e^{i(\mathbf{k}_R \cdot \mathbf{r} - \omega t)} = \tilde{A}_3 e^{i(\mathbf{k}_T \cdot \mathbf{r} - \omega t)},$$

where $\tilde{A}_1$, $\tilde{A}_2$, and $\tilde{A}_3$ are made of the amplitudes and polarization vectors of the waves, which do not depend on position or time, and $\mathbf{r} = x\hat{\mathbf{x}} + y\hat{\mathbf{y}}$ is a point of the boundary. The equation must hold at every point of the boundary and at every instant. In Section 21.6 the time dependence showed that the frequencies are equal: an equation such as $5\cos 3t = A\cos\omega t$ can hold at every instant only if $\omega = 3$. In the same way, the equation above can hold at every x and y only if the three exponents are equal on the whole boundary,

$$\mathbf{k}_I \cdot \mathbf{r} = \mathbf{k}_R \cdot \mathbf{r} = \mathbf{k}_T \cdot \mathbf{r} \qquad \text{for every } x \text{ and } y,$$

so the three wave vectors have the same x components, and the same y components, which are zero. Three laws follow.

First, since $\mathbf{k}_R$ and $\mathbf{k}_T$ have no y components, the incident, reflected, and transmitted waves all travel in one plane, the plane of $\mathbf{k}_I$ and the normal, which is called the **plane of incidence**.

Second, the reflected wave travels in the same material as the incident wave, with the same frequency, so its wave vector has the same length k_1, and it has the same x component: $k_1 \sin\theta_R = k_1 \sin\theta_I$, or

$$\theta_R = \theta_I.$$

This is the **law of reflection**.

Third, the transmitted wave has $k_T = k_2$, and $k_1 \sin\theta_I = k_2 \sin\theta_T$. Since $k = n\omega/c$ by (21.5),

$$\boxed{n_1 \sin\theta_I = n_2 \sin\theta_T}. \tag{21.19}$$

This is the **law of refraction**, or **Snell's law**.

These three laws are the basis of geometrical optics: ray tracing, lenses, and the formation of images all follow from them. Their derivation used nothing but the form of the waves and the fact that something must be continuous at the boundary, so they hold for waves of every kind. A wave bends at a boundary because it travels at different speeds on the two sides. The part of each wave front that enters the slower material first slows down first, and the wave front turns, as a car pulls toward the side whose wheels have run onto a soft shoulder.

When $n_1 > n_2$, as for light going from glass into air, Snell's law gives $\sin\theta_T = (n_1/n_2)\sin\theta_I$, which reaches 1 at the **critical angle**,

$$\sin\theta_c = \frac{n_2}{n_1}.$$



Willebrord Snellius (1580–1626, Dutch) Willebrord Snel van Royen, who wrote under the Latin name Snellius, was born in Leiden, and in 1613 he succeeded his father as professor of mathematics at the University of Leiden. In 1615 he measured the size of the Earth by triangulation, through a chain of triangles between fourteen points, mostly church towers, and he published the result in his *Eratosthenes Batavus* (1617). He found the law of refraction in 1621 but did not publish it. Ibn Sahl had found it about 984, and Descartes published it in 1637. (Wikipedia)

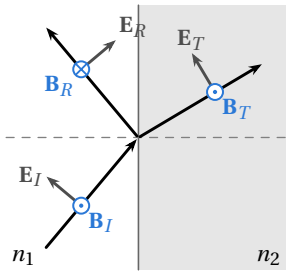


Figure 21.21 The fields of p-polarization, with the electric fields (black) in the plane of incidence. The directions are the ones assumed in the text; the reflected wave's **E** is chosen so that it is parallel to the incident wave's at normal incidence.



Augustin-Jean Fresnel (1788–1827, French) Fresnel was born in Broglie, in Normandy, studied at the École polytechnique and the École des ponts et chaussées, and worked all his life as a civil engineer for the French state. His wave theory of diffraction won the French Academy's prize in 1819, and by supposing that the waves of light are purely transverse he explained polarization. He read the equations named after him to the Academy in January 1823, and in July of that year the first lighthouse lens of his design was lit at Cordouan, at the mouth of the Gironde. He died of tuberculosis at thirty-nine, eight days after Arago brought him the Royal Society's Rumford Medal. (Wikipedia)

There the transmitted wave travels along the boundary. At larger angles of incidence no real angle θ_T satisfies Snell's law, and the wave is reflected completely. This total internal reflection is taken up in Section 22.1. For glass with $n = 1.50$ in air, $\theta_c = 41.8°$.

21.7.2 The Fresnel equations

The laws of optics give the directions of the reflected and transmitted waves. The boundary conditions also give their amplitudes, and for those the polarization matters. Every polarization is a combination of two kinds. In **p-polarization** the electric field lies in the plane of incidence (p for parallel), and in **s-polarization** it is perpendicular to the plane of incidence (s for *senkrecht*, the German word for perpendicular). At normal incidence there is no plane of incidence, and the two cannot be told apart.

Take p-polarization. The electric field of each wave lies in the xz plane, perpendicular to its wave vector, and its magnetic field, $(1/v)\hat{\mathbf{k}} \times \mathbf{E}$ by (21.10), lies along $\pm\hat{\mathbf{y}}$ (Figure 21.21). With $\theta_R = \theta_I$, the amplitudes $\tilde{\mathbf{E}}_0$ and $\tilde{\mathbf{B}}_0$ of the three waves, in the form (21.9), are

$$\begin{aligned}
&\text{incident:} && \tilde{E}_{0I}(\cos\theta_I\,\hat{\mathbf{x}} - \sin\theta_I\,\hat{\mathbf{z}}), && \frac{1}{v_1}\tilde{E}_{0I}\,\hat{\mathbf{y}},\\
&\text{reflected:} && \tilde{E}_{0R}(\cos\theta_I\,\hat{\mathbf{x}} + \sin\theta_I\,\hat{\mathbf{z}}), && -\frac{1}{v_1}\tilde{E}_{0R}\,\hat{\mathbf{y}},\\
&\text{transmitted:} && \tilde{E}_{0T}(\cos\theta_T\,\hat{\mathbf{x}} - \sin\theta_T\,\hat{\mathbf{z}}), && \frac{1}{v_2}\tilde{E}_{0T}\,\hat{\mathbf{y}}.
\end{aligned}$$

The direction of the reflected electric field is a choice: if the field really points the other way, $\tilde{E}_{0R}$ comes out negative. This choice makes it parallel to the incident field at normal incidence, so that the results agree there with those of Section 21.6.

At $z = 0$ the three exponentials are equal, by the laws of optics, so they cancel. The parallel components are the x component of **E** and the y component of **B**, and the conditions on them are

$$(\tilde{E}_{0I} + \tilde{E}_{0R})\cos\theta_I = \tilde{E}_{0T}\cos\theta_T, \qquad \frac{1}{\mu_1 v_1}(\tilde{E}_{0I} - \tilde{E}_{0R}) = \frac{1}{\mu_2 v_2}\tilde{E}_{0T}.$$

With β of (21.15) and

$$\alpha \equiv \frac{\cos\theta_T}{\cos\theta_I}, \tag{21.20}$$

they are $\tilde{E}_{0I} + \tilde{E}_{0R} = \alpha\tilde{E}_{0T}$ and $\tilde{E}_{0I} - \tilde{E}_{0R} = \beta\tilde{E}_{0T}$. Adding them gives $\tilde{E}_{0T}$, and then either one gives $\tilde{E}_{0R}$:

$$\boxed{r_p = \frac{\alpha - \beta}{\alpha + \beta}, \qquad t_p = \frac{2}{\alpha + \beta}}. \tag{21.21}$$

These are the **Fresnel equations** for p-polarization. The conditions on $\varepsilon E^{\perp}$ and $B^{\perp}$ add nothing new (Exercise P21.14). At normal incidence $\alpha = 1$, and (21.21) becomes (21.16). Sources differ in the direction they choose for the reflected field

in p-polarization. With the opposite choice r_p has the opposite sign, and it is $-(1-\beta)/(1+\beta)$ at normal incidence, so check which choice a source makes before you compare signs.

For the powers, notice that the beams cross the boundary at angles. A patch of area A on the boundary receives the incident power $I_I A\cos\theta_I$, since the part of the incident beam that falls on it has the cross section $A\cos\theta_I$, and the transmitted beam carries the power $I_T A\cos\theta_T$ away from it. (The reflected beam leaves at the incident angle.) So

$$R_p = |r_p|^2 = \left(\frac{\alpha-\beta}{\alpha+\beta}\right)^2, \qquad T_p = \frac{Z_1}{Z_2}\frac{\cos\theta_T}{\cos\theta_I}|t_p|^2 = \frac{4\alpha\beta}{(\alpha+\beta)^2}, \tag{21.22}$$

and again $R_p + T_p = 1$: the power that reaches each patch of the boundary leaves it.

Figure 21.22 shows these for light going from air into glass. At normal incidence $R_p = 0.04$, as in Example 21.3, and r_p is negative: the reflected wave is inverted. As θ_I approaches 90°, which is called grazing incidence, $\cos\theta_I \to 0$ and $\alpha \to \infty$, so $r_p \to 1$ and $t_p \to 0$: nearly everything is reflected. That is why a lake or a wet road reflects so well when you look along it toward the horizon. Between the two is an angle at which $\alpha = \beta$ and $r_p = 0$, so that p-polarized light is not reflected at all. This is **Brewster's angle**, and for materials that are not magnetic

$$\tan\theta_B = \frac{n_2}{n_1} \tag{21.23}$$

(Exercise P21.16). For glass, $\theta_B = 56.3°$. Below Brewster's angle r_p is negative, and above it r_p is positive.

Unpolarized light that strikes a surface at Brewster's angle reflects only its s part, so the reflected light is completely s-polarized, and light reflected near Brewster's angle is mostly s-polarized. For light reflecting off of a horizontal surface such as water, s-polarized light has its electric field horizontal, parallel to the surface. Polarizing sunglasses block horizontally polarized light, which is why they cut the glare from water and roads.

The Fresnel equations for s-polarization come from the same steps, and Tutorial 21.1 works them out. A wave with any other polarization is a sum of an s wave and a p wave, each reflected and transmitted with its own coefficients. For unpolarized light the reflectance is the average of R_s and R_p (Section 21.3).

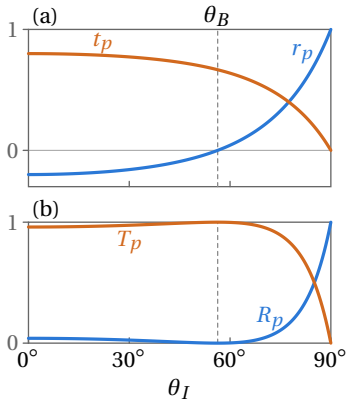


Figure 21.22 The Fresnel equations for p-polarization, (21.21) and (21.22), for light going from air into glass with $n = 1.50$: (a) r_p and t_p; (b) R_p and T_p, against the angle of incidence. At Brewster's angle θ_B (dotted) no p-polarized light is reflected.

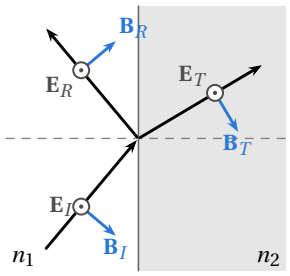


Figure 21.23 The setup of Tutorial 21.1: s-polarized light, with the electric fields of all three waves out of the page and the magnetic fields (blue) in the plane of incidence.

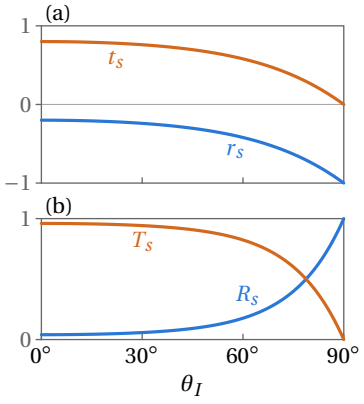


Figure 21.24 The Fresnel equations for s-polarization, for light going from air into glass with $n = 1.50$: (a) r_s and t_s; (b) R_s and T_s, against the angle of incidence. Compare Figure 21.22.

Tutorial 21.1 Fresnel equations for s-polarization

Printable version

Background

Section 21.7.2 derived the Fresnel equations for p-polarized light, whose electric field lies in the plane of incidence. In this tutorial you derive them for s-polarized light, whose electric field is perpendicular to the plane of incidence, by the same steps. Use the axes of that section: the boundary is the plane $z = 0$, the plane of incidence is the xz plane, and $\hat{\mathbf{y}}$ points out of the page. The incident wave travels in material 1 at the angle θ_I from the normal, the reflected wave at $\theta_R = \theta_I$, and the transmitted wave at the angle θ_T given by Snell's law (21.19). Use α and β as defined in (21.20) and (21.15).

Part 1 — The fields

Figure 21.23 shows the electric field of each wave pointing out of the page, along $+\hat{\mathbf{y}}$. If one of them really points into the page, its amplitude will come out negative.

(a) Why do the magnetic fields of the incident and transmitted waves point down, toward $-x$, while the magnetic field of the reflected wave points up?

(b) Write the incident, reflected, and transmitted electric and magnetic fields in complex exponential form. For each wave give the wave vector in terms of its angle, the direction of each field as its own unit vector, and the amplitude of $\mathbf{B}$ in terms of the amplitude of $\mathbf{E}$.

Part 2 — The amplitudes

(a) Impose the boundary conditions on $\mathbf{E}^{\parallel}$ and on $\mathbf{B}^{\parallel}/\mu$ at $z = 0$ to get two equations that involve the three electric field amplitudes. What happens to the exponential factors? Why does the condition on $\varepsilon E^{\perp}$ give nothing here?

(b) Solve the two equations for the ratios of the amplitudes, $r_s = \tilde{E}_{0R}/\tilde{E}_{0I}$ and $t_s = \tilde{E}_{0T}/\tilde{E}_{0I}$, in terms of α and β.

(c) Find the reflectance R_s and the transmittance T_s. (Remember the factor that the cross sections of the beams contributed for p-polarization.) Show that $R_s + T_s = 1$.

Part 3 — Wrap-up

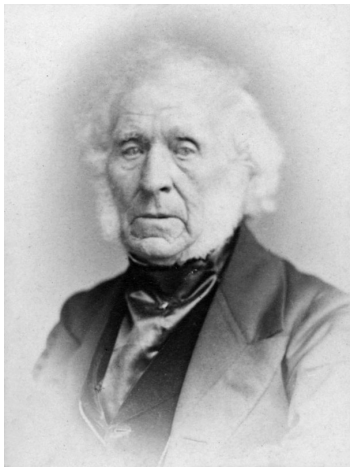
Figure 21.24 shows r_s, t_s, R_s, and T_s for light going from air into glass with $n = 1.50$.

(a) Check your r_s and t_s at normal incidence against (21.16), and show that their grazing incidence behavior matches Figure 21.24.

(b) For light going from air into glass with $n = 1.50$ at 45°, find R_s, and compare it with R_p from (21.22). Which polarization does a glass surface reflect more strongly?

21.8 Summary

In a uniform linear material with no free charges or currents, Maxwell's equations give waves that travel at $v = c/n$. Sinusoidal plane waves are the building blocks. At the boundary between two materials, the boundary conditions must hold at every instant and at every point of the boundary, and that fixes both the directions and the amplitudes of the reflected and transmitted waves.



David Brewster (1781–1868, Scottish) Brewster entered the University of Edinburgh at twelve to train for the ministry, and he was licensed to preach in the Church of Scotland, but he made his career in science. From 1808 to 1830 he edited the *Edinburgh Encyclopaedia*. Étienne-Louis Malus had found in 1808 that reflection polarizes light; in 1815 Brewster showed that the angle at which reflected light is polarized depends on the index of refraction (21.23), and the Royal Society gave him its Copley Medal that year. He invented the kaleidoscope, patented in 1817, and pressed for lenses in Britain's lighthouses. He was knighted in 1831, and he was principal of the University of St Andrews and then of the University of Edinburgh. (Wikipedia)

Waves in matter and sinusoidal waves

$\nabla^2\mathbf{E} = \mu\varepsilon\dfrac{\partial^2\mathbf{E}}{\partial t^2}$, and the same for $\mathbf{B}$ — every component obeys the wave equation

$v = \dfrac{1}{\sqrt{\mu\varepsilon}} = \dfrac{c}{n}, \quad n = \sqrt{\mu_r\varepsilon_r} \approx \sqrt{\varepsilon_r}$ — index of refraction; ε and μ at the wave's frequency

$\tilde{f} = \tilde{A}e^{i(kz-\omega t)}, \quad \tilde{A} = Ae^{i\phi}$ — amplitude, wave number, angular frequency, phase constant; the time-delay convention

$\lambda = 2\pi/k, \quad T = 2\pi/\omega, \quad \nu = 1/T$ — wavelength, period, frequency

$\dfrac{\omega}{k} = \dfrac{c}{n}$ — phase velocity; in a material $\lambda = \lambda_0/n$

$A\cos(kz-\omega t) + A\cos(-kz-\omega t) = 2A\cos kz\cos\omega t$ — standing wave; nodes half a wavelength apart

A wave on a string meets a knot

$r = \dfrac{v_2 - v_1}{v_2 + v_1}, \quad t = \dfrac{2v_2}{v_2 + v_1}$ — displacement and slope continuous; $r < 0$ (inverted) when $v_2 < v_1$

$R = r^2, \quad T = \dfrac{v_1}{v_2}t^2, \quad R + T = 1$ — power $\propto A^2/v$; amplitudes are not conserved, energy is

Polarization

$\tilde{\mathbf{f}} = \left(\tilde{A}_x\hat{\mathbf{x}} + \tilde{A}_y\hat{\mathbf{y}}\right)e^{i(kz-\omega t)}$ — a transverse wave has two independent polarizations

$\tilde{A}_y/\tilde{A}_x$ real — linear polarization

$\tilde{A}_y = -i\tilde{A}_x$ — right-handed circular polarization: clockwise as seen by someone the wave is coming toward (engineers call it left-handed)

$\tilde{A}_y = +i\tilde{A}_x$ — left-handed: counterclockwise

unpolarized light — powers are averages over two perpendicular polarizations

Plane waves, energy, and momentum

$\tilde{\mathbf{E}} = \tilde{\mathbf{E}}_0\, e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}, \quad |\mathbf{k}| = \omega/v$ — wave fronts are planes perpendicular to $\mathbf{k}$

$\mathbf{k}\cdot\tilde{\mathbf{E}}_0 = 0, \quad \tilde{\mathbf{B}}_0 = \dfrac{1}{\omega}\mathbf{k}\times\tilde{\mathbf{E}}_0$ — transverse; $\mathbf{E}$ and $\mathbf{B}$ in phase, $B = E/v$

$Z = \sqrt{\mu/\varepsilon} = \mu v,$
$Z_0 = \mu_0 c = 376.73\,\Omega$ — wave impedance, E/H

$\mathbf{S} = v\,u\hat{\mathbf{k}}, \quad I = \frac{1}{2}\varepsilon v E_0^2 = \dfrac{E_0^2}{2Z}$ — the energy travels at v, half of it in each field; for a given E_0, n times more power in a material

$\mathbf{g} = \mathbf{S}/c^2, \quad P_{\text{rad}} = I/c$ or $2I/c$ — momentum density in vacuum; radiation pressure on an absorber or a mirror

Normal incidence

$\beta = \dfrac{\mu_1 v_1}{\mu_2 v_2} = \dfrac{\mu_1 n_2}{\mu_2 n_1} = \dfrac{Z_1}{Z_2}$ — $\beta = n_2/n_1$ when $\mu_1 = \mu_2$

$r = \dfrac{1-\beta}{1+\beta}, \quad t = \dfrac{2}{1+\beta}$ — $\mathbf{E}^{\parallel}$ and $\mathbf{B}^{\parallel}/\mu$ continuous; the frequency is the same on both sides

$R = r^2, \quad T = \beta t^2, \quad R + T = 1$ — inverted reflection when $n_2 > n_1$; 4% from each air–glass surface

$r = -1$ — perfect conductor: a standing wave, with $\mathbf{E}$ zero at the mirror

Oblique incidence

$\theta_R = \theta_I, \quad n_1 \sin\theta_I = n_2 \sin\theta_T$ — from equal phases along the boundary; all three waves in the plane of incidence

$\sin\theta_c = n_2/n_1$ — critical angle, for $n_1 > n_2$

$r_p = \dfrac{\alpha-\beta}{\alpha+\beta}, \quad t_p = \dfrac{2}{\alpha+\beta},$
$\alpha = \dfrac{\cos\theta_T}{\cos\theta_I}$ — Fresnel equations for p-polarization: $\mathbf{E}$ in the plane of incidence

$R_p = r_p^2, \quad T_p = \alpha\beta\, t_p^2$ — the factor α from the cross sections of the beams

$\tan\theta_B = n_2/n_1$ — Brewster's angle: no reflection of p-polarized light

Exercises

Exercises for 21.1 The wave equation in vacuum and in matter

P21.1 A material that is linear, isotropic, and not magnetic has a permittivity $\varepsilon(\mathbf{r})$ that changes from place to place but not with time, and it has no free charges or currents. (Lenses whose index of refraction changes gradually inside the glass are made this way.)

(a) Show that $\nabla\cdot\mathbf{E} = -(1/\varepsilon)\,\mathbf{E}\cdot\nabla\varepsilon$, so that $\nabla\cdot\mathbf{E}$ is not zero where ε changes, even though there is no free charge. What charge makes it nonzero?

(b) Show that $\mathbf{E}$ obeys

$$\nabla^2\mathbf{E} - \mu_0\varepsilon\,\frac{\partial^2\mathbf{E}}{\partial t^2} = -\nabla\left(\frac{1}{\varepsilon}\,\mathbf{E}\cdot\nabla\varepsilon\right).$$

(c) Estimate the size of each term for a wave of wavelength λ in a material whose ε changes by a fraction of itself over a distance L. When can the right side be neglected?

Exercises for 21.2 Sinusoidal waves

P21.2 The power carried by a wave on a string. A string under tension F carries the wave $f(z,t) = A\cos(kz-\omega t)$.

(a) The part of the string to the left of a point pulls on the part to the right with a force whose sideways component is $-F\,\partial f/\partial z$ (for small slopes). Explain the sign, and show that the power delivered to the part on the right is $P = -F\,(\partial f/\partial z)(\partial f/\partial t)$.

(b) Show that $\langle P\rangle = F\omega^2A^2/2v = \frac{1}{2}\mu\omega^2A^2v$, where μ is the string's mass per unit length.

P21.3 Two strings, with wave speeds v_1 ($z<0$) and v_2 ($z>0$) and the same tension F, are joined by a knot of mass m at $z=0$ (Figure 21.25).

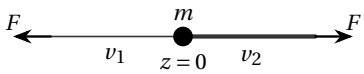


Figure 21.25 The two strings of Exercise P21.3, joined by a knot of mass m.

(a) Explain why the displacement is still continuous at the knot, and show that the second boundary condition becomes $F\left[\partial f/\partial z\right]_{z=0^-}^{z=0^+} = m\,\partial^2 f(0,t)/\partial t^2$.

(b) Find the complex reflection and transmission coefficients $\tilde{r}$ and $\tilde{t}$ for a sinusoidal wave of angular frequency ω coming from the left. They depend on ω.

(c) Show that $|\tilde{r}|^2 + (v_1/v_2)|\tilde{t}|^2 = 1$ for every m and ω.

(d) What happens as $m \to 0$, and as $m \to \infty$? For $v_1 = v_2$, what is the phase of $\tilde{r}$ at high and at low frequencies?

P21.4 Partly standing waves. On the string of Section 21.2.1 (Figure 21.4), the incident and reflected waves to the left of the knot add up.

(a) Show that the amplitude with which the string oscillates at a point $z < 0$ is $|\tilde{A}_I||1 + r\, e^{-2ik_1 z}|$.

(b) Show that this amplitude ranges between $|\tilde{A}_I|(1+|r|)$ and $|\tilde{A}_I|(1-|r|)$, and that neighboring maxima are half a wavelength apart. Is there a maximum or a minimum at the knot when $v_2 > v_1$? When $v_2 < v_1$?

(c) The ratio of the largest amplitude to the smallest, $(1+|r|)/(1-|r|)$, is called the *standing-wave ratio*. Show that a measurement of it determines v_2/v_1, apart from a choice between two values, and explain how the position of a minimum settles the choice.

Exercises for 21.3 Polarization

P21.5 Polarization states. For each of the following complex amplitudes of a wave $\tilde{\mathbf{E}}_0 e^{i(kz-\omega t)}$, write the real field at $z = 0$ as a function of time, sketch the path of its tip as seen by someone the wave is coming toward, and name the polarization (linear, circular, or elliptical, and for the last two, right- or left-handed in the convention of this book):

(a) $\tilde{\mathbf{E}}_0 = E_0\left(\hat{\mathbf{x}} - \hat{\mathbf{y}}\right)$;

(b) $\tilde{\mathbf{E}}_0 = E_0\left(\hat{\mathbf{y}} - i\hat{\mathbf{x}}\right)$;

(c) $\tilde{\mathbf{E}}_0 = E_0\left(2\hat{\mathbf{x}} + i\hat{\mathbf{y}}\right)$;

(d) $\tilde{\mathbf{E}}_0 = E_0\left(\hat{\mathbf{x}} + e^{-i\pi/4}\hat{\mathbf{y}}\right)$. Along which directions do the axes of this ellipse lie, and what is the ratio of their lengths?

P21.6 Linear polarization from circular polarizations.

(a) Show that the sum of a left-handed and a right-handed circularly polarized wave with the same amplitude, $E_0(\hat{\mathbf{x}} + i\hat{\mathbf{y}})e^{i(kz-\omega t)}$ and $E_0(\hat{\mathbf{x}} - i\hat{\mathbf{y}})e^{i(kz-\omega t)}$, is linearly polarized, and find the direction of its polarization.

(b) Show that giving the two waves complex amplitudes $E_0 e^{i\phi}$ and $E_0 e^{-i\phi}$ turns the direction of the linear polarization by the angle ϕ. Which way?

(c) What do you get when the two circular waves have different amplitudes?

Exercises for 21.4 Plane electromagnetic waves

P21.7 A plane wave in vacuum has angular frequency ω, amplitude E_0, and phase constant zero. Write the real fields **E** and **B**, giving the Cartesian components of **k** and of the polarization vector $\hat{\mathbf{n}}$, when the wave

(a) travels in the $-y$ direction and is polarized along $\hat{\mathbf{x}}$;

(b) travels in the direction from the origin toward the point $(1,2,2)$, with its electric field parallel to the xz plane (Figure 21.26). ($\hat{\mathbf{n}}$ must lie in the xz plane and be perpendicular to **k**.)

(c) For the wave of part (b), check that **E**, **B**, and **k** are mutually perpendicular and that $\mathbf{E} \times \mathbf{B}$ points along **k**.

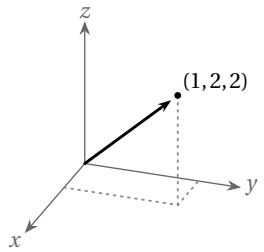


Figure 21.26 The point $(1,2,2)$ of Exercise P21.7 (b), and the direction from the origin toward it.

Exercises for 21.5 Energy and momentum in waves

P21.8 A solar sail. A solar sail, or light sail, is a proposed way of propelling a spacecraft with the radiation pressure of sunlight on a large mirror. Take a perfectly reflecting square sail 800 m on a side, facing the Sun, with a total mass (sail and spacecraft) of 200 kg (Figure 21.27).

(a) Find the force of sunlight on the sail as a function of its distance r from the Sun, and its value at Earth's distance. Compare it with the Sun's gravitational pull on the sail, found as in Example 21.2.

(b) The sail starts from rest at Earth's distance from the Sun and moves straight away from the Sun. (Ignore Earth's gravity: say that rockets have freed the sail from it.) Find its distance from the Sun as a function of time for 180 days by stepping forward in time, as in Section 10.1.5, and plot it. Where in the solar system has it gone by then?

(c) Use conservation of energy to find the speed that the sail approaches far from the Sun, and check your program against it.

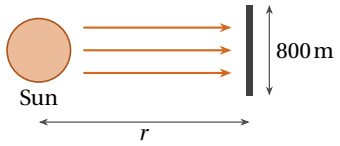


Figure 21.27 The solar sail of Exercise P21.8, facing the Sun (not to scale).

P21.9 Levitation by light. A small, perfectly reflecting mirror of mass 1.0 mg lies flat, and a laser beam aimed straight up from below holds it up against gravity (Figure 21.28).

(a) What laser power is needed? Compare it with the power of familiar electric appliances.

(b) What power would be needed if the mirror were replaced by a black disk of the same mass, and what would happen to the disk?

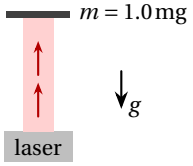


Figure 21.28 The mirror of Exercise P21.9, held up by a laser beam.

P21.10 Radiation pressure at an angle. A plane wave of intensity I in vacuum strikes a flat surface, its direction of travel making the angle θ with the normal to the surface (Figure 21.29).

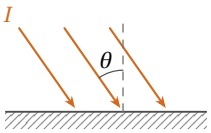


Figure 21.29 The wave and the surface of Exercise P21.10.

(a) The surface absorbs the wave completely. Show that the force per unit area perpendicular to the surface is $(I/c)\cos^2\theta$, and find the force per unit area along the surface. (Why are there two factors of $\cos\theta$?)

(b) The surface is a perfect mirror. Show that the force per unit area is $(2I/c)\cos^2\theta$, perpendicular to the surface.

Exercises for 21.6 Reflection and transmission at normal incidence

P21.11 The standing wave in front of a mirror. A plane wave in vacuum with the real amplitude E_0 reflects from a perfect mirror at $z = 0$, so that in front of the mirror the fields are (21.18) (Figure 21.19).

(a) Find the energy density $u(z,t)$. Show that its average over a period is the same everywhere in front of the mirror, although the averages of its electric and magnetic parts are not, and that the energy in each half wavelength does not change in time.

(b) Find the Poynting vector. Show that it averages to zero at every point, and describe how energy moves during one period.

(c) The fields are zero inside the mirror. Use the boundary condition on $\mathbf{B}^{\parallel}$ (Table 19.1) to find the surface current density $\mathbf{K}$ in the mirror. By the reasoning of Exercise P4.10, the force per unit area on a surface current is $\mathbf{K} \times \frac{1}{2}(\mathbf{B}_{\text{above}} + \mathbf{B}_{\text{below}})$. Show that its average is $2I/c$, the radiation pressure on a mirror of Section 21.5.1.

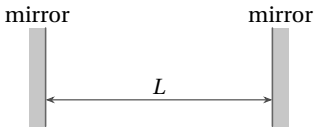


Figure 21.30 The two mirrors of Exercise P21.11 (d).

(d) Two parallel mirrors a distance L apart make the cavity of a laser (Figure 21.30). Which wavelengths and frequencies can form standing waves between them? For $L = 30.0\,\text{cm}$, find the spacing between neighboring frequencies, and the number of half wavelengths between the mirrors for the red light of a helium–neon laser, $\lambda = 632.8\,\text{nm}$.

P21.12 A boundary that is not at $z = 0$. Repeat the calculation of Section 21.6 with the boundary between material 1 and material 2 at the plane $z = d$, keeping the incident wave $\tilde{E}_{0I}\, e^{i(k_1 z - \omega t)}\,\hat{\mathbf{x}}$.

(a) Sketch the boundary and the three waves, with the direction of travel of each and the directions of its electric and magnetic fields. Then write all six fields with their exponentials.

(b) Apply the boundary conditions at $z = d$. Which factors cancel, and which do not?

(c) Show that $\tilde{E}_{0R} = r\,e^{2ik_1 d}\,\tilde{E}_{0I}$ and $\tilde{E}_{0T} = t\,e^{i(k_1-k_2)d}\,\tilde{E}_{0I}$, where $r = (1-\beta)/(1+\beta)$ and $t = 2/(1+\beta)$ are the coefficients (21.16) for a boundary at $z = 0$. Explain the phase factor of the reflected wave by the extra distance it travels, compared with a wave reflected at $z = 0$.

(d) Show that the fractions of the power reflected and transmitted are the same as for a boundary at $z = 0$, and explain why they must be.

P21.13 Light that bounces inside a window. A glass plate in air reflects the fraction R of the power that reaches either of its surfaces (Figure 21.31). Light that enters the plate bounces back and forth between the surfaces, and some of it leaves at each bounce. Ignore interference between these beams, which is reasonable for a thick plate and light that is not from a laser, so that their powers add.

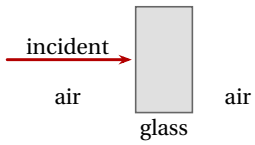


Figure 21.31 The glass plate of Exercise P21.13.

(a) Show that the plate reflects the fraction $2R/(1+R)$ of the incident power and transmits the fraction $(1-R)/(1+R)$. (Sum a geometric series.)

(b) Evaluate these for glass with $n = 1.50$ at normal incidence, and compare them with the fraction $(1-R)^2$ that gets through when the bouncing is ignored.

(c) A stack of N such plates, with air between them, transmits the fraction $(1-R)/[1+(2N-1)R]$. Check this for $N = 1$, and for $N = 2$ by combining two plates as you combined two surfaces. How many plates of glass with $n = 1.50$ transmit less than half the light?

Exercises for 21.7 Oblique incidence: the laws of optics and the Fresnel equations

P21.14 The other two boundary conditions. Section 21.7.2 used only the conditions on $\mathbf{E}^{\parallel}$ and $\mathbf{B}^{\parallel}/\mu$ to find the Fresnel equations for p-polarization.

(a) Show that the condition on $B^{\perp}$ holds automatically.

(b) Show that the condition on $\varepsilon E^{\perp}$ gives $\tilde{E}_{0I} - \tilde{E}_{0R} = \beta\tilde{E}_{0T}$ again, once Snell's law is used. (Use $\varepsilon = n^2/\mu c^2$.)

P21.15 Snell's window. A fish at the depth d below the calm surface of a lake looks up (Figure 21.32).

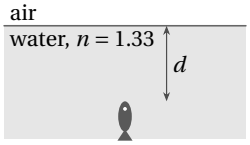


Figure 21.32 The fish of Exercise P21.15.

(a) The light from the whole sky reaches the fish within a cone. Find the half angle of the cone for water with $n = 1.33$. (Think about how the light which reaches the water at close to grazing incidence will bend.)

(b) What is the radius of the circle on the surface through which the fish sees the sky, for $d = 2.0\,\text{m}$? What does the fish see when it looks up at the surface outside this circle?

P21.16 Brewster's angle. Take $\mu_1 = \mu_2$.

(a) By (21.21), p-polarized light is not reflected when $\alpha = \beta$. Use this condition and Snell's law to derive (21.23), $\tan\theta_B = n_2/n_1$. (One way to start: find $\cos^2\theta_T$ in terms of θ_I and β from $\alpha = \beta$, and again from Snell's law, and set the two equal. That gives an equation for $\sin^2\theta_I$, from which you can find $\cos^2\theta_I$ and then $\tan^2\theta_I$.)

(b) Show that at Brewster's angle the reflected and transmitted waves travel at right angles to each other.

(c) By your results from Tutorial 21.1, s-polarized light is not reflected when $\alpha\beta = 1$. Show that this can happen only if $n_1 = n_2$, so that s-polarized light has no Brewster's angle.

(d) Find Brewster's angle for sunlight reflected from a calm lake, $n = 1.33$. In which direction does the electric field of the light reflected at that angle point?

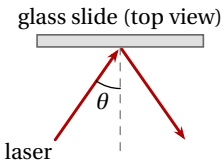


Figure 21.33 The laser beam and the glass slide of Exercise P21.17, seen from above.

P21.17 What a prediction needs. A laser beam travels parallel to an optical table, 5 cm above it, and reflects from a glass microscope slide that stands on the table (Figure 21.33).

(a) Describe what you would need to know to predict the fraction of the power that the front surface of the slide reflects, and the equations you would use.

(b) The light of the laser is polarized vertically, perpendicular to the table. Is it s- or p-polarized at the slide? Find the fraction that the front surface reflects for $n = 1.52$ and $\theta = 45°$, and compare it with the fraction for the other polarization.

(c) The back surface of the slide reflects too. Roughly what fraction of the light does the whole slide reflect? (Exercise P21.13 helps.)

P21.18 Reflection from diamond. Unpolarized light in air strikes a flat face of a diamond, $n = 2.42$. For reflected and transmitted powers, unpolarized light behaves like an equal mixture of s- and p-polarized light (Section 21.3). Use (21.22) and your results from Tutorial 21.1 to plot R and T for unpolarized light against the angle of incidence, from 0 to 90°. What fraction is reflected at normal incidence, and at Brewster's angle? Compare with glass, $n = 1.50$.

Exercises for 21.8 Summary

P21.19 True or false? Explain your reasoning.

(a) In glass with $n = 1.5$, light of a given frequency has a wavelength 1.5 times as long as in vacuum.

(b) When light goes from air into glass, its frequency changes but its speed does not.

(c) A wave on a string that reaches a heavier string is reflected upside down.

(d) A transmission coefficient can be larger than 1.

(e) A transverse wave can be circularly polarized, but a longitudinal wave cannot.

(f) In a plane wave in vacuum, the magnetic field reaches its largest value a quarter of a period after the electric field does.

(g) In this book's convention, right-handed circularly polarized light turns counterclockwise as seen by someone the light is coming toward.

(h) In the same beam of light, a mirror feels twice the radiation pressure that a black surface feels.

(i) At normal incidence, a surface between glass and air reflects a larger fraction of the light that crosses it from the glass than of the light that crosses it from the air.

(j) At Brewster's angle, a surface reflects no light at all.

(k) The law of reflection and Snell's law hold only for electromagnetic waves.

P21.20 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as a pulse on a string that reaches a junction of two strings, the path of the field for any complex amplitudes $\tilde{A}_x$ and $\tilde{A}_y$, the Fresnel coefficients for any two materials and both polarizations, the light reflected by a stack of glass plates, the standing waves in a laser cavity, or the flight of a solar sail.

Chapter 22

Complex Waves

Chapter 21 assumed that waves travel through materials that do not conduct and whose permittivity and permeability are real constants. Then the wave number $k = n\omega/c$ is real, and a wave travels on without losing any of its energy. This chapter drops those assumptions. In a conductor the electric field of a wave drives a current, the current heats the material, and the wave dies away as it travels. In any material whose charges respond to a field with a delay, part of the response is out of phase with the field, and the permittivity depends on the frequency. In all of these cases the wave number becomes complex, $\tilde{k} = k_{\text{real}} + ik_{\text{imag}}$, as Section 18.3 anticipated: its real part sets the wavelength, and its imaginary part sets how quickly the wave decays.

The chapter starts with a case that needs a complex wave number even though nothing absorbs any energy: total internal reflection, in which the transmitted wave cannot travel away from the boundary at all. Then come waves in conductors and their reflection from metals, the complex permittivity and conductivity that describe the response of every material, and the Lorentz model of how electrons respond to a field, which explains why the index of refraction of glass depends on the color of the light and why metals shine. The complex amplitudes and the rule $\partial/\partial t \to -i\omega$ of Section 18.2.4 are used throughout.

22.1 Total internal reflection

When light goes from a material with a larger index of refraction into one with a smaller index, $n_1 > n_2$, as from glass into air, Snell's law (21.19) has no real solution for θ_T once the angle of incidence passes the critical angle θ_c, where $\sin\theta_c = n_2/n_1$ (Section 21.7.1). The light is then reflected completely. This is called **total internal reflection**, and it is how an optical fiber keeps light inside it over many kilometers. But the boundary conditions must still be satisfied, and the incident and reflected waves cannot satisfy them by themselves. Something must be on the other side of the boundary. It is a wave with a complex wave vector.

22.1.1 The evanescent wave

Use the axes of Section 21.7: the boundary is the plane $z = 0$, with material 1 for $z < 0$ and material 2 for $z > 0$, and the plane of incidence is the xz plane. The laws of optics came from the requirement that the exponentials of the three waves match at every point of the boundary (Section 21.7.1), and that requirement has not changed. So the transmitted wave vector still has the same x component as the incident one,

$$k_{Tx} = k_1 \sin\theta_I \equiv k_x,$$

and no y component. In material 2 the transmitted wave must also satisfy the wave equation, which requires $\mathbf{k}_T \cdot \mathbf{k}_T = k_2^2$, or $k_x^2 + k_{Tz}^2 = k_2^2$. With $k = n\omega/c$ (21.5),

$$k_{Tz}^2 = k_2^2 - k_x^2 = \frac{\omega^2}{c^2}\left(n_2^2 - n_1^2 \sin^2\theta_I\right).$$

Below the critical angle the right side is positive, k_{Tz} is real, and the transmitted wave travels away from the boundary at the angle θ_T. Beyond the critical angle $n_1 \sin\theta_I > n_2$, the right side is negative, and k_{Tz} is imaginary: $k_{Tz} = \pm i\kappa$, with

$$\kappa = \frac{\omega}{c}\sqrt{n_1^2 \sin^2\theta_I - n_2^2} \tag{22.1}$$

real and positive. As in Section 18.3, choose the sign by the physics. The transmitted wave's exponential is $e^{i(k_x x + k_{Tz} z - \omega t)} = e^{\mp\kappa z} e^{i(k_x x - \omega t)}$, and the lower sign would make it grow without limit with the distance from the boundary, with nothing in material 2 to feed it. So $k_{Tz} = +i\kappa$, and each component of the transmitted fields is proportional to

$$e^{-\kappa z}\, e^{i(k_x x - \omega t)}. \tag{22.2}$$

With words, this is a wave that travels along the boundary, in the x direction, with the wave number k_x, while its amplitude falls off exponentially with the distance from the boundary (Figure 22.1). It is called an **evanescent wave**. Its amplitude falls by the factor $1/e$ in the distance $1/\kappa$, which is a fraction of a wavelength except close to the critical angle. For light going from glass with $n = 1.50$ into air, $1/\kappa$ is $0.45\lambda_0$ at $\theta_I = 45°$, where $\lambda_0 = 2\pi c/\omega$ is the wavelength in vacuum, and $0.14\lambda_0$ near grazing incidence. At the critical angle $\kappa = 0$, and the transmitted wave travels along the boundary, as Section 21.7.1 found.

Another way to say this is that Snell's law still holds, with an angle that is not real. The transmitted wave vector is $k_2(\sin\theta_T\,\hat{\mathbf{x}} + \cos\theta_T\,\hat{\mathbf{z}})$, as before, with

$$\sin\theta_T = \frac{k_x}{k_2} = \frac{n_1}{n_2}\sin\theta_I > 1, \qquad \cos\theta_T = \frac{k_{Tz}}{k_2} = \frac{i\kappa}{k_2}.$$

These still satisfy $\sin^2\theta_T + \cos^2\theta_T = 1$, since $k_x^2 - \kappa^2 = k_2^2$. The derivation of the Fresnel equations in Section 21.7.2 is algebra with $\sin\theta_T$ and $\cos\theta_T$, and nothing in it requires them to be real: the transmitted fields written there still satisfy

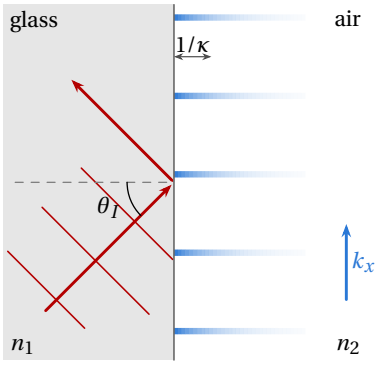


Figure 22.1 Total internal reflection at $\theta_I = 45°$ from glass ($n_1 = 1.50$, gray) into air. The crests of the incident wave (red) meet the boundary at the same places as the crests of the evanescent wave in the air (blue), which travels up along the boundary while it fades within about a wavelength. The shading of the blue crests is proportional to $e^{-\kappa z}$.

Maxwell's equations with these values. So the Fresnel equations still hold beyond the critical angle, with α replaced by the complex number

$$\tilde{\alpha} = \frac{\cos\theta_T}{\cos\theta_I} = \frac{i\kappa}{k_2\cos\theta_I} = i|\tilde{\alpha}|,$$

which is now purely imaginary.

22.1.2 The reflected wave and the energy

Take p-polarization, as in Section 21.7.2. With $\tilde{\alpha} = i|\tilde{\alpha}|$ the reflection coefficient (21.21) becomes the complex number

$$\tilde{r}_p = \frac{i|\tilde{\alpha}| - \beta}{i|\tilde{\alpha}| + \beta}. \tag{22.3}$$

The numerator and the denominator have the same magnitude, $\sqrt{|\tilde{\alpha}|^2 + \beta^2}$, so $|\tilde{r}_p| = 1$, and $R_p = |\tilde{r}_p|^2 = 1$: all of the power is reflected. That is the "total" in total internal reflection (Figure 22.2). The reflected wave is shifted in phase, though. The denominator lies in the first quadrant at the angle $\arctan(|\tilde{\alpha}|/\beta)$, and the numerator is the negative of its complex conjugate, at the angle $180° - \arctan(|\tilde{\alpha}|/\beta)$, so $\tilde{r}_p = e^{i\phi_p}$ with

$$\phi_p = 180° - 2\arctan\frac{|\tilde{\alpha}|}{\beta}.$$

At the critical angle $\tilde{\alpha} = 0$, so $\tilde{r}_p = -1$ and $\phi_p = 180°$. As θ_I approaches 90°, $|\tilde{\alpha}|$ grows without limit and ϕ_p goes to zero. (Like the sign of r_p in Section 21.7.2, the phase depends on the direction chosen for the reflected field.)

The transmission coefficient $\tilde{t}_p = 2/(i|\tilde{\alpha}| + \beta)$ is not zero: the evanescent wave has a definite amplitude, and for glass into air at 45°, $|\tilde{t}_p| = 2.4$. Yet no power is transmitted. The formula (21.22) for T_p does not apply here, since it assumed a transmitted beam that leaves the boundary at a real angle. The energy has to be found from the fields of the evanescent wave themselves. With $\cos\theta_T = i\kappa/k_2$ and $\sin\theta_T = k_x/k_2$, the transmitted fields of Section 21.7.2 become

$$\tilde{\mathbf{E}}_T = \tilde{E}_{0T}\left(\frac{i\kappa}{k_2}\hat{\mathbf{x}} - \frac{k_x}{k_2}\hat{\mathbf{z}}\right)e^{-\kappa z}e^{i(k_x x - \omega t)}, \qquad \tilde{\mathbf{B}}_T = \frac{1}{v_2}\tilde{E}_{0T}\,e^{-\kappa z}e^{i(k_x x - \omega t)}\hat{\mathbf{y}}. \tag{22.4}$$

By (20.15), with μ_2 in place of μ_0, the average Poynting vector has the components

$$\langle S_z\rangle = \frac{1}{2\mu_2}\,\mathrm{Re}\left[\tilde{E}_{Tx}\tilde{B}^*_{Ty}\right] = \frac{1}{2\mu_2}\,\mathrm{Re}\left[\frac{i\kappa}{k_2 v_2}\right]|\tilde{E}_{0T}|^2 e^{-2\kappa z} = 0,$$
$$\langle S_x\rangle = -\frac{1}{2\mu_2}\,\mathrm{Re}\left[\tilde{E}_{Tz}\tilde{B}^*_{Ty}\right] = \frac{k_x}{2\mu_2\omega}|\tilde{E}_{0T}|^2 e^{-2\kappa z}.$$

No energy crosses the boundary on average. The x component of $\mathbf{E}$ is 90° out of phase with $\mathbf{B}$, so energy flows into material 2 during half of each cycle and

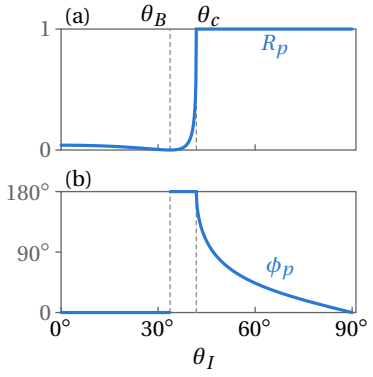


Figure 22.2 Reflection of p-polarized light going from glass ($n_1 = 1.50$) into air: (a) the reflectance R_p and (b) the phase ϕ_p of $\tilde{r}_p$, against the angle of incidence. Brewster's angle $\theta_B = 33.7°$ and the critical angle $\theta_c = 41.8°$ are dotted. Beyond θ_c, $R_p = 1$, and the phase falls from 180° to 0.

back out during the other half. Along the boundary energy does flow, in the $+x$ direction, but only within a distance of order $1/\kappa$ of it.

Total internal reflection is used wherever light must be reflected with no loss. The glass prisms in many binoculars reflect light that strikes their faces at 45°, beyond the critical angle of glass, with no metal coating at all, and an optical fiber guides light along its core because the light strikes the surrounding glass beyond the critical angle (Exercise P22.2). What happens when a second piece of glass is brought within a fraction of a wavelength of the boundary is the subject of Exercise P22.1.

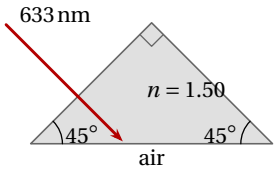
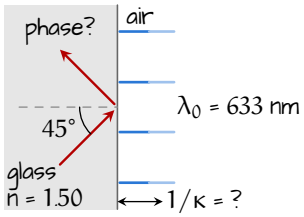


Figure 22.3 The prism of Example 22.1 and the laser beam, which enters perpendicular to one of its short faces.

Example 22.1

A beam from a helium–neon laser, with $\lambda_0 = 633\,\text{nm}$, enters a prism of glass with $n = 1.50$ whose cross section is a right isosceles triangle. It enters perpendicular to one of the short faces, so that it strikes the long face, which borders on air, at 45°. (a) Is it totally reflected? (b) How far beyond the long face does the evanescent wave reach before its amplitude has fallen by the factor $1/e$? Before its intensity has? (c) For p-polarized light, by what angle is the reflected wave shifted in phase?

Solution:



The first step is a sketch of what is given and what is asked: the long face of the prism with glass on one side and air on the other, the beam arriving at 45° from the normal, the reflected beam, whose phase we want, and the evanescent wave in the air, whose reach we want.

(a) The critical angle for glass into air is given by $\sin\theta_c = 1/1.50$, so $\theta_c = 41.8°$. The beam strikes the face at 45°, beyond the critical angle, so it is totally reflected. The prism works as a mirror, with no metal coating.

(b) By (22.1), with $n_1 = 1.50$, $n_2 = 1$, and $\omega/c = 2\pi/\lambda_0$,

$$\kappa = \frac{2\pi}{633\times10^{-9}\,\text{m}}\sqrt{(1.50)^2\sin^2 45° - 1} = \frac{2\pi}{633\times10^{-9}\,\text{m}}(0.354) = 3.51\times10^6\,\text{m}^{-1}.$$

The amplitude falls by $1/e$ in the distance $1/\kappa = 285\,\text{nm}$, less than half of the wavelength in vacuum. The intensity is proportional to the square of the amplitude, $e^{-2\kappa z}$, so it falls by $1/e$ in half that distance, $1/2\kappa = 142\,\text{nm}$.

(c) With $\mu_1 = \mu_2$, $\beta = n_2/n_1 = 0.667$. In air $k_2 = \omega/c$, so $\kappa/k_2 = 0.354$ from (b), and

$$|\tilde{\alpha}| = \frac{\kappa}{k_2\cos\theta_I} = \frac{0.354}{\cos 45°} = 0.500.$$

Then

$$\phi_p = 180° - 2\arctan\frac{0.500}{0.667} = 180° - 73.7° = 106.3°.$$

The reflected wave is shifted in phase by 106°, with the choice of direction for the reflected field made in Section 21.7.2. The intensity of the reflected beam is the same as that of the incident beam.

The results of this section were for p-polarization. Tutorial 22.1 works out s-polarization, continuing Tutorial 21.1.

Tutorial 22.1 Total internal reflection, s-polarized Printable version

Background

Section 22.1 found the reflected wave and the evanescent wave for p-polarized light beyond the critical angle. In this tutorial you do the same for s-polarized light, whose electric field is perpendicular to the plane of incidence. Use the axes and the notation of that section and of Tutorial 21.1: the boundary is the plane $z = 0$, material 1 (index n_1) is at $z < 0$ and material 2 (index $n_2 < n_1$) is at $z > 0$, the plane of incidence is the xz plane, both materials have the same permeability, and $\theta_I > \theta_c$. You will need your coefficients r_s and t_s from Tutorial 21.1. Beyond the critical angle they become the complex numbers $\tilde{r}_s$ and $\tilde{t}_s$.

Part 1 — The reflected wave

(a) Beyond the critical angle $\tilde{\alpha} = i|\tilde{\alpha}|$ is purely imaginary, while β is real. Use this in your r_s to show that $|\tilde{r}_s| = 1$. What fraction of the incident power is reflected?

(b) Find the phase ϕ_s of $\tilde{r}_s$ as a formula, and evaluate it for glass with $n = 1.50$ into air at $\theta_I = 45°$. Compare it with the phase ϕ_p of Example 22.1.

Part 2 — The evanescent wave

(a) Write the transmitted electric and magnetic fields in complex exponential form. Start from your fields of Tutorial 21.1, and replace $\sin\theta_T$ and $\cos\theta_T$ by their values beyond the critical angle, in terms of k_x, κ, and k_2.

(b) Take the real parts. Choose the zero of time so that $\tilde{E}_{0T}$ is real. Which component of **B** is in phase with **E**, and which is 90° out of phase with it?

(c) Check that your fields satisfy all four of Maxwell's equations in material 2.

Part 3 — The energy

(a) Find the average Poynting vector of the evanescent wave. Show that on average no energy crosses the boundary, and find the direction in which energy does flow.

Part 4 — Wrap-up

(a) Compare your results with those for p-polarization in Section 22.1. What is the same, and what differs?

(b) $\tilde{t}_s$ is not zero, yet no power is transmitted. Explain how both can be true.

22.2 Waves in conductors

Chapter 21 assumed that there are no free charges or currents. In a conductor that cannot be the whole story. When a wave arrives there is an electric field inside the conductor, and the free charges move in response to it. Their current takes energy from the wave, which dies away as it travels. This section finds how quickly.

22.2.1 Free charge and the damped wave equation

Take a conductor that is linear, isotropic, and uniform, with permittivity ε, permeability μ, and conductivity σ, so that $\mathbf{D} = \varepsilon\mathbf{E}$, $\mathbf{B} = \mu\mathbf{H}$, and the free current obeys Ohm's law, $\mathbf{J}_f = \sigma\mathbf{E}$ (Section 19.2). Maxwell's equations in matter (19.2) become

$$\nabla\cdot\mathbf{E} = \frac{\rho_f}{\varepsilon}, \qquad \nabla\cdot\mathbf{B} = 0, \qquad \nabla\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t}, \qquad \nabla\times\mathbf{B} = \mu\sigma\mathbf{E} + \mu\varepsilon\frac{\partial\mathbf{E}}{\partial t}.$$

Compared with (21.1), the new things are the free charge density in the first equation and the free current, the term $\mu\sigma\mathbf{E}$, in the last.

The free charge does not last. By the continuity equation (20.5), Ohm's law, and the first equation,

$$\frac{\partial\rho_f}{\partial t} = -\nabla\cdot\mathbf{J}_f = -\sigma\,\nabla\cdot\mathbf{E} = -\frac{\sigma}{\varepsilon}\rho_f,$$

so $\rho_f(t) = \rho_f(0)\,e^{-\sigma t/\varepsilon}$. This predicts that free charge placed inside a conductor should flow away to its surface on a time scale of ε/σ. For copper, with $\varepsilon = \varepsilon_0$ and $\sigma = 5.96\times10^7\,\mathrm{S/m}$ (Section 16.4.1), the time ε_0/σ is only $1.5\times10^{-19}\,\mathrm{s}$. That number should not be taken literally. Ohm's law describes the average drift of the electrons over many collisions with the atoms of the metal, and ε_0/σ is far shorter than the time between collisions; Exercise P22.3 shows what happens instead. The free charge inside the conductor *does* decay, just not in the way nor as fast as predicted here. However, the conclusion still stands: inside a conductor, you can assume $\rho_f = 0$, so that $\nabla\cdot\mathbf{E} = 0$.

☞ **Caution: Ohm's law does not hold on extremely short time scales.** Here the time scale predicted by ε_0/σ is about $10^{-19}\,\mathrm{s}$. Free charge inside a metal does not really die away as quickly as is predicted here; see Exercise P22.3.

Now take the curl of Faraday's law, as in Section 16.5, and use the identity (1.58) with $\nabla\cdot\mathbf{E} = 0$:

$$\nabla\times(\nabla\times\mathbf{E}) = -\nabla^2\mathbf{E} = -\frac{\partial}{\partial t}(\nabla\times\mathbf{B}) = -\mu\sigma\frac{\partial\mathbf{E}}{\partial t} - \mu\varepsilon\frac{\partial^2\mathbf{E}}{\partial t^2}.$$

So

$$\nabla^2\mathbf{E} = \mu\varepsilon\frac{\partial^2\mathbf{E}}{\partial t^2} + \mu\sigma\frac{\partial\mathbf{E}}{\partial t}, \tag{22.5}$$

and **B** obeys the same equation (take the curl of the last of Maxwell's equations instead). This is the **damped wave equation**: the wave equation (21.2) with one new term, a first derivative with respect to time. That term plays the part of the damping term $\Gamma\, dx/dt$ of the oscillator in Example 18.2: it takes energy out of the wave.

22.2.2 The complex wave number and the skin depth

Look for a wave of one frequency that travels along z, $\tilde{\mathbf{E}} = \tilde{\mathbf{E}}_0\, e^{i(\tilde{k}z-\omega t)}$. In (22.5) the Laplacian becomes $-\tilde{k}^2$ and $\partial/\partial t$ becomes $-i\omega$, so $-\tilde{k}^2 = -\mu\varepsilon\omega^2 - i\mu\sigma\omega$, or

$$\tilde{k}^2 = \mu\varepsilon\omega^2 + i\mu\sigma\omega. \tag{22.6}$$

Since $\tilde{k}^2$ is complex, so is $\tilde{k} = k_{\text{real}} + i\,k_{\text{imag}}$, and the wave is

$$\tilde{\mathbf{E}} = \tilde{\mathbf{E}}_0\, e^{-k_{\text{imag}}z}\, e^{i(k_{\text{real}}z-\omega t)},$$

as in Section 18.3: k_{real} sets the wavelength, $2\pi/k_{\text{real}}$, and the speed of the wave fronts, ω/k_{real}, while k_{imag} sets how quickly the amplitude decays.

To find $\tilde{k}$, call the right side of (22.6) K for short. Squaring $\tilde{k} = k_{\text{real}} + i\,k_{\text{imag}}$ gives $k_{\text{real}}^2 - k_{\text{imag}}^2 = \operatorname{Re} K$ and $2k_{\text{real}}k_{\text{imag}} = \operatorname{Im} K$. Then $(k_{\text{real}}^2 + k_{\text{imag}}^2)^2 = (k_{\text{real}}^2 - k_{\text{imag}}^2)^2 + 4k_{\text{real}}^2k_{\text{imag}}^2 = |K|^2$, so $k_{\text{real}}^2 + k_{\text{imag}}^2 = |K|$, and adding and subtracting the first equation gives

$$k_{\text{real}} = \sqrt{\frac{|K| + \operatorname{Re} K}{2}}, \qquad k_{\text{imag}} = \sqrt{\frac{|K| - \operatorname{Re} K}{2}}. \tag{22.7}$$

Of the two square roots of K, this is the one with $k_{\text{imag}} \geq 0$, a wave that decays as it travels toward $+z$, and it is right whenever $\operatorname{Im} K \geq 0$, so that k_{real} and k_{imag} have the same sign. Here $|K| = \mu\varepsilon\omega^2\sqrt{1 + (\sigma/\varepsilon\omega)^2}$, and

$$k_{\text{real}} = \omega\sqrt{\frac{\mu\varepsilon}{2}}\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}+1\right]^{1/2}, \qquad k_{\text{imag}} = \omega\sqrt{\frac{\mu\varepsilon}{2}}\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}-1\right]^{1/2}. \tag{22.8}$$

Everything depends on the ratio $\sigma/\varepsilon\omega$, which compares the free current with the displacement current: their amplitudes are $\sigma|\tilde{\mathbf{E}}|$ and $\varepsilon\omega|\tilde{\mathbf{E}}|$. When $\sigma/\varepsilon\omega \ll 1$ the material is a **poor conductor**, nearly an insulator, and $k_{\text{real}} \approx \omega\sqrt{\mu\varepsilon}$, as in Chapter 21, while k_{imag} is small (Exercise P22.4). When $\sigma/\varepsilon\omega \gg 1$ the material is a **good conductor**. Then $\tilde{k}^2 \approx i\mu\sigma\omega = \mu\sigma\omega\, e^{i\pi/2}$, so

$$\tilde{k} = \sqrt{\mu\sigma\omega}\, e^{i\pi/4} = \sqrt{\frac{\mu\sigma\omega}{2}}\,(1+i),$$

and $k_{\text{real}} = k_{\text{imag}}$. The distance in which the amplitude falls by the factor $1/e$, $\delta \equiv 1/k_{\text{imag}}$, is called the **skin depth**, and for a good conductor

$$\boxed{\delta = \sqrt{\frac{2}{\mu\sigma\omega}}}. \tag{22.9}$$

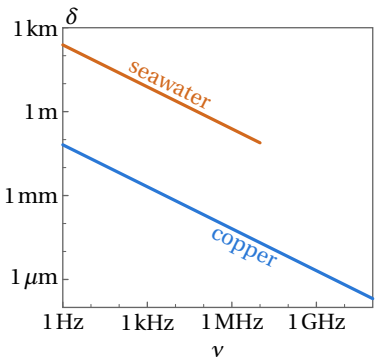


Figure 22.4 The skin depth (22.9) of copper and of seawater against the frequency, on logarithmic scales. Both lines fall as $1/\sqrt{\nu}$. Seawater is shown up to 10 MHz, where it is still a good conductor.

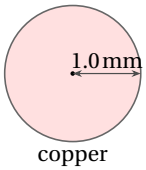


Figure 22.5 A cross section of the copper wire of Example 22.2.

The wavelength inside the conductor is $2\pi/k_{\text{real}} = 2\pi\delta$, so in one of its own wavelengths the amplitude falls by the factor $e^{-2\pi} = 0.0019$: the wave is gone almost before it has started. Figure 22.4 shows the skin depth of two conductors over a wide range of frequencies.

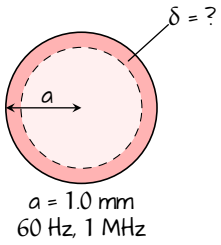
For copper, $\sigma/\varepsilon_0\omega$ is $(1.1 \times 10^{18}\,\text{Hz})/\nu$, where $\nu = \omega/2\pi$, so copper is a good conductor at every frequency up to the infrared, where its conductivity itself begins to change (Section 22.4). Seawater conducts far less well, with $\sigma = 4.29\,\text{S/m}$ at 15°C,[1] and its skin depth is 31 m at 60 Hz and 2.4 m at 10 kHz. This is why radio contact with a submerged submarine is possible only at very low frequencies.

When a current in a wire alternates, the same thing happens inside the wire: the fields, and with them the current, are confined to a layer about one skin depth thick at the surface. This is the skin effect, which Example 15.5 described in words; Example 22.2 estimates its size.

Example 22.2

A straight copper wire has a radius of 1.0 mm. (a) Find the skin depth of copper at 60 Hz, the frequency of the electric power grid in North America, and at 1.0 MHz, in the AM radio band. (b) Estimate the ratio of the wire's resistance at each of these frequencies to its resistance for a steady current.

Solution:



The first step is a sketch of what is given and what is asked: a cross section of the wire, radius 1.0 mm, with the current crowded into a layer about δ thick at its surface, where δ is still to be found.

(a) Copper is a good conductor at both frequencies, so with $\sigma = 5.96 \times 10^7\,\text{S/m}$ and $\mu = \mu_0$, (22.9) gives, at 60 Hz,

$$\delta = \sqrt{\frac{2}{(4\pi \times 10^{-7}\,\text{H/m})(5.96 \times 10^7\,\text{S/m})(2\pi)(60\,\text{s}^{-1})}} = 8.4 \times 10^{-3}\,\text{m} = 8.4\,\text{mm}.$$

The skin depth is proportional to $1/\sqrt{\omega}$, so at 1.0 MHz, which is 1.67×10^4 times higher, it is smaller by the factor $\sqrt{1.67 \times 10^4} = 129$:

$$\delta = \frac{8.4\,\text{mm}}{129} = 6.5 \times 10^{-5}\,\text{m} = 65\,\mu\text{m}.$$

[1] This is the conductivity of standard seawater, of practical salinity 35, at 15°C and atmospheric pressure: *Gibbs SeaWater Oceanographic Toolbox*, "gsw_C3515," https://teos-10.org/pubs/gsw/html/gsw_C3515.html. It is larger in warmer or saltier water.

(b) At 60 Hz the skin depth is more than eight times the radius, so the current fills the wire nearly uniformly, and the resistance is very nearly its value for a steady current. At 1.0 MHz the skin depth is much less than the radius. The surface of the wire then looks flat over distances of a few skin depths, and the current density falls off with the depth below the surface as for the wave in this section. The current behaves as if it flowed uniformly in a tube one skin depth thick at the surface (Exercise P22.6 shows that this is exact when δ is much less than the radius). For a wire of length L and radius a, the resistance of that tube is $L/(\sigma 2\pi a\delta)$, and the resistance for a steady current is $L/(\sigma\pi a^2)$, so their ratio is

$$\frac{R_{AC}}{R_{DC}} = \frac{\pi a^2}{2\pi a\delta} = \frac{a}{2\delta} = \frac{1.0\times10^{-3}\,\text{m}}{2(6.5\times10^{-5}\,\text{m})} = 7.7.$$

At 1 MHz the wire has nearly eight times the resistance it has for a steady current. That is why wires for high frequencies are often made of many fine, insulated strands, or of a thin layer of copper or silver on a cheaper core: the inside of a thick wire carries almost no current anyway.

22.2.3 The fields in a conductor

The fields of a wave in a conductor are still transverse, and still perpendicular to each other, for the same reasons as in Section 21.4. With $\nabla \to i\tilde{k}\hat{\mathbf{z}}$, the two divergence equations give $\hat{\mathbf{z}}\cdot\tilde{\mathbf{E}}_0 = 0$ and $\hat{\mathbf{z}}\cdot\tilde{\mathbf{B}}_0 = 0$, and Faraday's law gives

$$\tilde{\mathbf{B}}_0 = \frac{\tilde{k}}{\omega}\hat{\mathbf{z}}\times\tilde{\mathbf{E}}_0, \tag{22.10}$$

as in (21.10). But now $\tilde{k}$ is complex, and that changes two things. First, the ratio of the sizes of the fields is $|\tilde{\mathbf{B}}_0|/|\tilde{\mathbf{E}}_0| = |\tilde{k}|/\omega$, which is not $1/v$ for the speed $v = \omega/k_{\text{real}}$ of the wave fronts. Second, the fields are no longer in phase. Write $\tilde{k} = |\tilde{k}|\,e^{i\phi}$, with $\tan\phi = k_{\text{imag}}/k_{\text{real}}$. Then $\tilde{\mathbf{B}}_0$ has the phase of $\tilde{\mathbf{E}}_0$ plus ϕ: the magnetic field lags behind the electric field by the angle ϕ. In a good conductor $\phi = 45°$ (Figure 22.6), and $|\tilde{\mathbf{B}}_0|/|\tilde{\mathbf{E}}_0| = \sqrt{\mu\sigma/\omega}$, far larger than in an insulator; for copper at 1 MHz, $c|\tilde{\mathbf{B}}_0|/|\tilde{\mathbf{E}}_0| = 1.0\times10^6$.

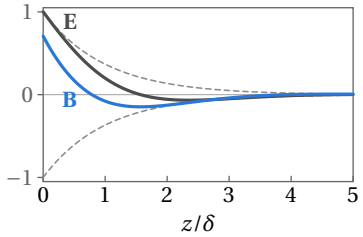


Figure 22.6 The fields of a wave in a good conductor at one instant, against the depth z below the surface, in units of the skin depth δ: **E** (black) and **B** (blue), each in units of its amplitude at the surface, and the envelope $\pm e^{-z/\delta}$ (dashed). **B** lags behind **E** by 45°, so its pattern trails **E**'s by an eighth of a wavelength.

So most of the energy of the wave is in its magnetic field. The average energy densities are $\langle u_E\rangle = \frac{1}{4}\varepsilon|\tilde{\mathbf{E}}|^2$ and $\langle u_B\rangle = \frac{1}{4}|\tilde{\mathbf{B}}|^2/\mu$ (Section 21.5, with (18.16)), so

$$\frac{\langle u_B\rangle}{\langle u_E\rangle} = \frac{|\tilde{k}|^2}{\mu\varepsilon\omega^2} = \sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2},$$

which is about 10^{12} for copper at 1 MHz.

By (20.15), with μ in place of μ_0, the average Poynting vector of the wave is

$$\langle\mathbf{S}\rangle = \frac{1}{2\mu}\,\text{Re}\left[\tilde{\mathbf{E}}\times\tilde{\mathbf{B}}^*\right] = \frac{k_{\text{real}}}{2\mu\omega}|\tilde{\mathbf{E}}_0|^2\,e^{-2k_{\text{imag}}z}\hat{\mathbf{z}}.$$

The intensity falls off twice as fast as the amplitude, as $e^{-2k_{\text{imag}}z}$, and $2k_{\text{imag}}$ is called the **absorption coefficient**. The power that the wave loses goes into heating the conductor: by Poynting's theorem in matter (20.16), the power lost from the fields per unit volume is the rate $\mathbf{E}\cdot\mathbf{J}_f$ at which they do work on the free charges, and Exercise P22.6 shows that $-d\langle S_z\rangle/dz$ is exactly the average of σE^2.

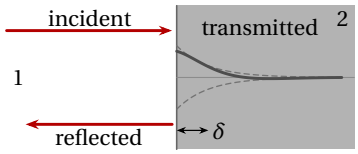


Figure 22.7 A wave in an insulator, material 1, strikes the surface of a conductor, material 2 (gray), at normal incidence. Inside the conductor the transmitted wave dies away within a few skin depths δ. The solid curve is its electric field at one instant, and the dashed curves are the envelope $\pm|\tilde{E}_{0T}|e^{-z/\delta}$.

22.3 Reflection from a conductor

Why do metals reflect so well? Take a plane wave in an insulator, material 1, that strikes the flat surface of a conductor, material 2, at normal incidence (Figure 22.7). The derivation of Section 21.6 goes through step by step, except that the transmitted wave has the complex wave number $\tilde{k}_2$, and by (22.10) its magnetic field is $(\tilde{k}_2/\omega)\tilde{E}_T$ instead of $\tilde{E}_T/v_2$. So the coefficients (21.16) still hold,

$$\tilde{r} = \frac{1-\tilde{\beta}}{1+\tilde{\beta}}, \qquad \tilde{t} = \frac{2}{1+\tilde{\beta}},$$

with β replaced by the complex number

$$\tilde{\beta} = \frac{\mu_1 v_1 \tilde{k}_2}{\mu_2 \omega}. \tag{22.11}$$

Both coefficients are now complex, so the reflected and transmitted waves are shifted in phase by angles other than 0 and 180°.

It is often convenient to describe a material by its **complex index of refraction**, defined from $\tilde{k}$ just as n was defined from k in (21.5):

$$\tilde{n} \equiv \frac{c\tilde{k}}{\omega} = n + i\kappa. \tag{22.12}$$

☞ **Caution: notation for $\tilde{k}$ and $\tilde{n}$.** This book writes $\tilde{k} = k_{\text{real}} + ik_{\text{imag}}$ and $\tilde{n} = n + i\kappa$, with n and κ both positive in a material that absorbs. Other sources write k and κ for k_{real} and k_{imag}; $\mathcal{N}$ (a calligraphic N) for $\tilde{n}$; and k (as most tables of optical constants do) or κ for the imaginary part of $\tilde{n}$. So k and κ can each mean two different things. Sources that assume an $e^{i\omega t}$ time dependence (Section 18.5) often write $\tilde{n} = n - ik$, with the same positive k.

Written as $\omega/\tilde{k} = c/\tilde{n}$, this is my favorite equation, (21.5), once again, now with a complex wave number and a complex index of refraction. The imaginary part κ of $\tilde{n}$ is called the **extinction coefficient**.[2] With $v_1 = c/n_1$, $\tilde{\beta} = \mu_1\tilde{n}_2/\mu_2 n_1$, and for materials that are not magnetic

$$\tilde{r} = \frac{n_1 - \tilde{n}_2}{n_1 + \tilde{n}_2}, \qquad \tilde{t} = \frac{2n_1}{n_1 + \tilde{n}_2}. \tag{22.13}$$

The reflected wave travels in material 1, so the reflectance is $R = |\tilde{r}|^2$, as before. Only the size of $\tilde{r}$ matters, since the intensity of the reflected wave does not depend on its phase. The transmitted wave's intensity just inside the surface is, by (20.15) with $\tilde{\mathbf{B}}_T = (\tilde{k}_2/\omega)\hat{\mathbf{z}}\times\tilde{\mathbf{E}}_T$, $\frac{1}{2}\text{Re}(\tilde{k}_2)|\tilde{E}_{0T}|^2/\mu_2\omega$, and the incident wave's is $\frac{1}{2}|\tilde{E}_{0I}|^2/\mu_1 v_1$. Their ratio is the transmittance, so

$$R = |\tilde{r}|^2, \qquad T = \text{Re}(\tilde{\beta})\,|\tilde{t}|^2. \tag{22.14}$$

[2] This κ is not the decay constant κ of Section 22.1.1.

When $\tilde{\beta}$ is real these are (21.17). Exercise P22.7 shows that $R+T=1$ for any complex $\tilde{\beta}$: the power that is not reflected enters the conductor, and there all of it is absorbed, within a few skin depths of the surface.

In the limit of a perfect conductor, $\sigma \to \infty$, both $\tilde{k}_2$ and $\tilde{\beta}$ grow without limit, so $\tilde{r}=-1$ and $\tilde{t}=0$: the wave is reflected completely, and inverted, as by the mirror of Section 21.6.1. Ordinary mirrors are made this way. A thin layer of silver or aluminum on the glass does the reflecting, and the glass is there only to hold it.

A good conductor is close to that limit. In air, with $n_1=1$, $\mu=\mu_0$ and $\sigma \gg \varepsilon_0\omega$, $\tilde{k}_2^2 \approx i\mu_0\sigma\omega$, so $\tilde{n}_2^2 = c^2\tilde{k}_2^2/\omega^2 \approx i\sigma/\varepsilon_0\omega$, and $\tilde{n}_2 \approx (1+i)g$ with $g=\sqrt{\sigma/2\varepsilon_0\omega} \gg 1$. Then

$$R=\frac{|1-\tilde{n}_2|^2}{|1+\tilde{n}_2|^2}=\frac{(g-1)^2+g^2}{(g+1)^2+g^2}\approx 1-\frac{2}{g}=1-2\sqrt{\frac{2\varepsilon_0\omega}{\sigma}}, \tag{22.15}$$

to first order in the small quantity $1/g$.[3] For copper at 10 GHz this is $R=1-2.7\times 10^{-4}$: copper reflects all but 0.03% of the power of these microwaves, which is why the walls of microwave waveguides and cavities are made of copper or plated with silver (Chapter 23). The formula does not apply to visible light, at whose frequencies the conductivity of a metal is no longer its value for steady currents; Section 22.5 explains why metals reflect visible light well too.

Example 22.3

Light traveling in air strikes, at normal incidence, a material whose complex index of refraction is $\tilde{n}=1+2i$. (a) What fraction of its power is reflected, and by what angle is the reflected wave shifted in phase? (b) What if the index of refraction were purely imaginary, $\tilde{n}=2i$?

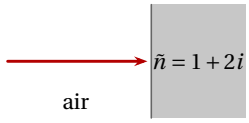
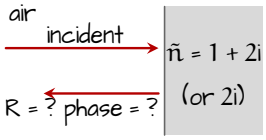


Figure 22.8 The light and the material of Example 22.3.

Solution:



The first step is a sketch of what is given and what is asked: the incident wave in air, the material with its complex index of refraction, and the reflected wave, whose power and phase we want.

(a) By (22.13) with $n_1=1$,

$$\tilde{r}=\frac{1-(1+2i)}{1+(1+2i)}=\frac{-2i}{2+2i}.$$

In polar form the numerator is 2 at the angle $-90°$, and the denominator is $2\sqrt{2}$ at $45°$, so

$$\tilde{r}=\frac{2\,e^{-i90°}}{2\sqrt{2}\,e^{i45°}}=\frac{1}{\sqrt{2}}\,e^{-i135°}.$$

The reflectance is $R=|\tilde{r}|^2=1/2$: half of the power is reflected, and the reflected wave lags the incident wave by 135°. The other half of the power enters the material, where it is absorbed.

[3] This is called the Hagen–Rubens relation.

(b) Now

$$\tilde{r} = \frac{1-2i}{1+2i}.$$

The numerator is the complex conjugate of the denominator, so they have the same magnitude, and $R = |\tilde{r}|^2 = 1$: all of the power is reflected. Nothing here depends on the 2. For any purely imaginary $\tilde{n}$ the numerator is the conjugate of the denominator, and $R = 1$. A material whose index of refraction is purely imaginary reflects everything, although it is not a perfect conductor. Section 22.5 shows that a metal's index of refraction is nearly imaginary at visible frequencies.

22.4 Complex permittivity and complex conductivity

Chapter 21 described insulators by a real permittivity, and Sections 22.2 and 22.3 described conductors by a real conductivity, as if every material responded to a field at once and in the same way at every frequency. Real materials are more complicated in two ways. First, their charges have mass, so they cannot follow a field that changes quickly: their response lags behind the field, by an amount that depends on the frequency. Second, the line between conductors and insulators is not sharp. Every real material is partly a conductor and partly an insulator. Even the best insulator absorbs a little: the extremely pure glass of the best optical fibers still loses about 3% of the power of the light that passes through it in each kilometer, most of it to scattering but some to absorption (Exercise P22.10), and an insulator carries a small current when a steady voltage is applied to it. And the electrons of a metal, which carry a steady current freely, also polarize the metal when the field oscillates quickly (Section 22.5.3). The split between conductors and insulators that began in Chapters 4 and 9 is really a statement about which part of a material's response is larger at the frequency in question. This section describes both parts with complex numbers.

! **Partly both.** Every real material is partly a conductor and partly an insulator. Which description fits depends on the frequency.

22.4.1 Polarization that lags the field

In Chapter 9 the polarization of a linear dielectric was proportional to the field, $\mathbf{P} = \varepsilon_0 \chi_e \mathbf{E}$. That cannot be exactly true when the field changes with time. The driven oscillator of Example 18.2 shows what happens instead: driven at the frequency ω, it oscillates at that frequency, but its displacement lags behind the force by a phase that depends on ω, and so does its amplitude. The charges in a material behave the same way. For fields that oscillate at one frequency, with complex amplitudes $\tilde{\mathbf{E}}$ and $\tilde{\mathbf{P}}$, the lag is described by a complex susceptibility $\tilde{\chi}_e$ and a **complex permittivity**:

☞ **Caution: notation for $\tilde{\varepsilon}_r$.** This book writes $\tilde{\varepsilon}_r = \varepsilon_r' + i\varepsilon_r''$. Many sources write ε for $\tilde{\varepsilon}_r$; ε_r, ε_1, or ε' for its real part; and ε_2 or ε'' for its imaginary part. Others write K or $\tilde{K}$, K_r or K', and K_i or K''.

$$\tilde{\mathbf{P}} = \varepsilon_0 \tilde{\chi}_e \tilde{\mathbf{E}}, \qquad \tilde{\mathbf{D}} = \varepsilon_0 \tilde{\varepsilon}_r \tilde{\mathbf{E}}, \qquad \tilde{\varepsilon}_r = 1 + \tilde{\chi}_e = \varepsilon_r' + i\varepsilon_r''. \tag{22.16}$$

Both $\tilde{\chi}_e$ and $\tilde{\varepsilon}_r$ depend on ω, and the phase of $\tilde{\chi}_e$ is the angle by which the polarization lags behind the field. These relations hold only between complex

amplitudes at one frequency. For a field that is not sinusoidal they make no sense as they stand, because the polarization at one instant depends on the field at earlier instants. So wherever a complex $\tilde{\varepsilon}_r$ appears in this chapter, the time derivatives have already been replaced by $-i\omega$ (18.14).

The imaginary part of $\tilde{\varepsilon}_r$ is what absorbs energy. The field does work on the bound charges through the polarization current $\mathbf{J}_p = \partial\mathbf{P}/\partial t$ (19.1), at the rate $\mathbf{E}\cdot\mathbf{J}_p$ per unit volume (20.10). With $\partial/\partial t \to -i\omega$ and the time average (18.16),

$$\left\langle \mathbf{E}\cdot\frac{\partial\mathbf{P}}{\partial t}\right\rangle = \frac{1}{2}\mathrm{Re}\left[(-i\omega\tilde{\mathbf{P}})\cdot\tilde{\mathbf{E}}^*\right] = \frac{1}{2}\omega\varepsilon_0\varepsilon_r''|\tilde{\mathbf{E}}|^2. \tag{22.17}$$

The part of the polarization in phase with the field, which ε_r' describes, takes energy from the field during part of each cycle and gives it back during the rest. The part 90° behind the field, which ε_r'' describes, takes energy on average, and the material turns it into heat. A material left to itself cannot give energy to a wave, so $\varepsilon_r'' \geq 0$: $\tilde{\varepsilon}_r$ lies in the upper half of the complex plane, although its real part can have either sign.

In a material that does not conduct and is not magnetic, the derivation of the wave equation in Section 21.1 goes through at one frequency with $\varepsilon_0\tilde{\varepsilon}_r$ in the place of ε, so $\tilde{k}^2 = \tilde{\varepsilon}_r\omega^2/c^2$, and with the complex index of refraction (22.12),

$$\tilde{n}^2 = \tilde{\varepsilon}_r. \tag{22.18}$$

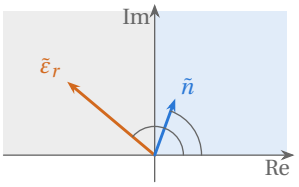


Figure 22.9 The complex plane. Every material's $\tilde{\varepsilon}_r$ lies in the upper half (shaded); $\tilde{n} = \sqrt{\tilde{\varepsilon}_r}$, at half the angle, lies in the first quadrant (blue), even when $\varepsilon_r' < 0$, as here.

Squaring $\tilde{n} = n + i\kappa$ gives $\varepsilon_r' = n^2 - \kappa^2$ and $\varepsilon_r'' = 2n\kappa$, and the roots (22.7), with K replaced by $\tilde{\varepsilon}_r$, give n and κ from $\tilde{\varepsilon}_r$. Since $\varepsilon_r'' \geq 0$, n and κ have the same sign, and with the root that decays both are positive: $\tilde{n}$ and $\tilde{k}$ lie in the first quadrant, even where ε_r' is negative (Figure 22.9). Where ε_r' is negative and ε_r'' is small, $\tilde{n}$ is nearly imaginary, and by Example 22.3 the material reflects nearly all of the light that strikes it.

The average Poynting vector is still $\langle\mathbf{S}\rangle = \frac{1}{2}\mathrm{Re}[\tilde{\mathbf{E}}\times\tilde{\mathbf{H}}^*]$ (20.15), which needs only the fields' complex amplitudes. What does change is the energy stored in the material: it is no longer $\frac{1}{2}\varepsilon E^2$, and finding it requires knowing how $\tilde{\varepsilon}_r$ depends on ω, a subject beyond this book.

22.4.2 Conduction at high frequencies

Ohm's law fails in the same way when the field changes quickly. In Drude's scattering model of Section 19.3, the average velocity $\mathbf{v}$ of the conduction electrons obeys (19.13),

$$m\frac{d\mathbf{v}}{dt} = -e\mathbf{E} - \frac{m}{\tau}\mathbf{v},$$

where the last term stands for collisions with the atoms of the material, which remove an electron's extra velocity in the time τ on average. In a steady field $d\mathbf{v}/dt = 0$, and the current density $\mathbf{J}_f = -Ne\mathbf{v}$, with N conduction electrons per unit volume, is Ohm's law, $\mathbf{J}_f = \sigma_0\mathbf{E}$, with the DC conductivity $\sigma_0 = Ne^2\tau/m$ of

(19.15). At one frequency $d/dt \to -i\omega$, so $(-i\omega + 1/\tau)\tilde{\mathbf{v}} = -(e/m)\tilde{\mathbf{E}}$, and the free current is $\tilde{\mathbf{J}}_f = \tilde{\sigma}\tilde{\mathbf{E}}$ with the **complex conductivity**

$$\tilde{\sigma} = \frac{\sigma_0}{1 - i\omega\tau}. \tag{22.19}$$

For a good metal at room temperature τ is of the order of 10^{-14} s (Exercise P22.3 finds it for copper), so $\omega\tau \ll 1$ and $\tilde{\sigma} \approx \sigma_0$ at all frequencies up to the infrared. The results of Sections 22.2 and 22.3 for radio waves and microwaves stand. At visible frequencies, though, $\omega\tau \gg 1$, and $\tilde{\sigma} \approx i\sigma_0/\omega\tau$ is nearly imaginary. The electrons' velocity then lags behind the field by nearly 90°, and their current, like the current in an inductor, mostly takes energy from the field and gives it back each cycle instead of absorbing it.

22.4.3 One current, two descriptions

Should the response of a metal's electrons be described as a current, with $\tilde{\sigma}$, or as a polarization, with $\tilde{\varepsilon}_r$? At any frequency other than zero, either will do. If the conduction electrons move by $\tilde{\mathbf{x}}$ from their average positions, their current density is $\tilde{\mathbf{J}}_f = -Ne(-i\omega\tilde{\mathbf{x}}) = -i\omega(-Ne\tilde{\mathbf{x}})$, and $-Ne\tilde{\mathbf{x}}$ is a dipole moment per unit volume, a polarization. So the current of the free electrons is the rate of change of their own polarization, just as $\mathbf{J}_p$ is for the bound charges. At one frequency the Ampère–Maxwell law in matter (19.2) is

$$\nabla \times \tilde{\mathbf{H}} = \tilde{\mathbf{J}}_f - i\omega\tilde{\mathbf{D}} = (\tilde{\sigma} - i\omega\varepsilon_0\tilde{\varepsilon}_r)\tilde{\mathbf{E}}.$$

☞ *Caution*: engineers, with $e^{j\omega t}$ (Section 18.5), write $\tilde{\varepsilon} = \varepsilon' - j\varepsilon''$ and $\tilde{n} = n - j\kappa$, with this book's κ and with $\varepsilon'' = \varepsilon_0\varepsilon_r''$. They often include the conductivity in ε'', as in (22.20).

All of the response can be put into an effective permittivity, $\nabla \times \tilde{\mathbf{H}} = -i\omega\varepsilon_0\tilde{\varepsilon}_{r,\text{eff}}\tilde{\mathbf{E}}$, with

$$\tilde{\varepsilon}_{r,\text{eff}} = \tilde{\varepsilon}_r + \frac{i\tilde{\sigma}}{\varepsilon_0\omega}, \tag{22.20}$$

or all of it into a total conductivity, $\nabla \times \tilde{\mathbf{H}} = \tilde{\sigma}_{\text{tot}}\tilde{\mathbf{E}} - i\omega\varepsilon_0\tilde{\mathbf{E}}$, with $\tilde{\sigma}_{\text{tot}} = \tilde{\sigma} - i\omega\varepsilon_0(\tilde{\varepsilon}_r - 1)$, which leaves out only the displacement current of empty space. With $\tilde{\varepsilon}_{r,\text{eff}}$ in the place of $\tilde{\varepsilon}_r$, every result for insulators holds for conductors too: $\tilde{k}^2 = \tilde{\varepsilon}_{r,\text{eff}}\omega^2/c^2$ and $\tilde{n}^2 = \tilde{\varepsilon}_{r,\text{eff}}$. For a real σ and a real $\varepsilon = \varepsilon_0\varepsilon_r$ this is (22.6) again. Tables of the optical constants of metals list $\tilde{n}$, and so $\tilde{\varepsilon}_{r,\text{eff}}$, without separating the parts, and Section 22.5 describes metals the same way.

Which part of the current is called conduction and which polarization is therefore bookkeeping, as long as $\omega \neq 0$. Only at $\omega = 0$ is there a difference that matters: a conductor carries a steady current, while the charges of an insulator can only be displaced a finite distance. In $\tilde{\varepsilon}_{r,\text{eff}}$ that difference appears as the term $i\sigma_0/\varepsilon_0\omega$, which grows without limit as $\omega \to 0$. Exercise P22.9 uses the two descriptions together to find the collision time of the electrons in sodium from its optical constants.

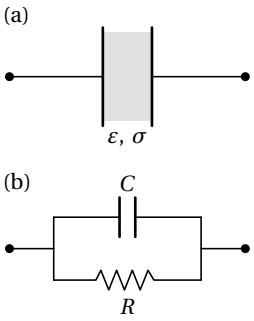


Figure 22.10 (a) A capacitor filled with a material that conducts a little. (b) The same current flows in an ideal capacitor in parallel with a resistor.

A capacitor filled with a material that conducts a little shows both descriptions in one circuit (Figure 22.10). Between plates of area A a distance d apart, the current is the conduction current plus the displacement current, $I = (\sigma A/d)V +$

$(\varepsilon A/d)\,dV/dt$: the same current as in an ideal capacitor, $C = \varepsilon A/d$, in parallel with a resistor, $R = d/\sigma A$. At one frequency $\tilde{I} = -i\omega\tilde{C}\tilde{V}$, with the complex capacitance

$$\tilde{C} = \frac{A}{d}\left(\varepsilon + \frac{i\sigma}{\omega}\right) = \frac{\varepsilon_0 A}{d}\,\tilde{\varepsilon}_{r,\text{eff}}.$$

Its real part stores energy, as the capacitor does, and its imaginary part dissipates it, as the resistor does. Engineers describe real capacitors this way, by the ratio $\varepsilon_r''/\varepsilon_r'$ of their dielectric, called its **loss tangent**.

22.5 The Lorentz oscillator model

Why does the permittivity of a material depend on frequency the way it does? The simplest model that answers the question, the **Lorentz oscillator model**, treats each charge in the material as a particle on a spring, driven by the field of the wave. It is named for Hendrik Lorentz, who from 1892 described matter as made of small charged particles that respond to electric and magnetic fields (his portrait is in Section 10.1.2). With the spring removed, the same model describes the free electrons of a metal.

22.5.1 Bound electrons

Take an electron, with charge q and mass m, held near its place in an atom by a force $-m\omega_0^2 x$ proportional to its displacement x, slowed by a damping force $-m\Gamma\,dx/dt$, and pushed by the field of a wave, qE. (The damping stands for every way in which the electron loses the energy of its oscillation; the model does not say what they are.) Its equation of motion is that of the driven oscillator of Example 18.2, with the force qE, so at one frequency its displacement is

$$\tilde{x} = \frac{q}{m}\,\frac{\tilde{E}}{\omega_0^2 - \omega^2 - i\omega\Gamma}. \qquad \text{(with } e^{-i\omega t}\text{, Section 18.5)}$$

The displaced electron and its atom form a dipole, $\tilde{p} = q\tilde{x}$. With N such electrons per unit volume the polarization is $\tilde{P} = Nq\tilde{x}$, and comparing with (22.16) gives $\tilde{\chi}_e$, so

$$\tilde{\varepsilon}_r = 1 + \frac{\omega_p^2}{\omega_0^2 - \omega^2 - i\omega\Gamma}, \qquad \omega_p^2 = \frac{Nq^2}{m\varepsilon_0}. \tag{22.21}$$

The constant ω_p has the units of a frequency and is called the **plasma frequency**, for a reason that Section 22.5.3 explains. At $\omega = 0$ the electron's displacement is in phase with the field, $p = (q^2/m\omega_0^2)E$, so the atomic polarizability of Section 9.1.5 is $\alpha = q^2/m\omega_0^2$: a stiff spring gives a small polarizability. The model takes the field acting on each electron to be the average field in the material. In a dense material the field that acts on one atom differs from the average, as Tutorial 9.1 showed; including the difference shifts the resonant frequency but keeps the form of (22.21), so in practice its parameters are fitted to measurements.

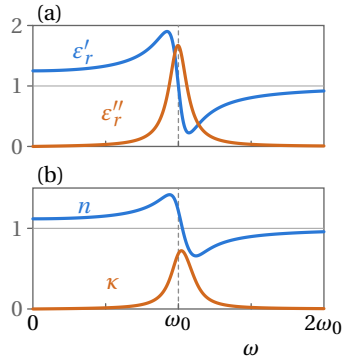


Figure 22.11 The Lorentz model (22.21) near a resonance, with $\omega_p = 0.5\,\omega_0$ and $\Gamma = 0.15\,\omega_0$: (a) ε_r' and ε_r'', (b) n and κ.

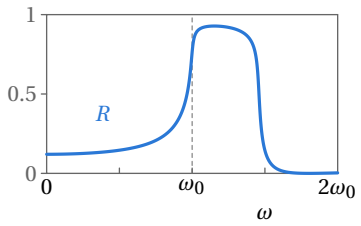


Figure 22.12 The reflectance at normal incidence from air of a material with one strong resonance, $\tilde{\varepsilon}_r = \varepsilon_\infty + \omega_p^2/(\omega_0^2 - \omega^2 - i\omega\Gamma)$, with $\varepsilon_\infty = 2$, $\omega_p = 1.5\omega_0$, and $\Gamma = 0.02\omega_0$. In the band just above ω_0, where ε_r' is negative, most of the light is reflected.

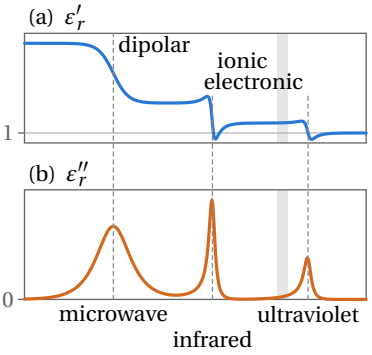


Figure 22.13 A schematic of $\tilde{\varepsilon}_r$ of a material with polar molecules, ions, and electrons, against the frequency on a logarithmic scale: (a) ε_r' falls in a step at each kind of response; (b) ε_r'' has a peak at each.

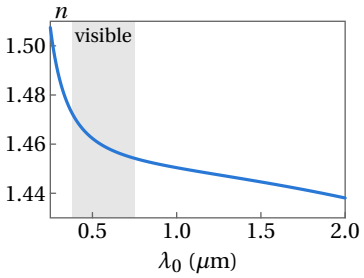


Figure 22.14 The index of refraction of fused silica against the vacuum wavelength, from the formula of Malitson (footnote). The visible range is shaded.

Figure 22.11 shows $\tilde{\varepsilon}_r$ and $\tilde{n}$ near a resonance. Far below ω_0 the electron follows the field in phase, and $\tilde{\varepsilon}_r$ is real and larger than 1. Far above, it cannot keep up: it moves 180° out of phase with the push, with a small amplitude, and $\tilde{\varepsilon}_r$ is real and slightly less than 1. Near ω_0 the electron lags by about 90°, and ε_r'' has a peak whose width is about Γ: the material absorbs in a band of frequencies around the resonance, an absorption line. By (22.21), $\varepsilon_r'' = \omega_p^2\omega\Gamma/[(\omega_0^2 - \omega^2)^2 + \omega^2\Gamma^2]$, which is positive at every frequency, as it must be. When the resonance is strong enough, ε_r' becomes negative in a band just above ω_0, and there the material reflects strongly (Figure 22.12 and Exercise P22.12).

A real material has several kinds of charges that respond, each with its own resonant frequency ω_j, damping rate Γ_j, and share f_j of the electrons, called its **oscillator strength**. Their polarizations add:

$$\tilde{\varepsilon}_r = \varepsilon_\infty + \omega_p^2 \sum_j \frac{f_j}{\omega_j^2 - \omega^2 - i\omega\Gamma_j}. \tag{22.22}$$

Here ω_p is computed with all N of the electrons that are included, and ε_∞, which would be 1 if every responding charge were included, stands for resonances at frequencies far above those of interest, whose charges still polarize the material a little. Each resonance adds its own absorption line and its own step in ε_r' (Figure 22.13). The electrons of atoms resonate in the ultraviolet. The ions of a crystal such as salt follow the same equation, with their own charge and mass; they are thousands of times heavier than electrons, and they resonate in the infrared. Polar molecules respond differently, by turning against the jostling of their neighbors. In water they follow the field only up to microwave frequencies, and water absorbs microwaves strongly, which is how a microwave oven heats food. Every material absorbs at some frequencies: a material that is transparent to visible light is one whose absorption lines all lie outside the visible range.

22.5.2 Dispersion

Below a resonance, n increases with frequency, as the drive approaches the resonance. The dependence of the index of refraction on frequency is called **dispersion**, and an index that increases with frequency is called **normal dispersion**. Glass is transparent in the visible because its nearest absorption lines are in the ultraviolet, and below them its index increases toward the blue (Figure 22.14). Fused silica, a glass of pure silicon dioxide, has $n = 1.455$ at 700 nm (red) and 1.470 at 400 nm (violet).[4] That is why a prism bends violet light more than red light, and spreads white light into a spectrum. Inside an absorption line n decreases with frequency, which is called **anomalous dispersion**; there the material absorbs strongly, so little light gets through. Dispersion also means that waves of different frequencies travel at different speeds, so a pulse made of many of them changes shape as it travels. How fast the pulse itself travels is taken up in Section 23.4.

[4] From the formula fitted to measurements at 20°C in "Interspecimen comparison of the refractive index of fused silica," I. H. Malitson, J. Opt. Soc. Am. 55, 1205–1209 (1965). https://doi.org/10.1364/JOSA.55.001205

Far below all of its resonances, where every $\omega_j \gg \omega$ and the damping does not matter, the index of a transparent material can be expanded in powers of ω. For a gas, whose ε_r is close to 1, $\tilde{n} = \sqrt{\tilde{\varepsilon}_r} \approx 1 + \frac{1}{2}\tilde{\chi}_e$; with $(1-\omega^2/\omega_j^2)^{-1} = 1+\omega^2/\omega_j^2+\omega^4/\omega_j^4+\cdots$ in each term of (22.22) and $\omega = 2\pi c/\lambda_0$, the index becomes

$$n \approx A + \frac{B}{\lambda_0^2} + \frac{C}{\lambda_0^4}, \tag{22.23}$$

with constants A, B, and C that depend on the material. This is **Cauchy's formula**, which Augustin-Louis Cauchy proposed in the 1830s, decades before anyone knew of electrons. Often the first two terms are enough, and the same form, with different constants, describes glasses as well (Exercise P22.11 fits it to air).

22.5.3 Free electrons and the plasma frequency

The conduction electrons of a metal are not bound to any atom, so there is no spring: $\omega_0 = 0$. Then (22.21) becomes

$$\tilde{\varepsilon}_r = 1 - \frac{\omega_p^2}{\omega^2 + i\omega\Gamma}, \tag{22.24}$$

with N the number of conduction electrons per unit volume and $q = -e$. This is the Drude model of Section 19.3, extended to oscillating fields in Section 22.4.2, in another form. With $\Gamma = 1/\tau$ and $\sigma_0 = Ne^2\tau/m = \varepsilon_0\omega_p^2\tau$, (22.24) is exactly $1 + i\tilde{\sigma}/\varepsilon_0\omega$ with $\tilde{\sigma}$ from (22.19): the effective permittivity (22.20) of a metal whose only response is that of its free electrons.

The plasma frequency has a simple meaning. Shift all of the conduction electrons of a slab of metal by a small distance x perpendicular to its faces, relative to the positive ions (Figure 22.15). That leaves the surface charge densities $-Nex$ on one face and $+Nex$ on the other, which produce the field Nex/ε_0 inside, pointing so as to pull the electrons back. Each electron then obeys $m\,d^2x/dt^2 = -(Ne^2/\varepsilon_0)x$, so the electrons oscillate back and forth at the frequency ω_p. A gas of free charges, called a plasma, oscillates the same way, and that gives ω_p its name. For aluminum, with three conduction electrons per atom, $N = 1.81 \times 10^{29}\,\mathrm{m}^{-3}$ and $\omega_p = 2.40 \times 10^{16}\,\mathrm{s}^{-1}$, which corresponds to the vacuum wavelength $2\pi c/\omega_p = 78.5\,\mathrm{nm}$, far in the ultraviolet. For most metals ω_p is in the ultraviolet. At frequencies far above Γ, which for a metal includes visible light, the damping hardly matters, and $\tilde{\varepsilon}_r \approx 1 - \omega_p^2/\omega^2$ is nearly real. The electrons then move 180° out of phase with the force on them, so their polarization points opposite to the field, and ε_r is less than 1. Below ω_p it is negative: $\tilde{n}$ is nearly imaginary, and by Example 22.3 nearly all of the light is reflected. That is why metals shine. Above ω_p, ε_r is positive but less than 1, $\tilde{n}$ is nearly real and less than 1, and the metal becomes transparent. In 1933 R. W. Wood found that thin films of the alkali metals, opaque to visible light, are transparent in the ultraviolet; sodium, for example, transmits light of wavelengths shorter than 210 nm.[5] Real metals also have bound

[5] "Remarkable optical properties of the alkali metals," R. W. Wood, Phys. Rev. 44, 353–360 (1933). https://doi.org/10.1103/PhysRev.44.353



Augustin-Louis Cauchy (1789–1857, French) Cauchy's *Cours d'analyse* of 1821, written for students of the École Polytechnique, was the first rigorous study of when infinite series converge, and he laid the foundations of the theory of functions of a complex variable and of the theory of elasticity. He left Paris after the revolution of July 1830 and lived in Turin and Prague until 1838. His theory of the dispersion of light, in which light is a vibration of an elastic medium, appeared in the 1830s. (Wikipedia)

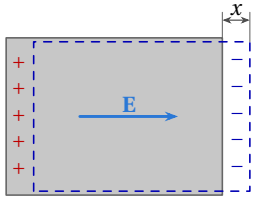


Figure 22.15 The conduction electrons of a slab of metal, shifted to the right relative to the ions, leave surface charge on the faces. Its field pulls them back, and they oscillate at ω_p.

electrons, whose resonances add terms like those of (22.22) to (22.24). In copper and gold these absorb part of the visible spectrum, which is why those metals have their colors.

So a metal behaves differently in three ranges of frequency. Far below Γ it is the good conductor of Section 22.2; between Γ and ω_p it reflects nearly everything; above ω_p it is transparent. Tutorial 22.2 works through all three for sodium. The free electrons of the upper atmosphere make the same distinction for radio waves.

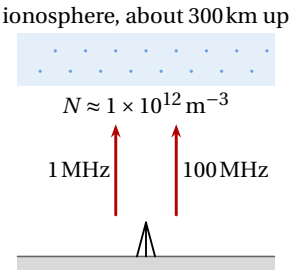
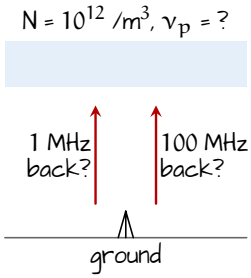


Figure 22.16 The ionosphere of Example 22.4, far above a radio transmitter on the ground that sends waves straight up.

Example 22.4

Sunlight ionizes part of the upper atmosphere. At its densest, about 300 km above the ground, this layer, the ionosphere, typically has about 1×10^{12} free electrons per cubic meter in the daytime.[6] (a) Find the plasma frequency $\nu_p = \omega_p/2\pi$ of these electrons. (b) A transmitter on the ground sends radio waves straight up. Which of them are turned back toward the ground: AM radio waves at 1 MHz, or FM radio waves at 100 MHz? (c) Find the index of refraction of the densest part of the ionosphere at 100 MHz.

Solution:



The first step is a sketch of what is given and what is asked: the ground, the layer of free electrons far above it, and the two radio waves going up, whose fate we want.

(a) By (22.21) with $q = -e$,

$$\omega_p = \sqrt{\frac{Ne^2}{m\varepsilon_0}} = \sqrt{\frac{(1\times10^{12}\,\mathrm{m}^{-3})(1.602\times10^{-19}\,\mathrm{C})^2}{(9.109\times10^{-31}\,\mathrm{kg})(8.854\times10^{-12}\,\mathrm{F/m})}} = 5.64\times10^{7}\,\mathrm{s}^{-1},$$

so $\nu_p = \omega_p/2\pi = 9.0\,\mathrm{MHz}$.

(b) The air at that height is so thin that the electrons rarely collide: Γ is far smaller than ω for both waves, so by (22.24)

$$\varepsilon_r \approx 1 - \frac{\omega_p^2}{\omega^2} = 1 - \frac{\nu_p^2}{\nu^2}.$$

[6] The density depends on the latitude, the season, and the sun's eleven-year cycle of activity. At middle latitudes the daytime peak is about $6 \times 10^{11}\,\mathrm{m}^{-3}$ near the minimum of the cycle and about $5 \times 10^{12}\,\mathrm{m}^{-3}$ near its maximum, and at night it is roughly ten times smaller: Figure 9-1 of F. J. Rich and Su. Basu, "Ionospheric Physics," Chapter 9 of *Handbook of Geophysics and the Space Environment*, Adolph S. Jursa, ed. (Air Force Geophysics Laboratory, 1985).

At 1 MHz, $\varepsilon_r = 1 - 81 = -80$, so $\tilde{n} = 8.9i$ is imaginary, and the wave is reflected. The density of electrons rises gradually with height, so the wave is not reflected at a sharp surface: it turns back at the height where the plasma frequency of the electrons there equals its own frequency, below the densest part of the layer. At 100 MHz, ε_r is positive at every height, and the wave goes on into space.

(c) At 100 MHz,

$$n = \sqrt{1 - \left(\frac{9.0\,\text{MHz}}{100\,\text{MHz}}\right)^2} = \sqrt{0.992} = 0.996.$$

The ionosphere hardly affects these waves. Instruments called ionosondes measure the density of the ionosphere this way: they send pulses straight up at a range of frequencies and find the highest frequency that comes back. Satellites communicate with the ground at frequencies far above ν_p for the same reason that FM signals escape.

The Drude model also describes a metal, whose electrons are about 10^{16} times denser. Tutorial 22.2 follows one metal through the three ranges of frequency; Figure 22.17 shows how suddenly a metal's reflectance falls at ω_p.

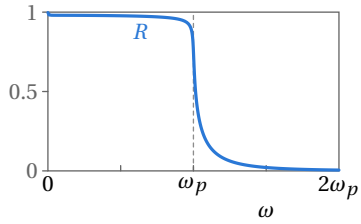


Figure 22.17 The reflectance at normal incidence from air of a free-electron metal, (22.24) with $\Gamma = 0.01\,\omega_p$. Below ω_p nearly all of the light is reflected; above it, little is.

Tutorial 22.2 A metal across the spectrum

Printable version

Background

In this tutorial you use the Drude model (22.24) to follow one metal, sodium, from microwaves to the ultraviolet. Sodium has one conduction electron per atom, 2.5×10^{28} per cubic meter, and its resistivity at 20°C is $4.77 \times 10^{-8}\,\Omega\text{m}$ (Table 19.2). Keep enough digits in each complex number to tell which of its parts is small. Where $\tilde{\varepsilon}_r$ is nearly real and negative, $n = \varepsilon_r''/2\kappa$ is more accurate on a calculator than the roots (22.7).

Part 1 — The parameters

(a) Find the plasma frequency ω_p of sodium and the vacuum wavelength $2\pi c/\omega_p$. In what part of the spectrum is it?

(b) Use the connection between (22.24) and the conductivity of Section 22.4.2 to find the damping rate Γ from the resistivity. How many times smaller than ω_p is it?

Part 2 — Microwaves

(a) For microwaves at 10 GHz, compare ω with Γ, and simplify (22.24) accordingly. Find $\tilde{\varepsilon}_r$. Which of its parts is larger, and by how much?

(b) Find $\tilde{n}$, and the distance in which the amplitude of the wave falls by the factor $1/e$. Compare that distance with the skin depth (22.9) computed with the DC conductivity.

(c) Find the reflectance at normal incidence from air.

Part 3 — Visible light

(a) For light with $\lambda_0 = 500\,\text{nm}$, compare ω with Γ and with ω_p, simplify (22.24), and find $\tilde{\varepsilon}_r$.

(b) Find $\tilde{n}$, the reflectance, and the distance in which the amplitude falls by $1/e$.

Part 4 — Ultraviolet

(a) Repeat Part 3 for $\lambda_0 = 150\,\text{nm}$. Is n larger or smaller than 1?

(b) How far does this light travel in sodium before its intensity falls by $1/e$? Could a film of sodium $1\,\mu\text{m}$ thick transmit much of it? Could it transmit the light of Part 3?

Part 5 — Wrap-up

(a) Using your results, explain in a few sentences why sodium looks like a shiny metal.

(b) At what wavelength does the model predict that sodium becomes transparent? Compare it with Wood's measurement (Section 22.5.3).



Ralph Kronig (1904–1995, German) Ralph de Laer Kronig grew up in Dresden and earned his doctorate at Columbia University in New York in 1925. Early that year he had suggested that the electron spins, but he did not publish the idea after Wolfgang Pauli rejected it; months later George Uhlenbeck and Samuel Goudsmit proposed it independently. After a period as Pauli's assistant in Zurich and some years at Groningen, he was professor of theoretical physics at the technical university in Delft from 1939 to 1969, and its rector from 1959 to 1962. The Kronig–Penney model (1931) of electrons in a crystal also bears his name. (Wikipedia)

22.5.4 Causality and the Kramers–Kronig relations

Figure 22.13 shows a pattern that every material follows, not only the model of this section: wherever ε_r'' has a peak, ε_r' steps down as the frequency rises through it. The reason is **causality**: a material cannot respond to a field before the field arrives. That fact alone ties the real and imaginary parts of $\tilde{\varepsilon}_r$ together, so that if ε_r'' is known at every frequency, ε_r' is determined, and the other way around. For a material that does not conduct at zero frequency,

$$\begin{aligned} \varepsilon_r'(\omega) - 1 &= \frac{2}{\pi}\,\mathrm{P}\int_0^\infty \frac{\omega'\varepsilon_r''(\omega')}{\omega'^2 - \omega^2}\,d\omega', \\ \varepsilon_r''(\omega) &= -\frac{2\omega}{\pi}\,\mathrm{P}\int_0^\infty \frac{\varepsilon_r'(\omega') - 1}{\omega'^2 - \omega^2}\,d\omega'. \end{aligned} \tag{22.25}$$

These are the **Kramers–Kronig relations**, published by Ralph Kronig in 1926 and, independently, by Hendrik Kramers in 1927.[7] The P means the **principal value**: the integrand grows without limit at $\omega' = \omega$, with opposite signs on the two sides, so the integral is taken from 0 to $\omega - \eta$ and from $\omega + \eta$ to ∞, where the large contributions from the two sides cancel, and then $\eta \to 0$. Appendix C derives the relations from causality.

[7] "On the Theory of Dispersion of X-Rays," R. de L. Kronig, J. Opt. Soc. Am. 12, 547–557 (1926). https://doi.org/10.1364/JOSA.12.000547 and "La diffusion de la lumière par les atomes," H. A. Kramers, Atti del Congresso Internazionale dei Fisici (Como) 2, 545–557 (1927).

At $\omega = 0$ the first relation gives

$$\varepsilon_r(0) - 1 = \frac{2}{\pi}\int_0^\infty \frac{\varepsilon_r''(\omega')}{\omega'}\,d\omega',$$

so every frequency at which a material absorbs adds to its static dielectric constant. Far above an absorption peak, on the other hand, the peak's contribution to the first integral is small, and negative, since there $\omega'^2 - \omega^2 < 0$. So each peak raises ε_r' at frequencies below it and not above it, which is why each step in Figure 22.13 lines up with a peak. And a material that absorbed nothing at any frequency would have $\varepsilon_r' = 1$ at every frequency: only empty space is transparent at all frequencies. The relations hold as written in either sign convention for the time dependence, since ε_r'' is positive in both (Section 18.5). For a conductor, whose ε_r'' grows as $\sigma/\varepsilon_0\omega$ at low frequencies (22.20), they need a correction for the steady current, which this book does not go into. Exercise P22.14 checks the first relation for the Lorentz model on a computer.

22.6 Summary

A wave that decays as it travels, or that cannot travel away from a boundary at all, has a complex wave number: its real part sets the wavelength and its imaginary part the decay. In matter the decay comes from the response of the charges, which lags behind the field and depends on the frequency. Every real material is partly a conductor and partly an insulator, and complex response functions describe both parts at once.

Total internal reflection

$\kappa = \dfrac{\omega}{c}\sqrt{n_1^2\sin^2\theta_I - n_2^2}$, $e^{-\kappa z}\,e^{i(k_x x-\omega t)}$	beyond the critical angle: the evanescent wave travels along the boundary and decays away from it
$\tilde\alpha = i\lvert\tilde\alpha\rvert$, $\quad \lvert\tilde r_p\rvert = 1$	all of the power is reflected, shifted in phase; on average no energy crosses the boundary

Waves in conductors

$\nabla^2\mathbf{E} = \mu\varepsilon\dfrac{\partial^2\mathbf{E}}{\partial t^2} + \mu\sigma\dfrac{\partial\mathbf{E}}{\partial t}$	damped wave equation; free charge dies out, so $\rho_f = 0$
$\tilde k^2 = \mu\varepsilon\omega^2 + i\mu\sigma\omega$, $\quad \tilde k = k_{\text{real}} + i\,k_{\text{imag}}$	the wave is $e^{-k_{\text{imag}}z}\,e^{i(k_{\text{real}}z-\omega t)}$
$\delta = \sqrt{\dfrac{2}{\mu\sigma\omega}}$	skin depth of a good conductor ($\sigma \gg \varepsilon\omega$), where $k_{\text{real}} = k_{\text{imag}} = 1/\delta$
$\tilde{\mathbf{B}}_0 = \dfrac{\tilde k}{\omega}\,\hat{\mathbf{z}}\times\tilde{\mathbf{E}}_0$	**B** lags behind **E** by the phase of $\tilde k$, 45° in a good conductor; the energy is mostly magnetic
intensity $\propto e^{-2k_{\text{imag}}z}$	$2k_{\text{imag}}$ is the absorption coefficient; the lost power heats the conductor

Reflection from a conductor

$\tilde{n} = \dfrac{c\tilde{k}}{\omega} = n + i\kappa$ — complex index of refraction; κ is the extinction coefficient (not the κ of total internal reflection)

$\tilde{r} = \dfrac{n_1 - \tilde{n}_2}{n_1 + \tilde{n}_2}, \quad \tilde{t} = \dfrac{2n_1}{n_1 + \tilde{n}_2}$ — normal incidence, materials not magnetic

$R = |\tilde{r}|^2, \quad T = \mathrm{Re}(\tilde{\beta})\,|\tilde{t}|^2, \quad R + T = 1$ — a purely imaginary $\tilde{n}$ gives $R = 1$

$R \approx 1 - 2\sqrt{2\varepsilon_0\omega/\sigma}$ — good conductor in air

Complex permittivity and conductivity

$\tilde{\mathbf{P}} = \varepsilon_0\tilde{\chi}_e\tilde{\mathbf{E}}, \quad \tilde{\mathbf{D}} = \varepsilon_0\tilde{\varepsilon}_r\tilde{\mathbf{E}}$ — complex amplitudes at one frequency only

$\left\langle \mathbf{E}\cdot\dfrac{\partial\mathbf{P}}{\partial t}\right\rangle = \dfrac{1}{2}\omega\varepsilon_0\varepsilon_r''|\tilde{\mathbf{E}}|^2$ — $\varepsilon_r'' \geq 0$ absorbs; ε_r' may have either sign

$\tilde{n}^2 = \tilde{\varepsilon}_r$ — materials not magnetic; $\tilde{n}$ in the first quadrant

$\tilde{\sigma} = \dfrac{\sigma_0}{1 - i\omega\tau}, \quad \sigma_0 = \dfrac{Ne^2\tau}{m}$ — equal to σ_0 up to the infrared for a good metal

$\tilde{\varepsilon}_{r,\mathrm{eff}} = \tilde{\varepsilon}_r + \dfrac{i\tilde{\sigma}}{\varepsilon_0\omega}$ — conduction or polarization: bookkeeping, except at $\omega = 0$

The Lorentz oscillator model

$\tilde{\varepsilon}_r = 1 + \dfrac{\omega_p^2}{\omega_0^2 - \omega^2 - i\omega\Gamma}, \quad \omega_p^2 = \dfrac{Nq^2}{m\varepsilon_0}$ — bound charges; an absorption line near ω_0

$\tilde{\varepsilon}_r = \varepsilon_\infty + \omega_p^2\sum_j \dfrac{f_j}{\omega_j^2 - \omega^2 - i\omega\Gamma_j}$ — several resonances: electrons in the ultraviolet, ions in the infrared

$n \approx A + \dfrac{B}{\lambda_0^2} + \dfrac{C}{\lambda_0^4}$ — Cauchy's formula, far below the resonances; normal dispersion

$\tilde{\varepsilon}_r = 1 - \dfrac{\omega_p^2}{\omega^2 + i\omega\Gamma}$ — free electrons: a good conductor below Γ, a mirror between Γ and ω_p, transparent above ω_p

$\varepsilon_r'(\omega) - 1 = \dfrac{2}{\pi}\,\mathrm{P}\displaystyle\int_0^\infty \dfrac{\omega'\varepsilon_r''(\omega')}{\omega'^2 - \omega^2}\,d\omega'$ — Kramers–Kronig: causality ties ε_r' to ε_r''; each absorption peak raises ε_r' below it

Exercises

Exercises for 22.1 Total internal reflection

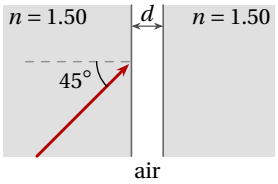


Figure 22.18 The two blocks of glass of Exercise P22.1, separated by a gap of air of width d, and the incident light.

P22.1 Frustrated total internal reflection. Light inside a block of glass with $n = 1.50$ strikes a flat face of the block at 45°, and a second block of the same glass is held parallel to that face, with a gap of air of width d between the two (Figure 22.18).

(a) When d is large, none of the light gets into the second block. Explain why some of it does when d is small, and about how small d must be.

(b) When κd is large, the fraction of the power that crosses the gap is roughly proportional to $e^{-2\kappa d}$. For light with $\lambda_0 = 633\,\text{nm}$, by what factor does the power that crosses fall when the gap is widened by 100 nm?

(c) Repeat (b) for light that strikes the face at 60°.

(d) This effect is the optical counterpart of an effect in quantum mechanics. Which one?

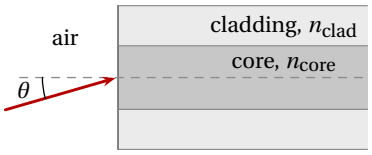


Figure 22.19 The end of the optical fiber of Exercise P22.2, cut along its axis: the core, the cladding, and a ray that enters from air at the angle θ from the axis.

P22.2 An optical fiber. A step-index optical fiber has a core of glass with the index of refraction n_{core}, surrounded by a cladding of glass with a slightly smaller index n_{clad}. Light enters the flat end of the fiber from air, at the angle θ from the fiber's axis (Figure 22.19).

(a) Show that a ray in a plane that contains the axis is trapped in the core by total internal reflection if $\sin\theta < \sqrt{n_{core}^2 - n_{clad}^2}$. This quantity is called the numerical aperture of the fiber.

(b) Typical values for a fiber with a large core are $n_{core} = 1.48$ and $n_{clad} = 1.46$. Find the largest angle θ at which light is trapped.

(c) A ray of light with $\lambda_0 = 850\,\text{nm}$ enters at half the angle of (b). How far into the cladding does its evanescent wave reach before its amplitude has fallen by the factor $1/e$? Why must the cladding be much thicker than this?

Exercises for 22.2 Waves in conductors

P22.3 Free charge in a metal. Section 22.2.1 found that, by Ohm's law, free charge inside a conductor dies away in the time ε_0/σ, which for copper is 1.5×10^{-19} s. Copper has $\sigma = 5.96 \times 10^7\,\text{S/m}$ and 8.5×10^{28} conduction electrons per cubic meter.

(a) In Drude's scattering model (Section 19.3), each electron is accelerated by the field and loses the extra velocity in collisions with the atoms, on average after the time τ. Then $\sigma = Ne^2\tau/m$, where N is the number of conduction electrons per unit volume. Find τ for copper, and the ratio $\tau/(\varepsilon_0/\sigma)$.

(b) On time scales shorter than τ Ohm's law does not hold. A better description is the equation of motion for the average velocity $\mathbf{v}$ of the conduction electrons,

$$m\frac{d\mathbf{v}}{dt} = -e\mathbf{E} - \frac{m}{\tau}\mathbf{v},$$

in which the last term stands for the collisions. With $\mathbf{J}_f = -Ne\mathbf{v}$, show that the free charge density in the metal obeys

$$\frac{d^2\rho_f}{dt^2} + \frac{1}{\tau}\frac{d\rho_f}{dt} + \omega_p^2\rho_f = 0, \qquad \omega_p^2 = \frac{Ne^2}{m\varepsilon_0}.$$

(Multiply the equation of motion by $-Ne$, take its divergence, and use the continuity equation and Gauss's law with ε_0.)

(c) Show that when the first term is dropped, this equation gives the decay time ε_0/σ of Section 22.2.1.

(d) Solve the full equation for copper. Show that the charge oscillates, and find the time in which the amplitude of its oscillation falls by the factor $1/e$. Compare it with ε_0/σ.

P22.4 A poor conductor.

(a) Show that when $\sigma = 0$, (22.8) gives the real wave number $k = \omega\sqrt{\mu\varepsilon}$ of Chapter 21.

(b) Now let σ be small but not zero, $\sigma \ll \varepsilon\omega$: a material that is nearly a perfect insulator but conducts a little. Expand (22.8) to the lowest order in $\sigma/\varepsilon\omega$ at which each of k_{real} and k_{imag} differs from its value for $\sigma = 0$. Show that k_{real} depends on ω but k_{imag} does not.

(c) Pure water at 20°C has the dielectric constant 80.1 (Table 9.2) and the resistivity $2.385 \times 10^5\,\Omega\,\text{m}$.[8] Taking its permittivity to keep its static value, find its skin depth, and the frequencies above which the result of (b) applies.

[8] "The fundamental conductivity and resistivity of water," T. S. Light, S. Licht, A. C. Bevilacqua, and K. R. Morash, Electrochem. Solid-State Lett. 8, E16–E19 (2005). https://doi.org/10.1149/1.1836121

P22.5 Silver plating for a microwave cavity.

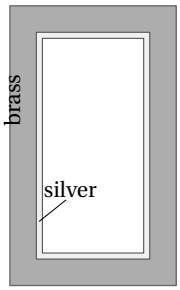


Figure 22.20 The microwave cavity of Exercise P22.5, cut along its axis: brass walls with a layer of silver on the inside (its thickness exaggerated).

(a) Show that for a good conductor the skin depth (22.9) can be written $\delta = \sqrt{\rho/\pi\nu\mu}$, where $\rho = 1/\sigma$ is the resistivity and ν the frequency. Notice that it does not depend on the permittivity.

(b) A student and I once designed a resonant cavity for microwaves near 10 GHz, a cylinder about 1.9 cm across and 3.8 cm tall (Figure 22.20).[9] It was machined from brass, and to make its walls reflect the microwaves with as little loss as possible we electroplated the inside with silver, which conducts better. Silver is expensive, so the plating should be no thicker than it needs to be. How thick must it be to be 10 skin depths thick at 10 GHz, so that the microwaves interact only with the silver and not with the brass? The resistivity of silver is $1.587 \times 10^{-8}\,\Omega\,\text{m}$ at 20°C (Table 19.2).

(c) How fast do the wave fronts of these microwaves travel in the silver? Find $v = \omega/k_{\text{real}}$.

P22.6 Where the energy goes. A plane wave travels along z in a conductor with real ε, μ, and σ.

(a) Show that the power that the wave loses per unit volume, $-d\langle S_z\rangle/dz$, with $\langle S_z\rangle$ from Section 22.2.3, equals $\langle \mathbf{E}\cdot\mathbf{J}_f\rangle = \frac{1}{2}\sigma|\tilde{\mathbf{E}}|^2$, the average rate at which the field does work on the free charges. (What is the imaginary part of $\tilde{k}^2$?)

(b) A good conductor fills the region $z > 0$, and the wave enters it at $z = 0$. Show that the current per unit width that flows in the conductor, $\tilde{K} = \int_0^\infty \tilde{J}_f\,dz$, and the power absorbed per unit area of its surface, $\langle S_z\rangle$ at $z = 0$, are related by

$$\langle S_z\rangle|_{z=0} = \frac{|\tilde{K}|^2}{2\sigma\delta}.$$

(c) Show that this is the power per unit area that the same current would dissipate if it flowed uniformly in a layer of thickness δ. This is why the resistance of a thick wire at high frequencies is that of a tube one skin depth thick (Example 22.2).

[9] "Resonant microwave cavity for 8.5–12 GHz optically detected electron spin resonance with simultaneous nuclear magnetic resonance," J. S. Colton and L. R. Wienkes, Rev. Sci. Instrum. 80, 035106 (2009). https://doi.org/10.1063/1.3095683

Exercises for 22.3 Reflection from a conductor

P22.7 Energy at the surface of a conductor.

(a) Show that the reflectance and the transmittance (22.14) satisfy $R + T = 1$ for any complex $\tilde{\beta}$.

(b) Light in air strikes copper at normal incidence at 1.0 GHz. What fraction of its power enters the copper? At what depth below the surface has that power fallen to 1% of its value just inside?

(c) What becomes of the power that enters the copper?

Exercises for 22.4 Complex permittivity and complex conductivity

P22.8 Sodium in blue light. The measured optical constants of sodium at the vacuum wavelength 460 nm are $n = 0.063$ and $\kappa = 1.86$.[10]

(a) Find $\tilde{\varepsilon}_r$. Which of its parts is larger, and what does the sign of its real part tell you about the polarization of the metal?

(b) Find the absorption coefficient, and the depth below the surface at which the intensity of the light has fallen by the factor $1/e$.

(c) Find the reflectance at normal incidence from air.

(d) Compare your $\tilde{\varepsilon}_r$ with the Drude model (22.24) at this wavelength, with the values of ω_p and Γ for sodium that you found in Tutorial 22.2. Which part agrees well with the model, and which does not?

P22.9 The collision time from optical data. Section 22.4.3 described the response of a metal either by an effective permittivity $\tilde{\varepsilon}_{r,\text{eff}}$ or by a total conductivity $\tilde{\sigma}_{\text{tot}}$.

(a) Show that $\tilde{\sigma}_{\text{tot}} = -i\omega\varepsilon_0(\tilde{\varepsilon}_{r,\text{eff}} - 1)$.

(b) Suppose that the free electrons are the only charges that respond, so that $\tilde{\sigma}_{\text{tot}}$ is the $\tilde{\sigma}$ of (22.19). Show that when $\omega\tau \gg 1$, the imaginary parts of the two expressions are equal if

$$\tau = \frac{\sigma_0}{\varepsilon_0\omega^2(1-\varepsilon_r')}.$$

(c) Find τ for sodium from the optical constants of Exercise P22.8 and the resistivity given in Tutorial 22.2. Compare it with the collision time $1/\Gamma$ that you found from the resistivity alone in the tutorial.

[10] From "Optical and photoelectric properties of alkali metals," J. Monin and G.-A. Boutry, Phys. Rev. B 9, 1309–1327 (1974), https://doi.org/10.1103/PhysRevB.9.1309, as tabulated at https://refractiveindex.info.

(d) Suppose that the -1 in (a) were left out. By what factor would τ be wrong? What current would then be attributed to the electrons, although they do not carry it?

P22.10 A very good insulator. The best optical fibers lose only 0.14 dB of the power of the light per kilometer at the vacuum wavelength 1550 nm.[11] (A loss of L decibels means that the power falls by the factor $10^{-L/10}$.) Most of this loss is light scattered sideways by tiny variations in the density of the glass, frozen in when it solidified; only a small part is absorbed. Treat it all as absorption here, so that your answers to (b) and (c) are upper limits.

(a) What fraction of the power is left after 100 km of this fiber?

(b) Find the absorption coefficient of the glass, and its κ.

(c) At this wavelength the glass has $n = 1.444$ (from the formula of Section 22.5.2). Find ε_r'', and the conductivity σ that, put into (22.20) with a real ε_r, would give the same ε_r''. Compare it with the conductivity of copper.

(d) Is the glass a conductor or an insulator?

Exercises for 22.5 The Lorentz oscillator model

P22.11 Cauchy's formula for air. The file `ch21-air-index.csv`, on the book's web site at https://em.byu.edu/data, lists the index of refraction of standard dry air at vacuum wavelengths from 300 to 1700 nm in steps of 20 nm. (Standard air is at 15°C and 101.325 kPa, with 450 parts per million of carbon dioxide.) After a header line, each line holds a wavelength in nanometers and the index at that wavelength, separated by a comma. The values were computed from the equations of Ciddor, which fit the best measurements of the index of air.[12]

(a) Fit the data to the first two terms of Cauchy's formula (22.23), $n = A + B/\lambda_0^2$, with λ_0 in micrometers. Plot the data with the fitted curve, and plot the residuals, the data minus the fit.

(b) Fit the data again with the third term, C/λ_0^4, included. By about what factor are the largest residuals smaller?

(c) Explain from Section 22.5.2 why A is slightly larger than 1, and why B is positive.

[11] "The first 0.14-dB/km loss optical fiber and its impact on submarine transmission," Y. Tamura, H. Sakuma, K. Morita, M. Suzuki, Y. Yamamoto, K. Shimada, Y. Honma, K. Sohma, T. Fujii, and T. Hasegawa, J. Lightwave Technol. 36, 44–49 (2018). https://doi.org/10.1109/JLT.2018.2796647

[12] "Refractive index of air: new equations for the visible and near infrared," P. E. Ciddor, Appl. Opt. 35, 1566–1573 (1996). https://doi.org/10.1364/AO.35.001566

P22.12 The reststrahlen band and the Lyddane–Sachs–Teller relation. The ions of a crystal such as sodium chloride resonate in the infrared, at ω_0. Near that frequency the electrons, whose resonances are in the ultraviolet, contribute only a constant, so that

$$\tilde{\varepsilon}_r = \varepsilon_\infty + \frac{\omega_p^2}{\omega_0^2 - \omega^2 - i\omega\Gamma}.$$

(a) Take $\Gamma = 0$. Show that ε_r is negative, so that the reflectance at normal incidence from air is $R = 1$, at frequencies between ω_0 and a frequency ω_L that you should find. Such a band is called a reststrahlen band, from the German for "residual rays": after several reflections from the crystal, only light in the band remains.

(b) Sketch ε_r against ω for $\Gamma = 0$. How does a small damping rate change the reflectance near the edges of the band?

(c) Show that the static value of the permittivity, $\varepsilon(0)$, satisfies

$$\frac{\omega_L^2}{\omega_0^2} = \frac{\varepsilon(0)}{\varepsilon_\infty}.$$

This is the Lyddane–Sachs–Teller relation, usually written with ω_T for ω_0.

(d) At ω_L the permittivity is zero, so $\tilde{\mathbf{D}} = \varepsilon_0\tilde{\varepsilon}_r\tilde{\mathbf{E}} = 0$ even if $\tilde{\mathbf{E}}$ is not zero. Show that a wave of polarization along its own direction of travel, $\tilde{\mathbf{P}} = \tilde{P}_0 e^{i(kx-\omega t)}\hat{\mathbf{x}}$, can then exist in the crystal with no free charge, with $\tilde{\mathbf{E}} = -\tilde{\mathbf{P}}/\varepsilon_0$. Why can a transverse electromagnetic wave not drive it? (This is why ω_L is called the frequency of the longitudinal oscillations of the crystal, and ω_T that of the transverse ones.)

P22.13 Sodium on logarithmic scales. Use the Drude model (22.24) with the values of ω_p and Γ for sodium that you found in Tutorial 22.2.

(a) Plot n and κ against ω on logarithmic scales, from $\omega = 0.01\,\Gamma$ to $\omega = 100\,\omega_p$. Mark Γ and ω_p, and identify the three ranges of frequency of Section 22.5.3.

(b) Show that for $\omega \ll \Gamma$, $n = \kappa = \omega_p/\sqrt{2\Gamma\omega}$, and compare with the slopes of your plot.

(c) Find how n and κ depend on ω for $\Gamma \ll \omega \ll \omega_p$ and for $\omega \gg \omega_p$, and compare with your plot.

P22.14 The Kramers–Kronig relations on a computer. Take the Lorentz model (22.21) with $\omega_p = \omega_0$ and $\Gamma = 0.1\,\omega_0$.

(a) Plot ε_r' and ε_r'' from 0 to $3\omega_0$.

(b) Evaluate the first of the Kramers–Kronig relations (22.25) numerically at about 50 frequencies from 0 to $3\omega_0$, using only ε_r'', and plot the results on top of your ε_r'. (The integrand is infinite at $\omega' = \omega$. Since $\mathrm{P}\int_0^\infty d\omega'/(\omega'^2 - \omega^2) = 0$, subtracting $\omega\varepsilon_r''(\omega)$ from its numerator does not change the integral, and the new integrand is finite. If you stop the integral at a frequency Ω instead of infinity, add back $\frac{1}{2}\varepsilon_r''(\omega)\ln[(\Omega-\omega)/(\Omega+\omega)]$, the integral of the subtracted term from 0 to Ω.)

(c) Check your value at $\omega = 0$ against the static dielectric constant that (22.21) gives.

(d) Show that far above the resonance the first relation gives $\varepsilon_r' - 1 \approx -(2/\pi\omega^2)\int_0^\infty \omega'\varepsilon_r''(\omega')\,d\omega'$. Evaluate the integral numerically, and check that it is $\pi\omega_p^2/2$, so that far above the resonance $\varepsilon_r' \approx 1 - \omega_p^2/\omega^2$, as for free electrons (22.24).

Exercises for 22.6 Summary

P22.15 True or false? Explain your reasoning.

(a) In total internal reflection, the fields on the far side of the boundary are zero.

(b) The skin depth of a good conductor does not depend on its permittivity.

(c) In a good conductor the electric and magnetic fields of a wave are in phase.

(d) A material whose index of refraction is purely imaginary reflects all of the light that strikes it at normal incidence.

(e) A material whose effective permittivity (22.20) is real at some frequency does not absorb energy from a wave of that frequency.

(f) The real part of the permittivity of a material can be negative.

(g) Glass is a perfect insulator at visible frequencies.

(h) Below its absorption lines, the index of refraction of glass decreases as the frequency increases.

(i) At any frequency other than zero, the current of a metal's free electrons can equally well be described as a polarization.

(j) According to the Drude model, a metal is transparent to waves whose frequency is far above its plasma frequency.

P22.16 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as light tunneling across a narrow gap between two prisms, the fields of a wave in a conductor at any frequency, the reflectance of a metal from radio waves to the ultraviolet, the permittivity of a material with several resonances, the spectrum spread out by a prism, or the path of a radio wave sent at an angle into the ionosphere.

Chapter 23

Confined Waves

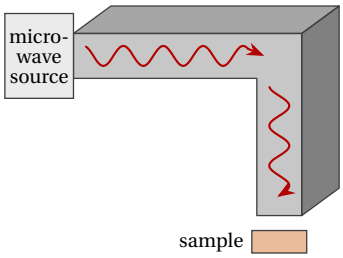


Figure 23.1 A waveguide carrying microwaves from their source to a sample.

The waves of Chapters 21 and 22 filled all of space, or met a single boundary between two materials. This chapter confines them. A hollow metal pipe, called a **waveguide**, can carry microwaves from the place where they are made to the place where they are needed, such as a sample in an experiment (Figure 23.1), and lose very little of them on the way, because its walls are good conductors and nothing passes through them. A closed metal box holds a standing wave, but only at certain frequencies, and a pair of conductors, such as a coaxial cable, can carry signals of any frequency, all the way down to zero. The walls change the waves in several ways. A wave confined between conducting walls cannot be a plane wave, it need not be transverse, it travels only above a minimum frequency, and its speed depends on its frequency even though the pipe is empty.

The chapter starts with the boundary conditions at the surface of a perfect conductor (Section 23.1), finds the general equations for a wave in a hollow pipe (Section 23.2), and solves them for a pipe of rectangular cross section (Section 23.3). A wave in a guide has two different speeds, which leads to the distinction between the phase velocity and the group velocity (Section 23.4). Then come pipes of round cross section, guides of any shape solved by computer, resonant cavities, and transmission lines (Sections 23.5–23.8). Throughout, the fields have the time dependence $e^{-i\omega t}$ of the time-delay convention (Section 18.2.4).

23.1 Boundary conditions at a perfect conductor

The walls of a waveguide are made of a good conductor, typically copper, brass, or aluminum, and are often plated with silver or gold. The fields of a wave reach into a good conductor only about a skin depth (22.9). For copper at 10 GHz, a typical frequency for a waveguide, the skin depth is 0.65 μm, while the guide itself is a few centimeters across. So the fields hardly enter the walls, and for finding the fields in the guide the walls can be treated as **perfect conductors**, the limit $\sigma \to \infty$ of Section 19.1.3. (The thin layer that the fields do enter is where real walls absorb a small part of the wave's energy. This chapter mostly ignores that loss.)

Inside a perfect conductor the electric field is zero, since any field would drive

an infinite current. The magnetic field inside is zero too, as long as it changes. Faraday's law gives $\partial \mathbf{B}/\partial t = -\nabla \times \mathbf{E} = 0$ inside, so the magnetic field there cannot change at all; a field that oscillates, as the field of a wave does, must therefore be zero there.

With both fields zero below the surface, the boundary conditions of Table 19.1 take a simple form. Let $\hat{\mathbf{n}}$ point out of the conductor, into the space where the waves are. Then just outside the surface,

$$\boxed{\mathbf{E}^{\parallel} = 0, \qquad B^{\perp} = 0}. \tag{23.1}$$

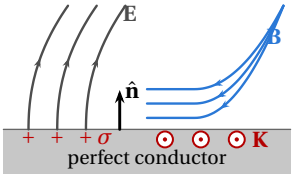


Figure 23.2 Just outside the surface of a perfect conductor (gray), the electric field (black) meets the surface at right angles, and the magnetic field (blue) runs along it. The electric field lines end on surface charge (+), and the magnetic field goes with a surface current **K**, here out of the page (⊙).

With words, electric field lines meet a perfect conductor at right angles, and magnetic field lines run along it (Figure 23.2). These two conditions restrict the fields. The other two conditions, (5.5) and (11.11) with zero fields below the surface, say what the surface charge and the surface current must be:

$$\sigma = \varepsilon_0 \hat{\mathbf{n}} \cdot \mathbf{E}, \qquad \mathbf{K} = \frac{1}{\mu_0} \hat{\mathbf{n}} \times \mathbf{B}, \tag{23.2}$$

with the fields evaluated just outside. (The second comes from taking the cross product of $\hat{\mathbf{n}}$ with both sides of $\mathbf{B} = \mu_0 \mathbf{K} \times \hat{\mathbf{n}}$, since $\hat{\mathbf{n}} \times (\mathbf{K} \times \hat{\mathbf{n}}) = \mathbf{K}$ for a $\mathbf{K}$ that is perpendicular to $\hat{\mathbf{n}}$.) The conductor supplies whatever charge and current the fields require. Those are not known in advance; once the fields are found, (23.2) gives them (Exercise P23.5).

The conditions (23.1) are a little like the infinite square well of quantum mechanics, whose walls force the wave function to be zero at its edges. Here the walls force $\mathbf{E}^{\parallel}$ and $B^{\perp}$ to be zero, and, as in the square well, only certain patterns fit between them.

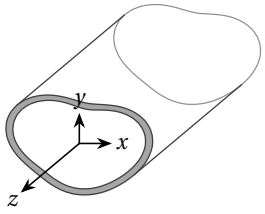


Figure 23.3 A hollow guide: a conducting pipe whose cross section, here of an arbitrary shape, is the same all along the z axis. The axes form a right-handed set, with z toward the viewer.

23.2 Waves in a hollow guide

Take a hollow pipe with perfectly conducting walls, whose cross section can have any shape but is the same all along its length, and let the z axis run along the pipe (Figure 23.3). The inside is empty. We look for waves that travel along the guide: fields of the form

$$\tilde{\mathbf{E}}(x,y,z,t) = \tilde{\mathbf{E}}_0(x,y)\, e^{i(kz-\omega t)}, \qquad \tilde{\mathbf{B}}(x,y,z,t) = \tilde{\mathbf{B}}_0(x,y)\, e^{i(kz-\omega t)}. \tag{23.3}$$

With words, whatever the fields do across the guide, at each point of the cross section they oscillate in z and t like a plane wave traveling along $+z$, with the angular frequency ω and the wave number k. The real fields are the real parts, as in Section 18.2.4. Unlike the amplitudes of a plane wave, though, $\tilde{\mathbf{E}}_0$ and $\tilde{\mathbf{B}}_0$ depend on x and y, because the fields must satisfy the boundary conditions (23.1) at the walls.

These fields cannot be purely transverse, with $\mathbf{E}$ and $\mathbf{B}$ perpendicular to z everywhere, as the fields of a plane wave are. A useful picture is that the wave bounces back and forth between the walls as it travels, like light in an optical

fiber (Exercise P22.2). It then travels at an angle to the z axis, and fields that are perpendicular to its direction of travel have components along z (Figure 23.4). Section 23.3.2 makes this picture exact for one wave. In general, all six components of the fields can be nonzero.

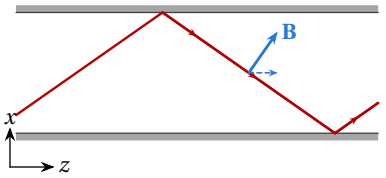


Figure 23.4 A wave that bounces between the walls of a guide travels at an angle to the z axis, so a field that is perpendicular to its direction of travel, such as the magnetic field **B** drawn here, has a component along z (dashed).

23.2.1 The longitudinal-to-transverse equations

Inside the guide there are no charges or currents, so the fields obey Maxwell's equations in empty space, (16.11). For the fields (23.3), each derivative with respect to z brings down a factor ik, and each derivative with respect to t a factor $-i\omega$ (Section 18.2.4). To keep the equations readable, from here on E_x, E_y, ..., B_z stand for the components of the amplitudes $\tilde{\mathbf{E}}_0(x,y)$ and $\tilde{\mathbf{B}}_0(x,y)$; the common factor $e^{i(kz-\omega t)}$ cancels from every equation. Then the two curl equations, $\nabla\times\mathbf{E}=-\partial\mathbf{B}/\partial t$ and $\nabla\times\mathbf{B}=(1/c^2)\,\partial\mathbf{E}/\partial t$, become six equations, one for each component:

$$\frac{\partial E_z}{\partial y}-ikE_y=i\omega B_x, \qquad \frac{\partial B_z}{\partial y}-ikB_y=-\frac{i\omega}{c^2}E_x, \tag{23.4}$$

$$ikE_x-\frac{\partial E_z}{\partial x}=i\omega B_y, \qquad ikB_x-\frac{\partial B_z}{\partial x}=-\frac{i\omega}{c^2}E_y, \tag{23.5}$$

$$\frac{\partial E_y}{\partial x}-\frac{\partial E_x}{\partial y}=i\omega B_z, \qquad \frac{\partial B_y}{\partial x}-\frac{\partial B_x}{\partial y}=-\frac{i\omega}{c^2}E_z. \tag{23.6}$$

The left column is Faraday's law and the right column is the Ampère–Maxwell law, and each row is one Cartesian component.

The first two rows can be solved for the transverse components, the x and y components, in terms of the longitudinal ones, E_z and B_z. For example, the right-hand equation of (23.4) gives B_y in terms of E_x and $\partial B_z/\partial y$, and the left-hand equation of (23.5) gives E_x in terms of B_y and $\partial E_z/\partial x$; eliminating B_y between them gives E_x. The results are four equations, which I call the **longitudinal-to-transverse equations**:

$$\begin{aligned}
E_x &= \frac{i}{(\omega/c)^2-k^2}\left(k\frac{\partial E_z}{\partial x}+\omega\frac{\partial B_z}{\partial y}\right),\\
E_y &= \frac{i}{(\omega/c)^2-k^2}\left(k\frac{\partial E_z}{\partial y}-\omega\frac{\partial B_z}{\partial x}\right),\\
B_x &= \frac{i}{(\omega/c)^2-k^2}\left(k\frac{\partial B_z}{\partial x}-\frac{\omega}{c^2}\frac{\partial E_z}{\partial y}\right),\\
B_y &= \frac{i}{(\omega/c)^2-k^2}\left(k\frac{\partial B_z}{\partial y}+\frac{\omega}{c^2}\frac{\partial E_z}{\partial x}\right).
\end{aligned} \tag{23.7}$$

The longitudinal-to-transverse equation for E_x

Solve the right-hand equation of (23.4) for B_y,

$$B_y = \frac{1}{ik}\left(\frac{\partial B_z}{\partial y} + \frac{i\omega}{c^2}E_x\right),$$

and the left-hand equation of (23.5) for E_x, and substitute:

$$E_x = \frac{1}{ik}\left(i\omega B_y + \frac{\partial E_z}{\partial x}\right) = \frac{1}{ik}\left[\frac{\omega}{k}\left(\frac{\partial B_z}{\partial y} + \frac{i\omega}{c^2}E_x\right) + \frac{\partial E_z}{\partial x}\right].$$

Multiply both sides by ik^2,

$$ik^2 E_x = \omega\frac{\partial B_z}{\partial y} + \frac{i\omega^2}{c^2}E_x + k\frac{\partial E_z}{\partial x},$$

and collect the terms with E_x on the left:

$$i\left(k^2 - \frac{\omega^2}{c^2}\right)E_x = k\frac{\partial E_z}{\partial x} + \omega\frac{\partial B_z}{\partial y}.$$

Dividing by $i(k^2 - \omega^2/c^2)$, and using $1/i = -i$, gives the first line of (23.7).

The other three come from the same two rows in the same way (Exercise P23.1).

Two things follow from (23.7). First, the fields in a guide need not be transverse: E_z and B_z can be nonzero. Second, once E_z and B_z are known, everything is known. (The equations fail when $(\omega/c)^2 = k^2$, as for a plane wave in empty space, which has $E_z = B_z = 0$: then both sides of each equation are $0/0$. Section 23.2.2 returns to that case.)

To find E_z and B_z, go back to the wave equation. Each Cartesian component of **E** and **B** obeys the wave equation (16.13) with $v = c$. For E_z of the form (23.3), $\partial^2/\partial z^2$ brings down $(ik)^2 = -k^2$ and $\partial^2/\partial t^2$ brings down $(-i\omega)^2 = -\omega^2$, so the wave equation becomes an equation in x and y alone. The same holds for B_z:

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\omega^2}{c^2} - k^2\right)E_z = 0, \qquad \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\omega^2}{c^2} - k^2\right)B_z = 0. \tag{23.8}$$

Each is a **Helmholtz equation** in two dimensions. It is Laplace's equation with one extra term, and the same methods solve it: separation of variables (Section 7.4) and relaxation on a grid (Section 23.6).

Each equation needs a boundary condition at the walls. For E_z the condition is direct. Every wall runs along the z axis, so E_z is parallel to every wall, and $\mathbf{E}^{\parallel} = 0$ gives $E_z = 0$ at the walls. For B_z it is less direct: the condition $B^{\perp} = 0$ is about the transverse components of **B**, and the longitudinal-to-transverse equations translate it into a condition on B_z. Tutorial 23.1 does this for a rectangular guide.

23.2.2 TE, TM, and TEM modes

E_z and B_z each obey their own Helmholtz equation with their own boundary condition, and the two problems are independent of each other. So there are two families of solutions. In one family, $E_z = 0$ and $B_z \neq 0$: the electric field is purely transverse, and these solutions are called **transverse electric (TE) modes**. In the other, $B_z = 0$ and $E_z \neq 0$: the magnetic field is purely transverse, and these are **transverse magnetic (TM) modes**. Each solution of either kind is called a **mode** of the guide. Just as any polarization of light can be written as a sum of s- and p-polarized light (Section 21.7.2), any field in the guide can be written as a sum of TE and TM modes, because the longitudinal-to-transverse equations are linear (Exercise P23.2).

A field with both $E_z = 0$ and $B_z = 0$, purely transverse like a plane wave, would be a **transverse electromagnetic (TEM) mode**. A hollow guide has no TEM mode.

Proof that a hollow guide has no TEM mode

With $E_z = B_z = 0$, the left-hand equation of (23.6) gives $\partial E_y/\partial x - \partial E_x/\partial y = 0$, and $\nabla \cdot \mathbf{E} = 0$ gives $\partial E_x/\partial x + \partial E_y/\partial y = 0$ (its third term, ikE_z, is zero). The first says that the transverse field (E_x, E_y) has no curl in the plane of the cross section, which has no holes in it, so it is a two-dimensional gradient field: $E_x = -\partial V/\partial x$ and $E_y = -\partial V/\partial y$ for some function $V(x, y)$. The second then says

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = 0:$$

V obeys Laplace's equation in the cross section. At the wall $\mathbf{E}^{\parallel} = 0$, so the derivative of V along the wall is zero, and V has the same value everywhere on the wall. A hollow guide has only one wall, so V has that one value on the whole boundary of the cross section, and by the first uniqueness theorem (Section 7.1.3) it has the same value everywhere inside. Then $E_x = E_y = 0$, so $\mathbf{E} = 0$, and Faraday's law, $i\omega\tilde{\mathbf{B}} = \nabla \times \tilde{\mathbf{E}} = 0$, gives $\mathbf{B} = 0$ as well. There is no wave at all.

The proof needs the guide to have only one conductor. Two separate conductors, such as the inner and outer conductors of a coaxial cable, can be at different potentials, so V need not be constant, and a TEM mode exists (Figure 23.5). That is the transmission line of Section 23.8.

Lord Rayleigh analyzed electromagnetic waves in hollow conducting tubes in 1897;[1] his portrait is in Section 28.2.4. Hollow waveguides came into practical use in the 1930s.[2]

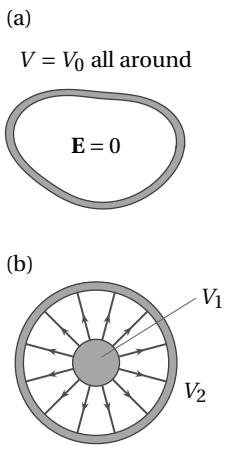


Figure 23.5 (a) The cross section of a hollow guide is bounded by one wall, so the potential of a TEM field would have one value all around it, and the field inside would be zero. (b) A coaxial cable has two conductors, which can be at different potentials, with a transverse electric field between them.

[1] "On the passage of electric waves through tubes, or the vibrations of dielectric cylinders," Lord Rayleigh, Phil. Mag. 43, 125–132 (1897). https://doi.org/10.1080/14786449708620969

[2] G. C. Southworth at Bell Telephone Laboratories and W. L. Barrow at the Massachusetts Institute of Technology developed them independently, and both reported their work in 1936: "Hyper-frequency wave guides—general considerations and experimental results," G. C. Southworth, Bell Syst. Tech. J. 15, 284–309 (1936), https://doi.org/10.1002/j.1538-7305.1936.tb00733.x, and "Transmission of electromagnetic waves in hollow tubes of metal," W. L. Barrow, Proc. IRE 24, 1298–1328 (1936). https://doi.org/10.1109/JRPROC.1936.227357

23.3 Rectangular waveguides

The most common guide has a rectangular cross section, with inside width a along x and inside height b along y, and by convention $a \geq b$ (Figure 23.6). Its walls are at $x = 0$, $x = a$, $y = 0$, and $y = b$. Tutorial 23.1 finds its TE modes. Work through it before reading on; the rest of this section uses its results.

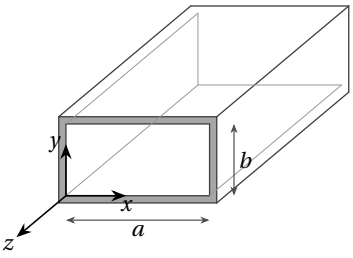


Figure 23.6 A rectangular waveguide, with inside width a and inside height b. The axes start at an inside corner of the guide, with z along it, toward the viewer.

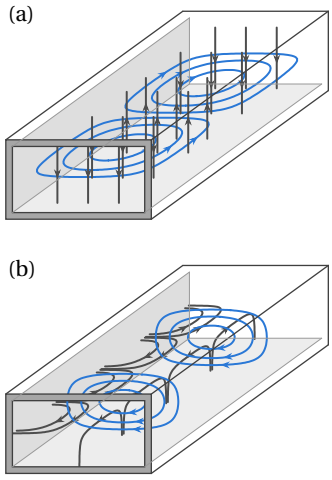


Figure 23.7 A representative TE mode and a representative TM mode of a rectangular waveguide at one instant, with the z axis along the tube. The black lines are electric field lines and the blue lines are magnetic field lines. The top and right walls are drawn transparent, to show the fields inside.

Tutorial 23.1 Rectangular waveguides

Printable version

Background

Waveguides are hollow metal tubes that can be used to carry electromagnetic waves, typically in the microwave range of frequencies. A rectangular waveguide with inside dimensions a and b is shown in Figure 23.6. The waves are confined between four walls, at $x = 0$ and $x = a$ (the two vertical walls) and at $y = 0$ and $y = b$ (the two horizontal walls).

Just as any polarization of light can be decomposed into s- and p-polarizations, any field inside a waveguide can be decomposed into two kinds of modes, called TE and TM. In a TE mode ("transverse electric"), the electric field has no z component, while the magnetic field has x, y, and z components. In a TM mode ("transverse magnetic"), the magnetic field has no z component, while the electric field has x, y, and z components.

If the z components of the fields are known, the x and y components follow from the longitudinal-to-transverse equations (23.7). For TE modes the E_z terms in them are all zero, and for TM modes the B_z terms are all zero. So once we find B_z (for TE modes) or E_z (for TM modes), the complete solution for all components of both fields is easily obtained. For TE modes we find B_z from its Helmholtz equation (23.8), which comes from the wave equation $\nabla^2\mathbf{B} = (1/c^2)\,\partial^2\mathbf{B}/\partial t^2$, each component of which is a wave equation of its own, by assuming the plane-wave-like dependence $e^{i(kz-\omega t)}$ on z and t, with some other dependence on x and y.

Part 1 — General

Figure 23.7 shows a representative TE mode and a representative TM mode, with the z axis along the tube. The black lines are electric field lines and the blue lines are magnetic field lines.

(a) Which picture is which mode? Explain your reasoning.

(b) Write down the four boundary conditions that come from Maxwell's equations. If you do not know any details about the free charge or free current at a surface, which two boundary conditions are most relevant?

(c) Describe how the two relevant boundary conditions are both satisfied by the TE and TM modes of Figure 23.7. Approximate the electric and magnetic fields inside the conductor as both being zero.

Part 2 — TE modes

(a) Now consider the TE modes specifically. As mentioned above, for the TE modes we focus on B_z. Assume that $B_z(x, y, z, t) = B_z(x, y)\, e^{i(kz-\omega t)}$, and show that the Helmholtz equation results from the wave equation. This notation is a little strange, because the B_z on the left is the complete field, while the B_z on the right is only its dependence on x and y across the guide, but unfortunately that is what is commonly done, including in this chapter.

(b) The Helmholtz equation is typically solved by separation of variables together with boundary conditions, so consider the boundary conditions further. For the TE modes we must consider the boundary condition on **B**. What constraint does it put on the components of **B** at the walls $x = 0$ and $x = a$? What constraint does it put on them at the walls $y = 0$ and $y = b$?

(c) The constraints you just found involve the x and y components of **B**, but the Helmholtz equation is for B_z. So use the longitudinal-to-transverse equations to connect the components to each other. Remembering that $E_z = 0$ for the TE modes, what do the longitudinal-to-transverse equations become?

(d) The TE modes of this guide have

$$B_z(x, y) = B_0 \cos\left(\frac{m\pi x}{a}\right) \cos\left(\frac{n\pi y}{b}\right),$$

where the integers m and n describe the details of the mode: how many nodes B_z has across the guide in the x and y directions. Each mode is labeled TE_{mn}, as in TE_{13} or TE_{25}. Show that this B_z, along with your previous results, makes the boundary conditions hold on the four walls. (It is enough to do $x = 0$ and $x = a$; the walls $y = 0$ and $y = b$ are similar.)

(e) This B_z looks as if it depends only on x and y, but as mentioned above, there is an underlying dependence on z and t as well. What is B_z when you include the z and t dependence explicitly? What do ω and k represent in your equation? It is not enough to say "angular frequency" and "wave number" here: what does the angular frequency mean in this context? What does the wave number mean?

Part 3 — Dispersion relation

Figure 23.8 shows k against ω for the TE modes of a waveguide 10 cm wide and 7 cm high. You will use it in step (c).

(a) Show that the full $B_z(x, y, z, t)$ of a TE mode does solve the Helmholtz equation, as long as k and ω are related by

$$k = \frac{1}{c}\sqrt{\omega^2 - \omega_{mn}^2}, \qquad \text{where} \qquad \omega_{mn} = c\pi\sqrt{\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2}.$$

(b) That equation for k as a function of ω is called the dispersion relation of the waveguide, and ω_{mn} is called the cutoff frequency of the mode. Note that ω_{mn} depends only on the mode numbers and on the size of the guide. What happens to k if ω is below the cutoff frequency of a mode? What does that mean for the physical behavior of the wave at that frequency?

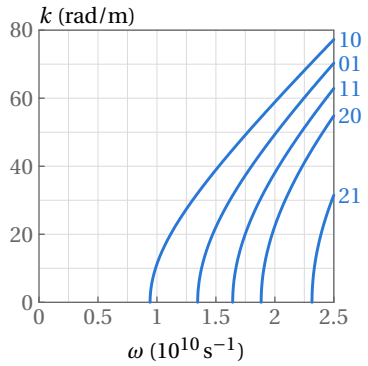


Figure 23.8 The wave number k against the angular frequency ω for the TE modes of a rectangular waveguide 10 cm wide and 7 cm high whose cutoff frequencies are below $2.5 \times 10^{10}\,\mathrm{s}^{-1}$. Each curve is labeled with its mode numbers mn.

(c) Look at the plot of k against ω in Figure 23.8. How many different TE modes can travel in that 10 cm × 7 cm guide at the frequency $\omega = 1 \times 10^{10}\,\mathrm{s}^{-1}$, and what are their wavelengths? How about at $\omega = 2 \times 10^{10}\,\mathrm{s}^{-1}$?

Part 4 — Wrap-up

(a) Check that each wavelength you found in step 3(c) is longer than the wavelength $2\pi c/\omega$ of a wave of the same frequency in empty space. Why should it be?

(b) As ω increases far above every cutoff frequency, what happens to k for each mode? Does that make sense?

23.3.1 The TE modes and their cutoff frequencies

Tutorial 23.1 found that the condition $B^\perp = 0$ at the walls becomes $\partial B_z/\partial x = 0$ at $x = 0$ and $x = a$, and $\partial B_z/\partial y = 0$ at $y = 0$ and $y = b$, and that

$$B_z(x, y) = B_0 \cos\left(\frac{m\pi x}{a}\right)\cos\left(\frac{n\pi y}{b}\right) \tag{23.9}$$

satisfies these conditions. This solution comes from separation of variables, as in Section 7.4.1. With $B_z = X(x)Y(y)$, the Helmholtz equation (23.8) divided by XY becomes

$$\frac{X''}{X} + \frac{Y''}{Y} = -\left(\frac{\omega^2}{c^2} - k^2\right).$$

The first term depends only on x and the second only on y, so each is a constant. Write $X''/X = -k_x^2$; then $X = A\cos k_x x + A'\sin k_x x$. The condition $X'(0) = 0$ requires $A' = 0$, and $X'(a) = 0$ then requires $\sin k_x a = 0$, so $k_x = m\pi/a$ with $m = 0, 1, 2, \ldots$. In the same way $Y = \cos(n\pi y/b)$, with $k_y = n\pi/b$. Each pair of integers (m, n) gives a mode, called TE_{mn}, and m and n count the half wavelengths of B_z across the width and the height of the guide (Figure 23.9).

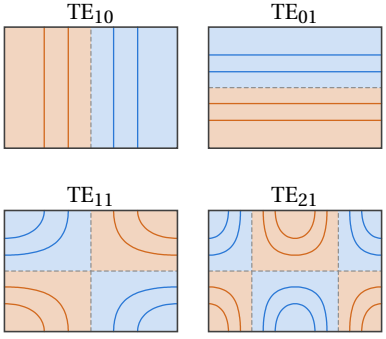


Figure 23.9 B_z of four TE modes of a rectangular guide, (23.9), in its cross section: orange where B_z is positive and blue where it is negative, with the nodes dashed. The mode numbers m and n count the half wavelengths of B_z across the width and the height.

The separation constants must add up to the right side: $(m\pi/a)^2 + (n\pi/b)^2 = \omega^2/c^2 - k^2$. Solved for k, this is the **dispersion relation** of the mode, the relation between its wave number and its frequency,

$$\boxed{k = \frac{1}{c}\sqrt{\omega^2 - \omega_{mn}^2}}, \tag{23.10}$$

where

$$\boxed{\omega_{mn} = c\pi\sqrt{\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2}} \tag{23.11}$$

is the **cutoff frequency** of the TE_{mn} mode. It depends only on the mode numbers and on the size of the guide.

The integers m and n cannot both be zero. Then B_z would be the same at every point of the cross section. But $\nabla \cdot \mathbf{B} = 0$ requires that as much magnetic flux

leave a thin slice of the guide, between z and $z+dz$, as enters it. None can leave through the walls, where $B^\perp = 0$, so the flux $\int B_z\,da$ through every cross section must be the same, and a uniform B_z that varies with z as e^{ikz} cannot satisfy that. So there is no TE_{00} mode.

For $a > b$ the mode with the lowest cutoff frequency is TE_{10}, with

$$\omega_{10} = \frac{\pi c}{a}. \tag{23.12}$$

It is called the **dominant mode** of the guide. The TM modes are found in the same way from E_z, which must be zero on all four walls (Exercise P23.4).

23.3.2 The TE_{10} mode

A rectangular guide is usually used at frequencies above ω_{10} but below the next cutoff frequency, where TE_{10} is the only mode that travels. For $m = 1$ and $n = 0$, $B_z = B_0\cos(\pi x/a)$ and $(\omega/c)^2 - k^2 = (\pi/a)^2$, and the longitudinal-to-transverse equations with $E_z = 0$ give the other components. With the factor $e^{i(kz-\omega t)}$ put back,

$$\begin{aligned}
\tilde{E}_y &= \frac{i\omega a}{\pi} B_0 \sin\frac{\pi x}{a}\, e^{i(kz-\omega t)},\\
\tilde{B}_x &= -\frac{ika}{\pi} B_0 \sin\frac{\pi x}{a}\, e^{i(kz-\omega t)},\\
\tilde{B}_z &= B_0 \cos\frac{\pi x}{a}\, e^{i(kz-\omega t)},
\end{aligned}$$

and $E_x = E_z = B_y = 0$. For a real B_0 the real fields are

$$\begin{aligned}
E_y &= -\frac{\omega a}{\pi} B_0 \sin\frac{\pi x}{a}\sin(kz-\omega t),\\
B_x &= \frac{ka}{\pi} B_0 \sin\frac{\pi x}{a}\sin(kz-\omega t),\\
B_z &= B_0 \cos\frac{\pi x}{a}\cos(kz-\omega t).
\end{aligned} \tag{23.13}$$

The electric field points straight across the guide, from one broad wall to the other. It is strongest at the middle and zero at the side walls, where it would otherwise be parallel to the wall. The magnetic field forms loops in planes parallel to the broad walls (Figure 23.10). Since $B_x = -(k/\omega)E_y$, the two are largest at the same places and times, while B_z is 90° out of phase with both.

The TE_{10} mode has a simple interpretation. Write $\sin(\pi x/a) = (e^{i\pi x/a} - e^{-i\pi x/a})/2i$ in $\tilde{E}_y$:

$$\tilde{E}_y = \frac{\omega a}{2\pi} B_0 \left[e^{i(\pi x/a + kz - \omega t)} - e^{i(-\pi x/a + kz - \omega t)} \right].$$

Each term is a plane wave, with the wave vector $\mathbf{k}_1 = (\pi/a)\hat{\mathbf{x}} + k\hat{\mathbf{z}}$ or $\mathbf{k}_2 = -(\pi/a)\hat{\mathbf{x}} + k\hat{\mathbf{z}}$. By the dispersion relation, both have the magnitude $\sqrt{(\pi/a)^2 + k^2} = \omega/c$, as

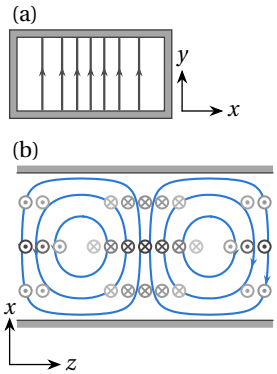


Figure 23.10 The fields of the TE_{10} mode (23.13) at one instant. (a) A cross section where the electric field (black) is strongest. (b) A view from above of one wavelength of the guide: the magnetic field (blue) forms loops in planes parallel to the broad walls, and the electric field points into (⊗) and out of (⊙) the page, each symbol shaded in proportion to the field's strength. The pattern travels along $+z$.

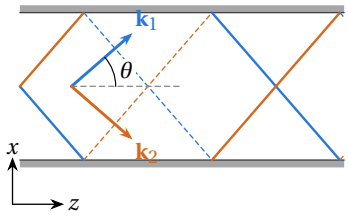


Figure 23.11 The TE_{10} mode as two plane waves, seen from above, with the side walls at the top and bottom (gray). Solid lines are crests and dashed lines are troughs of the wave traveling along $\mathbf{k}_1$ (blue) and of the wave traveling along $\mathbf{k}_2$ (orange). At each wall a crest of one wave meets a trough of the other, and their fields cancel. Drawn for $\omega = 1.53\,\omega_{10}$, so $\theta = 41°$.

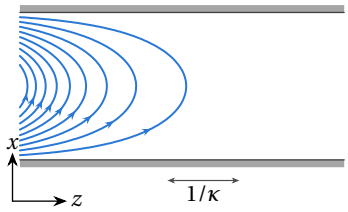


Figure 23.12 The magnetic field of the TE_{10} mode below its cutoff frequency, seen from above as in Figure 23.10b, at an instant when it is largest. It does not travel along the guide, but falls off by the factor $1/e$ in each distance $1/\kappa$; drawn for the guide of Example 23.1 at 5.0 GHz.

plane waves in empty space must. So the TE_{10} mode is two plane waves that travel at the angle θ on either side of the z axis (Figure 23.11), with

$$\sin\theta = \frac{\pi/a}{\omega/c} = \frac{\omega_{10}}{\omega}. \tag{23.14}$$

Each wave reflects from a side wall into the other one, and the two cancel at the walls, as $\mathbf{E}^{\parallel} = 0$ requires. This is the picture of a wave that bounces between the walls (Section 23.2), made exact. Near the cutoff frequency θ approaches 90°, and the waves bounce almost straight back and forth across the guide while they hardly travel along it. Far above it θ is small, and they travel nearly straight down the guide.

23.3.3 Below cutoff

If $\omega < \omega_{mn}$, the dispersion relation (23.10) gives an imaginary wave number. As for the transmitted wave beyond the critical angle in Section 22.1.1, the root is chosen by the physics: $k = i\kappa$, with

$$\kappa = \frac{1}{c}\sqrt{\omega_{mn}^2 - \omega^2} \tag{23.15}$$

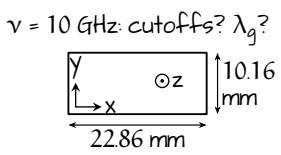
real and positive, so that the fields are proportional to $e^{ikz}e^{-i\omega t} = e^{-\kappa z}e^{-i\omega t}$. These fields do not travel along the guide. Each component oscillates in step at every z, and its amplitude falls off exponentially with z, by the factor $1/e$ in the distance $1/\kappa$: the fields form an evanescent wave, like the transmitted wave of total internal reflection (Figure 23.12). In the longitudinal-to-transverse equations, $k = i\kappa$ makes the transverse electric and magnetic fields 90° out of phase with each other, so the time-averaged Poynting vector along the guide is zero: on average no energy flows down the guide. In a guide with perfectly conducting walls nothing absorbs energy either. So a wave below cutoff that is sent into a guide is reflected back toward its source, just as light is in total internal reflection, and as radio waves below the plasma frequency are by the ionosphere (Example 22.4).

Each mode travels only above its own cutoff frequency, and below the lowest one, ω_{10}, a guide carries nothing at all: a guide is a high-pass filter. In a guide a few centimeters across, ω_{10} is in the gigahertz range, and that is why hollow waveguides are used for microwaves.

Example 23.1

A common rectangular waveguide for frequencies near 10 GHz, called WR-90, has inside dimensions 22.86 mm by 10.16 mm.[3] (a) Find the cutoff frequencies $\nu_{mn} = \omega_{mn}/2\pi$ of its three lowest TE modes. Over what range of frequencies does only one mode travel? (b) A 10.0 GHz wave travels in the TE_{10} mode. Find its wavelength in the guide, $\lambda_g = 2\pi/k$, and compare it with the wavelength in empty space. Find its phase velocity. (c) Over what distance does the field of a 5.0 GHz signal in this guide fall by the factor $1/e$?

Solution:



The first step is a sketch of what is given and what is asked: the cross section of the guide with its dimensions, the frequency, and the quantities we want.

(a) By (23.11), $\nu_{mn} = (c/2)\sqrt{(m/a)^2 + (n/b)^2}$. With $a = 22.86\,\text{mm}$ and $b = 10.16\,\text{mm}$,

$$\nu_{10} = \frac{c}{2a} = \frac{2.998\times10^8\,\text{m/s}}{2(0.02286\,\text{m})} = 6.557\,\text{GHz},$$
$$\nu_{20} = \frac{c}{a} = 13.11\,\text{GHz}, \qquad \nu_{01} = \frac{c}{2b} = 14.75\,\text{GHz},$$

and every other TE mode has a higher cutoff (TE_{11}, the next, has 16.15 GHz). Between 6.557 GHz and 13.11 GHz only the TE_{10} mode travels (Figure 23.13). Exercise P23.4 shows that the TM modes start higher still.

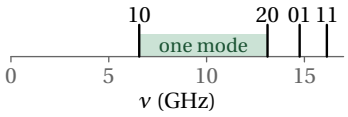


Figure 23.13 The cutoff frequencies of the four lowest TE modes of the WR-90 guide, labeled mn. In the shaded band only the TE_{10} mode travels.

(b) At 10.0 GHz, $\nu_{10}/\nu = 0.6557$, and by (23.10)

$$k = \frac{\omega}{c}\sqrt{1-\left(\frac{\nu_{10}}{\nu}\right)^2} = \frac{2\pi(10.0\times10^9\,\text{Hz})}{2.998\times10^8\,\text{m/s}}\sqrt{1-0.6557^2}$$
$$= (209.6\,\text{m}^{-1})(0.7550) = 158.2\,\text{m}^{-1},$$

so $\lambda_g = 2\pi/k = 39.7\,\text{mm}$, while in empty space $\lambda_0 = c/\nu = 30.0\,\text{mm}$. The wavelength in the guide is longer, by the factor $1/0.7550 = 1.32$. The phase velocity is

$$v_p = \frac{\omega}{k} = \frac{c}{0.7550} = 1.32\,c.$$

That is faster than light; Section 23.4 explains why that is not a problem.

(c) At 5.0 GHz, below ν_{10}, (23.15) gives

$$\kappa = \frac{2\pi}{c}\sqrt{\nu_{10}^2 - \nu^2} = \frac{2\pi}{2.998\times10^8\,\text{m/s}}\sqrt{(6.557\times10^9\,\text{Hz})^2 - (5.0\times10^9\,\text{Hz})^2}$$
$$= 88.9\,\text{m}^{-1},$$

[3] The name comes from the American (EIA) standard for waveguide sizes, in which the 90 gives the inside width in hundredths of an inch, 0.900 in; the inside height is 0.400 in. The international standard IEC 60153-2 calls the same guide R100. Dimensions and designations from Flann Microwave Ltd, *Rigid Rectangular Waveguide Data*, technical data sheet (undated), https://flann.com/wp-content/uploads/2025/01/Waveguide-Parameters.pdf.

so the field falls by $1/e$ in $1/\kappa = 11.2\,\text{mm}$. Over 10 cm of guide it falls by the factor $e^{-8.9} \approx 1.4 \times 10^{-4}$: a 5 GHz signal does not get through.

23.4 Phase velocity and group velocity

Example 23.1 found that the crests of a wave in a guide travel faster than light. That does not contradict relativity, because nothing travels with the crests: neither energy nor information moves along with them. Figure 23.11 shows how they can move so fast. Each plane wave's crests move at c along its own direction, at the angle θ to the axis, and the place where a crest meets a wall moves along the wall at $c/\cos\theta$, which is more than c. In the same way, where an ocean wave meets a straight beach at a small angle, the place where it breaks runs along the beach faster than the wave itself moves. The energy, on the other hand, travels with each plane wave at c along its own direction, so it moves down the guide at only $c\cos\theta$.

23.4.1 Beats and the group velocity

The speed of a signal, a change in a wave that carries information, can be seen in the simplest example: two waves of nearby frequencies ω_1 and ω_2, with wave numbers k_1 and k_2, that travel together. By the identity $\cos A + \cos B = 2\cos[(A-B)/2]\cos[(A+B)/2]$, their sum is

$$\cos(k_1 z - \omega_1 t) + \cos(k_2 z - \omega_2 t) = 2\cos\left(\frac{\Delta k\, z - \Delta\omega\, t}{2}\right)\cos\left(\bar{k}z - \bar{\omega}t\right), \qquad (23.16)$$

where $\bar{k}$ and $\bar{\omega}$ are the averages, and $\Delta k = k_2 - k_1$ and $\Delta\omega = \omega_2 - \omega_1$ the differences. The second factor is a wave at the average frequency, which travels at its phase velocity $\bar{\omega}/\bar{k}$. The first factor varies slowly with z and t. It is an envelope that groups the fast oscillations into bunches, called **beats** (Figure 23.14), and it travels at $\Delta\omega/\Delta k$. As the two frequencies approach each other, this speed becomes the derivative

$$\boxed{v_g = \frac{d\omega}{dk}}, \qquad (23.17)$$

the **group velocity**. A pulse, made of many waves whose frequencies are spread around a central one, travels in the same way at the group velocity at its central frequency, as long as the group velocity hardly changes over the frequencies the pulse contains. Where it does change over them, different parts of the pulse travel at different speeds, and the pulse spreads out as it travels.

For a mode of a guide, the dispersion relation (23.10) is $\omega^2 = \omega_{mn}^2 + c^2k^2$, so $2\omega\, d\omega = 2c^2 k\, dk$, and

$$v_g = \frac{d\omega}{dk} = \frac{c^2 k}{\omega} = c\sqrt{1 - \frac{\omega_{mn}^2}{\omega^2}}, \qquad (23.18)$$

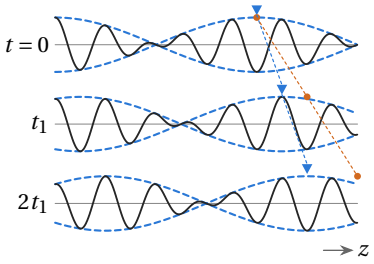


Figure 23.14 Beats: the sum (23.16) of two waves of nearby frequencies (black) and its envelope (blue, dashed), at three instants. Here the phase velocity is twice the group velocity, so the crests (one followed by the orange dot) move twice as fast as the envelope (its peak followed by the blue triangle).

which is less than c, while the phase velocity, $v_p = \omega/k = c/\sqrt{1-\omega_{mn}^2/\omega^2}$, is more than c (Figure 23.15). Their product is

$$v_p v_g = c^2. \tag{23.19}$$

For the TE_{10} mode, with $\sin\theta = \omega_{10}/\omega$ (23.14), the two velocities are $c/\cos\theta$ and $c\cos\theta$, the two speeds found above from the plane waves. At the cutoff frequency the group velocity is zero: the plane waves bounce straight across the guide and make no progress along it. For the guide of Example 23.1 at 10 GHz, $v_g = 0.755\,c$.

The free electrons of Section 22.5.3 have a dispersion relation of the same form. Without damping, (22.24) gives $\varepsilon_r = 1-\omega_p^2/\omega^2$, and $k = n\omega/c$ with $n^2 = \varepsilon_r$ gives $c^2k^2 = \omega^2 - \omega_p^2$, or

$$\omega^2 = \omega_p^2 + c^2k^2. \tag{23.20}$$

So a plasma acts like a waveguide whose cutoff frequency is ω_p. Below ω_p the wave number is imaginary, and the wave is evanescent and reflected, as in Example 22.4; above it, $v_p v_g = c^2$ again.

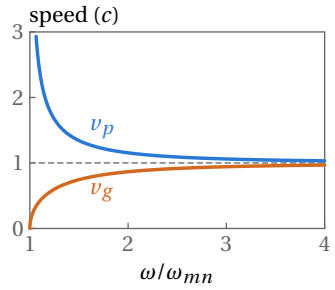


Figure 23.15 The phase velocity v_p and the group velocity v_g of a mode of a hollow guide, in units of c, against the frequency in units of the mode's cutoff frequency. The dashed line is c.

23.4.2 Group velocity in a material

In a transparent material with an index of refraction $n(\omega)$ that depends on the frequency, $k = n(\omega)\,\omega/c$, so $dk/d\omega = (n + \omega\,dn/d\omega)/c$, and

$$v_g = \frac{c}{n + \omega\,dn/d\omega}. \tag{23.21}$$

The denominator is called the **group index**. Where the dispersion is normal, n increases with frequency (Section 22.5.2), so $dn/d\omega > 0$, and a pulse travels more slowly than its crests, at less than c/n.

Inside an absorption line the dispersion is anomalous, $dn/d\omega < 0$, and the denominator of (23.21) can be less than 1, or even negative (Figure 23.16): the group velocity can be greater than c, or negative. This does not mean that signals travel faster than light. There the material absorbs strongly, and a pulse is reshaped so much as it travels that the group velocity no longer describes its motion. The front of any signal never travels faster than c.[4]

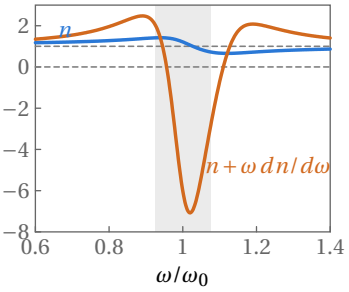


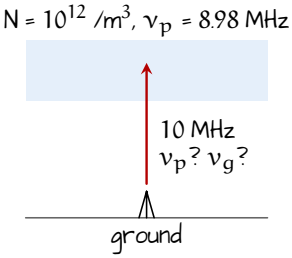
Figure 23.16 The index of refraction n (blue) and the group index (orange) of a material with one absorption line, the material of Figure 22.11. Near the center of the line (gray), where the material absorbs strongly, the group index is less than 1, and even negative.

[4] Arnold Sommerfeld and Léon Brillouin showed this in 1914 for a wave that is switched on suddenly in a material described by the Lorentz model: "Über die Fortpflanzung des Lichtes in dispergierenden Medien," A. Sommerfeld, Ann. Phys. 349, 177–202 (1914), https://doi.org/10.1002/andp.19143491002, and the paper of the same title by L. Brillouin, Ann. Phys. 349, 203–240 (1914). https://doi.org/10.1002/andp.19143491003

Example 23.2

A radio wave of frequency 10.0 MHz travels up into the ionosphere of Example 22.4, where the free electrons have the plasma frequency $\omega_p = 5.64 \times 10^7\ \mathrm{s}^{-1}$. As in that example, ignore collisions. (a) Find the phase velocity and the group velocity of the wave there. (b) Check the group velocity with (23.21).

Solution:



The first step is a sketch of what is given and what is asked: the layer of free electrons with its plasma frequency, and the 10 MHz wave going up into it, whose two velocities we want.

(a) The plasma frequency is $\nu_p = \omega_p/2\pi = 8.98\,\mathrm{MHz}$, so by (23.20)

$$\frac{ck}{\omega} = \sqrt{1 - \frac{\omega_p^2}{\omega^2}} = \sqrt{1 - \left(\frac{8.98\,\mathrm{MHz}}{10.0\,\mathrm{MHz}}\right)^2} = \sqrt{0.194} = 0.440.$$

Then

$$v_p = \frac{\omega}{k} = \frac{c}{0.440} = 2.27\,c, \qquad v_g = \frac{c^2 k}{\omega} = 0.440\,c.$$

(b) Here $n = ck/\omega = \sqrt{1 - \omega_p^2/\omega^2} = 0.440$, and

$$\omega \frac{dn}{d\omega} = \omega \frac{\omega_p^2/\omega^3}{\sqrt{1 - \omega_p^2/\omega^2}} = \frac{\omega_p^2}{\omega^2 n} = \frac{0.806}{0.440} = 1.83,$$

so $v_g = c/(0.440 + 1.83) = c/2.27 = 0.440\,c$, as in (a). In general $n + \omega_p^2/\omega^2 n = (n^2 + \omega_p^2/\omega^2)/n = 1/n$, so $v_g = nc$ for this dispersion relation. The dispersion is normal, since n increases with ω, even though n is less than 1.

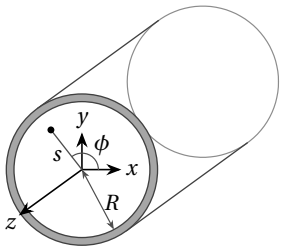


Figure 23.17 A cylindrical waveguide of inside radius R, with the cylindrical coordinates s and ϕ of a point in its cross section and the z axis along its center, toward the viewer.

23.5 Cylindrical waveguides

The other common shape for a waveguide is a pipe of round cross section, with inside radius R. Its natural coordinates are cylindrical ones (Section 1.1.3), with the z axis along the center of the pipe (Figure 23.17). Everything in Section 23.2 still holds: the fields have the form (23.3), E_z and B_z obey Helmholtz equations, and the other components follow from the longitudinal-to-transverse equations. Those equations were written with Cartesian components, but they can be put in a form that works in any coordinates. Write $\nabla_t = \hat{\mathbf{x}}\partial/\partial x + \hat{\mathbf{y}}\partial/\partial y$ for the transverse

part of the gradient, and $\mathbf{E}_t$ and $\mathbf{B}_t$ for the transverse parts of the fields. Then the four equations (23.7) are the x and y components of

$$\begin{aligned}\mathbf{E}_t &= \frac{i}{(\omega/c)^2 - k^2}\left(k\nabla_t E_z - \omega\hat{\mathbf{z}}\times\nabla_t B_z\right),\\ \mathbf{B}_t &= \frac{i}{(\omega/c)^2 - k^2}\left(k\nabla_t B_z + \frac{\omega}{c^2}\hat{\mathbf{z}}\times\nabla_t E_z\right),\end{aligned} \tag{23.22}$$

as you can check by writing out $\hat{\mathbf{z}}\times\nabla_t B_z = \hat{\mathbf{y}}\partial B_z/\partial x - \hat{\mathbf{x}}\partial B_z/\partial y$. In cylindrical coordinates the transverse gradient is $\nabla_t f = \hat{\mathbf{s}}\,\partial f/\partial s + \hat{\boldsymbol{\phi}}\,(1/s)\,\partial f/\partial\phi$, and $\hat{\mathbf{z}}\times\hat{\mathbf{s}} = \hat{\boldsymbol{\phi}}$ and $\hat{\mathbf{z}}\times\hat{\boldsymbol{\phi}} = -\hat{\mathbf{s}}$.

The two-dimensional Laplacian in the Helmholtz equation (23.8) is, in cylindrical coordinates, the left side of (7.49). Written out, the equation for E_z is

$$\frac{\partial^2 E_z}{\partial s^2} + \frac{1}{s}\frac{\partial E_z}{\partial s} + \frac{1}{s^2}\frac{\partial^2 E_z}{\partial\phi^2} + \left(\frac{\omega^2}{c^2} - k^2\right)E_z = 0, \tag{23.23}$$

and B_z obeys the same equation.

23.5.1 TM modes

For a TM mode, $B_z = 0$, and E_z, which is parallel to the wall, must be zero at $s = R$. Separate the variables: $E_z = S(s)\Phi(\phi)$. Substituting this into (23.23), dividing by $S\Phi$, and multiplying by s^2 gives

$$s^2\frac{S''}{S} + s\frac{S'}{S} + s^2\left(\frac{\omega^2}{c^2} - k^2\right) = -\frac{\Phi''}{\Phi}.$$

The left side depends only on s and the right side only on ϕ, so both are equal to a constant, which we call α^2. The ϕ equation, $\Phi'' = -\alpha^2\Phi$, has the solutions $\cos\alpha\phi$ and $\sin\alpha\phi$. Since ϕ and $\phi + 2\pi$ are the same point, Φ must repeat itself after 2π (Figure 23.18), and that requires α to be an integer, $\alpha = 0, 1, 2, \ldots$. A sine is a cosine shifted in angle, so by choosing where $\phi = 0$ we can take $\Phi = \cos\alpha\phi$. (For $\alpha \geq 1$ the solution with $\sin\alpha\phi$ is the same pattern turned through the angle $\pi/2\alpha$; it is a separate mode, with the same cutoff frequency.)

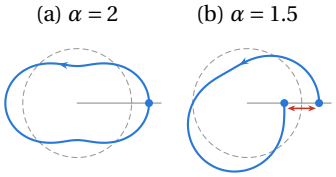


Figure 23.18 $\Phi = \cos\alpha\phi$ drawn around a circle: the curve is outside the dashed circle where $\Phi > 0$ and inside it where $\Phi < 0$. (a) For $\alpha = 2$, after one turn Φ comes back to the value it started with. (b) For $\alpha = 1.5$ it does not: it would have two values at the same point (red), which cannot be.

The s equation is

$$s^2 S'' + sS' + \left[s^2\left(\frac{\omega^2}{c^2} - k^2\right) - \alpha^2\right]S = 0.$$

The quantity $\omega^2/c^2 - k^2$ has units of $1/\text{length}^2$, so multiplying it by R^2 gives a dimensionless number, which we call u^2:

$$u = R\sqrt{\frac{\omega^2}{c^2} - k^2}.$$

(The name will make sense in a moment.) In terms of the dimensionless variable $x = us/R$, which is not the x coordinate, the s equation becomes $x^2\,d^2S/dx^2 +$

$x\,dS/dx + (x^2 - \alpha^2)S = 0$. That is Bessel's equation (6.19). Its solutions are $J_\alpha(x)$ and $Y_\alpha(x)$, and since Y_α is infinite at $x = 0$ (Figure 6.10), on the axis of the guide, only J_α is allowed: $S = J_\alpha(us/R)$.

At the wall, $E_z = 0$ for every ϕ, so $J_\alpha(u) = 0$: the number u must be one of the zeros of J_α, $u_{\alpha m}$ in the notation of Section 6.3. So the TM modes of a round guide are

$$\boxed{E_z(s,\phi) = E_0 J_\alpha\left(\frac{u_{\alpha m} s}{R}\right)\cos\alpha\phi}, \tag{23.24}$$

times $e^{i(kz-\omega t)}$. Since $u_{\alpha m}^2 = R^2(\omega^2/c^2 - k^2)$, each mode has the same dispersion relation as the modes of a rectangular guide, (23.10), with the cutoff frequency

$$\boxed{\omega_{\alpha m} = \frac{u_{\alpha m} c}{R}}. \tag{23.25}$$

Each pair of integers, $\alpha = 0, 1, 2, \ldots$ and $m = 1, 2, 3, \ldots$, gives a mode, called $\mathrm{TM}_{\alpha m}$. In the cross section, E_z has α nodal lines through the center, where $\cos\alpha\phi = 0$, and m nodal circles, counting the wall (Figure 23.19). The transverse fields then follow from (23.22) (Exercise P23.14).

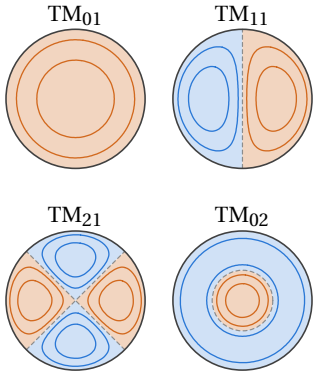


Figure 23.19 E_z of four TM modes of a round guide, (23.24), in its cross section: orange where E_z is positive and blue where it is negative, with the nodes dashed. The mode numbers count the nodal lines through the center (α) and the nodal circles, including the wall (m).

mode	$u_{\alpha m}$
TM_{01}	2.405
TM_{11}	3.832
TM_{21}	5.136
TM_{02}	5.520
TM_{31}	6.380
TM_{12}	7.016
TM_{41}	7.588
TM_{22}	8.417
TM_{03}	8.654
TM_{51}	8.771

Table 23.1 The ten lowest TM modes of a round guide, in order. The cutoff frequency of each is $u_{\alpha m}c/R$.

The mode with the lowest cutoff is TM_{01}, with $u_{01} = 2.405$. It does not depend on ϕ: its field $E_z = E_0 J_0(2.405\,s/R)$ is largest on the axis and falls to zero at the wall. The other modes follow in an order that their labels do not show (Table 23.1). The cutoffs grow with α as well as with m, so TM_{21} comes before TM_{02}, and TM_{41} before TM_{22}. In particular, the sixteen modes with α from 0 to 3 and m from 1 to 4 are not the sixteen lowest: TM_{41}, with $u_{41} = 7.588$, has a lower cutoff than ten of them. To list the lowest modes of a round guide, sort the zeros $u_{\alpha m}$ themselves.

The TE modes come from B_z in the same way. Their boundary condition is the one on the magnetic field, $B^\perp = 0$, and it leads to a different set of numbers in place of the zeros $u_{\alpha m}$ (Exercise P23.13).

23.6 Numerical methods: modes of a guide of any shape

Only a few shapes, such as rectangles and circles, give the Helmholtz equation solutions by separation of variables. For any other shape it can be solved on a grid, in the same way that Section 7.3 solved Laplace's equation. Cover the cross section with a square grid of spacing h, and replace each second derivative at a grid point by differences of the values at the neighboring points. For Laplace's equation that gave (7.10): the value at each point is the average of the values at its four neighbors. For E_z (or B_z) at the point (i, j), the Helmholtz equation (23.8) becomes

$$\frac{E_{i+1,j} + E_{i-1,j} + E_{i,j+1} + E_{i,j-1} - 4E_{i,j}}{h^2} + \left(\frac{\omega^2}{c^2} - k^2\right)E_{i,j} = 0.$$

For every mode, $\omega^2/c^2 - k^2 = \omega_c^2/c^2$ by the dispersion relation, where ω_c is the mode's cutoff frequency. So

$$4E_{i,j} - \left(E_{i+1,j} + E_{i-1,j} + E_{i,j+1} + E_{i,j-1}\right) = \lambda E_{i,j}, \qquad \lambda = \left(\frac{\omega_c h}{c}\right)^2. \tag{23.26}$$

This equation is not solved the way Laplace's equation was, by fixing the values on the boundary and letting the values inside relax. For a TM mode the boundary values are all zero, and with zero on the boundary the only solution of Laplace's equation is zero everywhere. Here the equation has nonzero solutions, but only for certain values of λ. Written out for all N points of the grid inside the guide at once, (23.26) is a matrix equation, $A\mathsf{e} = \lambda\mathsf{e}$, where e is the column of the N values of E_z and A is the $N \times N$ matrix whose rows are the left side of (23.26). This is an **eigenvalue problem**: each eigenvector of A is the pattern of a mode, and its eigenvalue λ gives the mode's cutoff frequency. The matrix is large, but each of its rows has at most five entries that are not zero, and computer programs written for such matrices find their smallest eigenvalues quickly, even for many thousands of points.

The boundary condition decides which neighbors appear in each row (Figure 23.20). For a TM mode, $E_z = 0$ at the walls, so the values at grid points on the walls are zero. For a TE mode, the condition is that the derivative of B_z across the wall is zero (as Tutorial 23.1 found for a rectangle): a value just beyond a wall is taken equal to its mirror image just inside, so that there is no difference across the wall. The TE problem then has one eigenvalue $\lambda = 0$, belonging to a uniform B_z, which Section 23.3.1 ruled out; it is discarded. By (23.22), the transverse electric field of a TE mode is perpendicular to $\nabla_t B_z$, so its field lines are the contours of B_z, which makes the computed modes easy to draw.

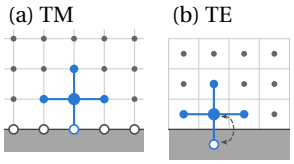


Figure 23.20 The grid near a wall (gray). Each row of (23.26) ties the value at a point (large dot) to its four neighbors (blue). (a) For a TM mode the values at the grid points on the wall (open circles) are zero. (b) For a TE mode, with one value at the center of each square as in Example 23.3, the value beyond the wall (open circle) is set equal to its mirror image just inside, here the point itself.

Before using a program on a new shape, test it on a shape whose answer is known. For the WR-90 guide of Example 23.1, with 80 grid squares across its height, the TE program described here gives the cutoff frequencies 6.557, 13.114, and 14.753 GHz, against the exact values 6.557, 13.114, and 14.754 GHz.

A shape that has no formula, and that engineers use, is the **ridge waveguide**: a rectangular guide with a metal ridge along the middle of one of its broad walls (Figure 23.21). Its lowest mode is like TE_{10}. At the cutoff frequency, where it does not travel along the guide, it oscillates across it, with its electric field running from one broad wall to the other, strongest at the middle, and its magnetic field strongest near the side walls. The middle acts like a capacitor and the sides like an inductor, as in a resonant circuit whose frequency is $1/\sqrt{LC}$. The ridge narrows the gap at the middle, which raises the capacitance and lowers the cutoff frequency. In the next mode, like TE_{20}, the electric field is zero at the middle, so the ridge changes it much less. A ridge therefore widens the range of frequencies in which only one mode travels.[5]

[5] "Properties of Ridge Wave Guide," S. B. Cohn, Proc. IRE 35, 783–788 (1947). https://doi.org/10.1109/JRPROC.1947.226277

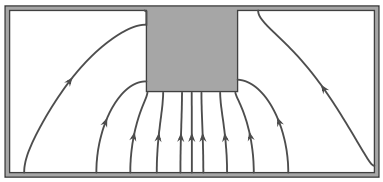


Figure 23.21 The lowest mode of the ridge guide of Example 23.3, computed on the grid: the electric field lines (black) in the cross section, which are contours of B_z, crowd into the gap below the ridge (gray).

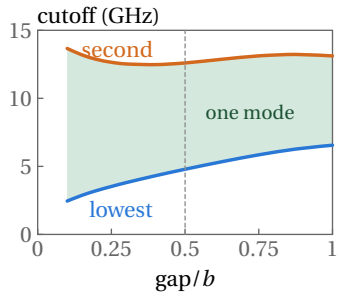
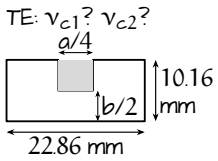


Figure 23.22 The cutoff frequencies of the two lowest modes of the ridge guide of Example 23.3, against the height of the gap below the ridge, in units of b. Without a ridge (gap b) they are those of the plain WR-90 guide. Only one mode travels between the two curves.

Example 23.3

A ridge guide has the inside dimensions of the WR-90 guide of Example 23.1, $a = 22.86\,\text{mm}$ and $b = 10.16\,\text{mm}$, and a ridge $a/4$ wide down the middle of its top wall that fills half of its height, leaving a gap $b/2$ below it. Find the cutoff frequencies of its two lowest TE modes numerically. Over what range of frequencies does only one mode travel, and how does that range compare with the plain guide's?

Solution:



The first step is a sketch of what is given and what is asked: the cross section with the ridge and its dimensions, and the two lowest cutoff frequencies we want.

Choose the grid spacing $h = b/80 = 0.127\,\text{mm}$, so that the walls and the edges of the ridge all fall on grid lines: the cross section is 180 squares by 80, and the ridge is 45 squares wide and 40 high. B_z gets one value at the center of each of the 12 600 squares that the ridge does not fill, and (23.26) with the TE boundary condition gives a $12\,600 \times 12\,600$ matrix. Its smallest eigenvalue is zero (the uniform B_z, discarded), and the next two give

$$\nu_{c1} = 4.79\,\text{GHz}, \qquad \nu_{c2} = 12.6\,\text{GHz}.$$

With half the grid spacing they change by less than 0.05%. So only one mode travels from 4.79 to 12.6 GHz, a range whose top is 2.63 times its bottom, against 2.00 for the plain guide (6.56 to 13.1 GHz). The ridge lowered the first cutoff by 27% and the second by only 4%. Figure 23.21 shows the computed lowest mode. Making the ridge deeper lowers the first cutoff further while the second hardly moves (Figure 23.22).

Tips for Mathematica and Python. In Mathematica, `NDEigensystem` finds the eigenvalues and eigenfunctions of $-\nabla^2$ on a region that you describe, such as a rectangle with a smaller rectangle removed (`RegionDifference`). Add a `DirichletCondition` for the TM modes; with no condition given, it uses the zero-derivative condition of the TE modes. In Python, build the matrix of (23.26) as a sparse matrix (`scipy.sparse`) and find its smallest eigenvalues with `eigsh`, from `scipy.sparse.linalg`. Asking for the eigenvalues nearest a small negative number (its `sigma` option) is fast and avoids trouble with the zero eigenvalue of the TE problem.

23.7 Resonant cavities

A waveguide closed off by conducting walls at both ends is a **resonant cavity**. Put the end walls at $z = 0$ and $z = d$. A wave traveling along $+z$ reflects from the wall

at $z = d$ and travels back, and the two waves together form a standing wave, as light does in front of the mirror of Section 21.6.1.

23.7.1 The resonant frequencies

Take a TM mode of the guide, $E_z \propto e^{i(kz-\omega t)}$, and add the same mode traveling the other way, with k replaced by $-k$, so that $E_z \propto (e^{ikz} + e^{-ikz})e^{-i\omega t} = 2\cos kz\, e^{-i\omega t}$. By (23.22), the transverse electric field is proportional to k, so it changes sign for the wave traveling the other way, and the sum has $\mathbf{E}_t \propto (e^{ikz} - e^{-ikz}) \propto \sin kz$. At the end walls the transverse field is parallel to the walls, so it must be zero at $z = 0$ and $z = d$ (Figure 23.23). It is zero at $z = 0$, and at $z = d$ only if $\sin kd = 0$:

$$k = \frac{\ell\pi}{d}, \qquad \ell = 0, 1, 2, \ldots.$$

For a TE mode, B_z is perpendicular to the end walls and must be zero there. The same argument gives $B_z \propto \sin kz$ and the same values of k, except that $\ell = 0$ is not allowed, since it would make B_z zero everywhere.

Each mode of the guide has $\omega^2 = \omega_c^2 + c^2k^2$, where ω_c is its cutoff frequency. So a cavity can hold that mode only at the frequencies

$$\boxed{\omega = \sqrt{\omega_c^2 + \left(\frac{\ell\pi c}{d}\right)^2}}. \tag{23.27}$$

A guide carries a mode at every frequency above its cutoff, but a cavity holds it only at these discrete **resonant frequencies**. Each mode of a cavity has three mode numbers, such as $\mathrm{TE}_{mn\ell}$ for a rectangular box or $\mathrm{TM}_{\alpha m\ell}$ for a cylinder.

The standing wave between two parallel mirrors (Figure 21.19) is the one-dimensional case: light traveling straight between mirrors a distance d apart has $\omega_c = 0$, and (23.27) gives $\omega = \ell\pi c/d$, the frequencies at which a whole number of half wavelengths fits between the mirrors. A laser's mirrors form such a cavity.

The simplest modes of all are the TM modes with $\ell = 0$. Then $k = 0$, nothing depends on z, and the resonant frequency is the cutoff frequency of the guide mode. At its cutoff frequency a mode of a guide does not travel along it, but bounces straight across it (Section 23.3.2), so end walls can be added anywhere.

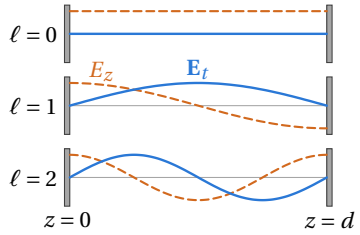
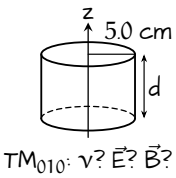


Figure 23.23 Standing waves between the end walls of a cavity (gray), for $\ell = 0$, 1 and 2. The transverse electric field (blue), proportional to $\sin(\ell\pi z/d)$, is zero at the walls; E_z of a TM mode (orange, dashed), proportional to $\cos(\ell\pi z/d)$, is largest there.

Example 23.4

A closed cylindrical can with conducting walls has inside radius $R = 5.0\,\mathrm{cm}$ and height d. (a) Find the frequency ν of its TM_{010} mode, and the electric and magnetic fields of the mode. (b) Where in the cavity are the electric and magnetic fields largest, and when? (c) Compare the fields with those of the capacitor of Example 16.3.

Solution:



The first step is a sketch of what is given and what is asked: the can with its radius and height, its axis along z, and the frequency and fields of the TM_{010} mode that we want.

(a) The mode is the TM_{01} mode of a round guide with $\ell = 0$, so $k = 0$. By (23.24), $\tilde{E}_z = E_0 J_0(u_{01} s/R)\, e^{-i\omega t}$, and by (23.27) and (23.25), $\omega = \omega_{01} = u_{01} c/R$:

$$\nu = \frac{u_{01} c}{2\pi R} = \frac{(2.405)(2.998 \times 10^8\,\text{m/s})}{2\pi(0.050\,\text{m})} = 2.29\,\text{GHz}.$$

The transverse fields come from (23.22) with $k = 0$, $B_z = 0$, and $(\omega/c)^2 - k^2 = (u_{01}/R)^2$. The electric field has no transverse part, since $\mathbf{E}_t$ is proportional to k. For the magnetic field, $\hat{\mathbf{z}} \times \nabla_t E_z = \hat{\boldsymbol{\phi}}\, \partial E_z/\partial s$, and $J_0' = -J_1$ (Section 6.3), so

$$\tilde{\mathbf{B}}_t = \frac{iR^2}{u_{01}^2}\frac{\omega}{c^2}\frac{u_{01}}{R} E_0 J_0'\Big(\frac{u_{01} s}{R}\Big) e^{-i\omega t}\,\hat{\boldsymbol{\phi}} = -\frac{i}{c} E_0 J_1\Big(\frac{u_{01} s}{R}\Big) e^{-i\omega t}\,\hat{\boldsymbol{\phi}},$$

where $\omega R/(c u_{01}) = 1$. For a real E_0 the real fields are

$$\mathbf{E} = E_0 J_0\Big(\frac{2.405\, s}{R}\Big)\cos\omega t\,\hat{\mathbf{z}}, \qquad \mathbf{B} = -\frac{E_0}{c} J_1\Big(\frac{2.405\, s}{R}\Big)\sin\omega t\,\hat{\boldsymbol{\phi}}.$$

They meet every boundary condition: $\mathbf{E}$ is perpendicular to the end walls and zero at the side wall, where $J_0(2.405) = 0$, and $\mathbf{B}$ runs parallel to every wall. Nothing depends on the height d.

(b) The electric field points along the axis. It is largest on the axis, where $J_0 = 1$, and falls to zero at the wall. The magnetic field circles the axis. It is zero on the axis and largest about three quarters of the way out to the wall, where J_1 has its maximum. The two are 90° out of phase: when the electric field is largest the magnetic field is zero, and a quarter of a period later the energy is all in the magnetic field (Figure 23.24). The energy moves back and forth between the electric field near the axis and the magnetic field near the wall, as it does between the capacitor and the inductor of a resonant circuit.

(c) Example 16.3 found the field between the plates of a capacitor at high frequency as a series in $x = \omega s/c$, which is $J_0(x)$, and at $\omega = 2.405\, c/R$ the field at the edge of the plates is zero. Its magnetic field $\mathbf{B}_1 = -(\omega s/2c^2) E_0 \sin\omega t\,\hat{\boldsymbol{\phi}}$ is the first term of the series for $\mathbf{B}$ above, since $J_1(x) \approx x/2$ for small x. So the TM_{010} mode is that capacitor at its resonant frequency: its plates are the end walls, and a conducting cylinder at the edge of the plates closes it off. At 2.29 GHz this cavity resonates near the frequency of a microwave oven.

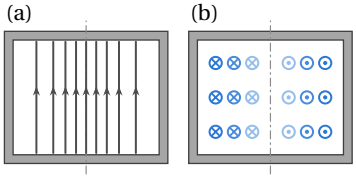


Figure 23.24 The TM_{010} mode of Example 23.4 in a cut through the axis of the cavity. (a) When the electric field (black) is largest; its lines are spaced to show its strength. (b) A quarter of a period later, when the magnetic field (blue) is largest; it circles the axis, out of the page on one side and into it on the other, each symbol shaded in proportion to the field.

The TE_{011} mode of a cylinder is another common mode. Its electric field circles the axis and is zero on it, while its magnetic field is strongest on the

axis. Cavities for electron spin resonance, which use the magnetic field of the microwaves to flip the spins of electrons in a sample, are often built to work in this mode (Figure 23.25), with the sample on the axis.

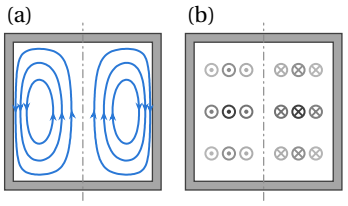


Figure 23.25 The TE_{011} mode of a cylindrical cavity in a cut through its axis. (a) When the magnetic field (blue) is largest: its lines form loops in the plane of the cut. (b) A quarter of a period later, when the electric field is largest: it circles the axis, out of the page on one side and into it on the other, and is zero on the axis.

23.7.2 Losses and the quality factor

Real walls are not perfect conductors. The currents that the fields require flow in a layer about a skin depth thick (Section 22.2.2), and each cycle they turn a small part of the stored energy into heat, so the oscillation dies away unless something feeds it. The **quality factor** Q of a cavity measures how slowly:

$$Q = \omega \frac{\text{energy stored}}{\text{power lost}}. \tag{23.28}$$

With words, the fraction of the stored energy lost in each cycle is $2\pi/Q$. A cavity driven near one of its resonant frequencies responds strongly only over a range of frequencies about ω/Q wide, so a large Q means a sharp resonance. The losses happen within a skin depth of the walls, while the energy fills the whole volume, so Q is of the order of the size of the cavity divided by the skin depth. For the cavity of Example 23.4 with copper walls and $d = R$, $Q = 1.8\times10^4$ (Exercise P23.18). Silver conducts a little better than copper, and the insides of some cavities are plated with it (Exercise P22.5).

23.8 Transmission lines

A hollow guide carries nothing below its lowest cutoff frequency, which for a guide a few centimeters across is in the gigahertz range. Signals of lower frequencies, all the way down to the steady current from a battery, travel on lines made of two separate conductors, such as a coaxial cable or a pair of parallel wires. Two conductors allow a TEM mode (Section 23.2.2), and that is what such a **transmission line** carries.

23.8.1 The TEM mode

With $E_z = B_z = 0$ the longitudinal-to-transverse equations do not apply (both sides are 0/0), so go back to the component equations (23.4)–(23.5). With $E_z = B_z = 0$, the left-hand equation of (23.5) gives $B_y = (k/\omega)E_x$ and the right-hand equation of (23.4) gives $B_y = (\omega/c^2k)E_x$. Both can hold only if $k^2 = \omega^2/c^2$. For a wave traveling along $+z$, $k = \omega/c$, and the four equations together say

$$\mathbf{B} = \frac{1}{c}\hat{\mathbf{z}}\times\mathbf{E}. \tag{23.29}$$

With words, a TEM wave travels at c whatever its frequency, with no cutoff and no dispersion, and its fields are related as in a plane wave (21.10). (In a line filled with a material of index n, it travels at c/n, as in Exercise P23.3.) The transverse electric field has no curl in the cross section (the left-hand equation of (23.6) with $B_z = 0$)

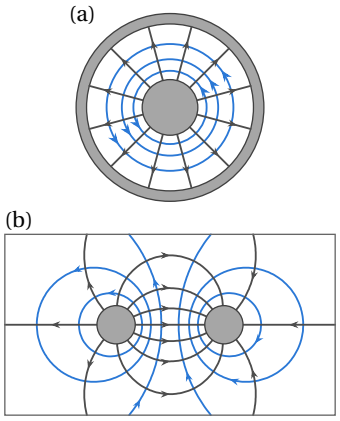


Figure 23.26 The electric (black) and magnetic (blue) fields of the TEM mode in the cross section of (a) a coaxial cable and (b) a pair of parallel wires, at one instant. The electric field lines run from one conductor to the other, and the magnetic field lines circle the conductors.

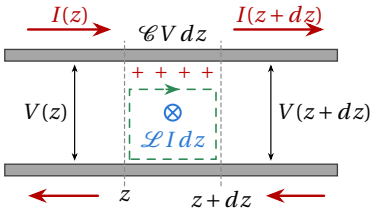


Figure 23.27 A short piece of a transmission line, between z and $z+dz$. The current I flows along the top conductor and back along the bottom one; the charge $\mathscr{C}V\,dz$ sits on the top conductor; and the magnetic flux $\mathscr{L}I\,dz$ passes through the loop (dashed).

and no divergence, so, as in the proof of Section 23.2.2, it is a two-dimensional gradient field, $\mathbf{E}_t = -\nabla_t V$, whose potential obeys Laplace's equation in the cross section and is constant on each conductor. So at each instant, and at each z, the electric field in the cross section is the electrostatic field of the two conductors charged to some voltage, and the magnetic field, perpendicular to it by (23.29), is the magnetostatic field of equal and opposite currents along them (Figure 23.26).

23.8.2 The telegrapher's equations

A line is usually described not by its fields but by the voltage $V(z,t)$ between its conductors and the current $I(z,t)$ along them, I in one conductor and $-I$ in the other (Figure 23.27). The voltage is the integral of $\mathbf{E}$ across the cross section from one conductor to the other, which does not depend on the path, since $\mathbf{E}_t$ is a gradient field there. Let $\mathscr{C}$ be the capacitance per unit length of the line and $\mathscr{L}$ its inductance per unit length ($\mathscr{C}$ and $\mathscr{L}$ are calligraphic capital letters C and L). For a coaxial cable they are found in Exercises P5.13 and P12.21.

Take a short piece of the line, between z and $z+dz$. First, the charge on its top conductor, $\mathscr{C}V\,dz$, changes only as current flows in at z and out at $z+dz$: $\partial(\mathscr{C}V\,dz)/\partial t = I(z) - I(z+dz)$. Second, take the loop that runs along the top conductor from z to $z+dz$, across to the bottom one, back along it, and across again. Along the perfect conductors $\mathbf{E}^{\|} = 0$, so the circulation of $\mathbf{E}$ around the loop is $V(z+dz) - V(z)$, and by Faraday's law it equals the negative of the rate of change of the flux through the loop, $\mathscr{L}I\,dz$. Dividing by dz,

$$\boxed{\frac{\partial V}{\partial z} = -\mathscr{L}\frac{\partial I}{\partial t}, \qquad \frac{\partial I}{\partial z} = -\mathscr{C}\frac{\partial V}{\partial t}}. \tag{23.30}$$

These are the **telegrapher's equations**. Oliver Heaviside, who worked out the theory of transmission lines (his portrait is in Section 2.4), also patented the coaxial cable, in 1880.

Taking the derivative of the first equation with respect to z and using the second gives $\partial^2 V/\partial z^2 = \mathscr{L}\mathscr{C}\,\partial^2 V/\partial t^2$, the wave equation (16.13) with the speed

$$v = \frac{1}{\sqrt{\mathscr{L}\mathscr{C}}}.$$

For any line this is c/n, the speed of the TEM wave, where n is the index of the material between the conductors (Exercise P23.19).

For a wave traveling along $+z$, $V = f(z-vt)$ for some function f, and the second telegrapher's equation gives $\partial I/\partial z = \mathscr{C}vf'$, so $I = \mathscr{C}vV = V/Z_c$, where

$$\boxed{Z_c = \sqrt{\frac{\mathscr{L}}{\mathscr{C}}}} \tag{23.31}$$

is the **characteristic impedance** of the line. (The symbol Z_0 is already taken by the impedance of free space, Section 21.4.1.) For a wave traveling along $-z$,

$I = -V/Z_c$. With words, a wave traveling one way along a line has a fixed ratio of voltage to current, set by the line's geometry and not by what is connected to its far end: to the source that launches it, the line looks like a resistor Z_c. The power the wave carries past a point is $VI = V^2/Z_c$.

For a coaxial cable with inner radius a, outer radius b, and a material of permittivity ε between the conductors, $\mathscr{C} = 2\pi\varepsilon/\ln(b/a)$ and $\mathscr{L} = (\mu_0/2\pi)\ln(b/a)$. (At these frequencies the currents flow on the surfaces of the conductors, so the magnetic field is only between them, as in part (b) of Exercise P12.21.) Then $v = 1/\sqrt{\mu_0\varepsilon} = c/n$, and

$$Z_c = \frac{1}{2\pi}\sqrt{\frac{\mu_0}{\varepsilon}}\ln\frac{b}{a} = \frac{Z_0}{2\pi n}\ln\frac{b}{a}. \tag{23.32}$$

The impedance of free space, 376.73 Ω, sets the scale, and the dimensions enter only through a logarithm, which changes slowly: for ratios b/a from 2 to 10 and insulators with n from 1 to 1.5, Z_c ranges only from about 28 to 140 Ω. Laboratory cables are usually made with $Z_c = 50\,\Omega$.

23.8.3 Pulses and reflections

Let a line end, at $z = 0$, in a resistor R_L connected across its two conductors, and let a pulse of any shape, $V_I f(t - z/v)$, arrive from the left. At the resistor, $V = R_L I$ at every instant, while the incident pulse alone has $V/I = Z_c$. Unless $R_L = Z_c$, a reflected pulse $V_R g(t + z/v)$ must travel back to the left, with $V/I = -Z_c$. At $z = 0$ the total voltage and current are $V = V_I f + V_R g$ and $I = (V_I f - V_R g)/Z_c$, and $V = R_L I$ at every instant requires $g = f$ and

$$r = \frac{V_R}{V_I} = \frac{R_L - Z_c}{R_L + Z_c}. \tag{23.33}$$

So the reflected pulse has the same shape as the incident one. The reflection coefficient is that of light at normal incidence, (21.16) with $\beta = Z_1/Z_2$, with the line's Z_c in place of Z_1 and the load's R_L in place of Z_2, and the algebra is the same as for a wave on a string at a knot (Section 21.2.1). Three cases stand out (Figure 23.28). At an open end, $R_L \to \infty$, $r = +1$: the voltage pulse comes back upright, and its current, $-V/Z_c$, cancels the incident current at the end, where no current can flow. At a short circuit, $R_L = 0$, $r = -1$: the voltage pulse comes back inverted, and the voltage at the end stays zero. And with $R_L = Z_c$, a **matched load**, $r = 0$: the resistor absorbs the whole pulse, just as more line would. Lines are ended in matched loads so that signals are not reflected back and forth along them. An antenna at the end of a line is a load equal to its radiation resistance (Section 28.2.3), and a line matched to it delivers all of the power to the antenna (Exercise P23.21).

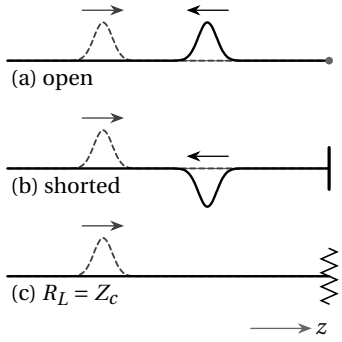


Figure 23.28 A voltage pulse on a transmission line (black), before (dashed) and after (solid) it reaches the end of the line at the right, where (a) the line is open, (b) it is shorted, and (c) a resistor $R_L = Z_c$ is connected across it.

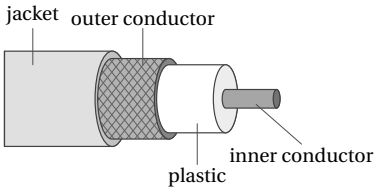
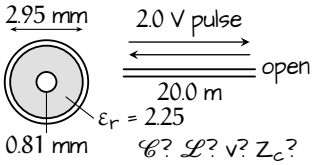


Figure 23.29 A coaxial cable cut away to show its layers: the inner conductor, the plastic between the conductors, the outer conductor (often a braid of fine wires), and a protective jacket of plastic.

Example 23.5

A coaxial cable (Figure 23.29) has an inner conductor 0.81 mm in diameter and an outer conductor whose inside diameter is 2.95 mm, with plastic of relative permittivity $\varepsilon_r = 2.25$ between them. (a) Find its capacitance and inductance per unit length, the speed of signals along it, and its characteristic impedance. (b) A pulse of 2.0 V is sent into one end of 20.0 m of this cable, whose far end is left open. What current does the pulse carry? When does its reflection get back to the start, and what are its voltage and current? (c) What fraction of the pulse's voltage comes back if a 50.0 Ω resistor is connected across the far end?

Solution:



The first step is a sketch of what is given and what is asked: the cross section of the cable with its two diameters and the plastic between them, and the cable's length with the pulse going out and coming back from the open end.

(a) The radii are $a = 0.405\,\text{mm}$ and $b = 1.475\,\text{mm}$, so $\ln(b/a) = \ln 3.642 = 1.293$. Then

$$\mathscr{C} = \frac{2\pi\varepsilon_r\varepsilon_0}{\ln(b/a)} = \frac{2\pi(2.25)(8.854\times10^{-12}\,\text{F/m})}{1.293} = 96.8\,\text{pF/m},$$

$$\mathscr{L} = \frac{\mu_0}{2\pi}\ln\frac{b}{a} = (2.000\times10^{-7}\,\text{H/m})(1.293) = 0.259\,\mu\text{H/m}.$$

The speed is $v = 1/\sqrt{\mathscr{L}\mathscr{C}} = 2.00\times10^8\,\text{m/s}$, which is $c/1.5 = 0.667\,c$, as it must be for $n = \sqrt{2.25} = 1.5$, and the characteristic impedance is

$$Z_c = \sqrt{\frac{\mathscr{L}}{\mathscr{C}}} = \frac{376.73\,\Omega}{2\pi(1.5)}(1.293) = 51.7\,\Omega.$$

(b) The pulse carries the current $I = V/Z_c = (2.0\,\text{V})/(51.7\,\Omega) = 38.7\,\text{mA}$. It travels 40.0 m there and back, so its reflection returns after $(40.0\,\text{m})/(2.00\times10^8\,\text{m/s}) = 200\,\text{ns}$. At the open end $r = +1$: the reflection has the voltage +2.0 V and the current $-V/Z_c = -38.7\,\text{mA}$, flowing back toward the start.

(c) By (23.33),

$$r = \frac{50.0\,\Omega - 51.7\,\Omega}{50.0\,\Omega + 51.7\,\Omega} = -0.016:$$

1.6% of the voltage comes back, inverted. A resistor close to Z_c absorbs nearly all of the pulse.

23.9 Summary

Conducting walls confine a wave by requiring $\mathbf{E}^{\parallel} = 0$ and $B^{\perp} = 0$ at every wall. In a hollow pipe only certain patterns, the modes, meet these conditions, and

each travels only above its own cutoff frequency, with a phase velocity above c and a group velocity below it. Closing the pipe at both ends leaves only discrete resonant frequencies, and two separate conductors carry a TEM wave at every frequency.

Perfect conductors and hollow guides

$\mathbf{E}^{\parallel} = 0, \quad B^{\perp} = 0$ — just outside a perfect conductor; the fields require the surface charge $\sigma = \varepsilon_0 \hat{\mathbf{n}} \cdot \mathbf{E}$ and current $\mathbf{K} = \hat{\mathbf{n}} \times \mathbf{B}/\mu_0$

$\tilde{\mathbf{E}} = \tilde{\mathbf{E}}_0(x, y)\, e^{i(kz-\omega t)}$ — a wave traveling along a guide; the same form for $\mathbf{B}$

$\mathbf{E}_t = \dfrac{i}{(\omega/c)^2 - k^2} (k \nabla_t E_z - \omega \hat{\mathbf{z}} \times \nabla_t B_z)$ — the longitudinal-to-transverse equations, with the similar one for $\mathbf{B}_t$: E_z and B_z give everything

$\left(\nabla_t^2 + \dfrac{\omega^2}{c^2} - k^2\right) E_z = 0$ — Helmholtz equation, the same for B_z; $E_z = 0$ at the walls

TE: $E_z = 0$; TM: $B_z = 0$; TEM: both — any field is a sum of TE and TM modes; a hollow guide has no TEM mode

Rectangular and round guides

$B_z = B_0 \cos \dfrac{m\pi x}{a} \cos \dfrac{n\pi y}{b}$ — TE_{mn} modes of an $a \times b$ guide; m and n not both zero

$\omega_{mn} = c\pi \sqrt{\left(\dfrac{m}{a}\right)^2 + \left(\dfrac{n}{b}\right)^2}$ — cutoff frequency; TE_{10}, with $\omega_{10} = \pi c/a$, is the lowest for $a > b$

$k = \dfrac{1}{c} \sqrt{\omega^2 - \omega_c^2}$ — dispersion relation of every mode; below cutoff $k = i\kappa$: the field is evanescent and the wave is reflected

$E_z = E_0 J_\alpha \left(\dfrac{u_{\alpha m} s}{R}\right) \cos \alpha\phi, \quad \omega_{\alpha m} = \dfrac{u_{\alpha m} c}{R}$ — $\text{TM}_{\alpha m}$ modes of a round guide of radius R; $u_{\alpha m}$ is the mth zero of J_α

Phase velocity and group velocity

$v_p = \dfrac{\omega}{k}, \quad v_g = \dfrac{d\omega}{dk}$ — the crests move at v_p; signals and energy at v_g

$v_p v_g = c^2$ — for $\omega^2 = \omega_c^2 + c^2 k^2$: modes of a guide, and a plasma with $\omega_c = \omega_p$

$v_g = \dfrac{c}{n + \omega\, dn/d\omega}$ — in a material; the denominator is the group index

Guides of any shape, and cavities

$4E_{i,j} - \big(E_{i+1,j} + E_{i-1,j} + E_{i,j+1} + E_{i,j-1}\big) = \lambda E_{i,j}$ — the Helmholtz equation on a grid, an eigenvalue problem; $\lambda = (\omega_c h/c)^2$

$\omega = \sqrt{\omega_c^2 + \left(\dfrac{\ell\pi c}{d}\right)^2}$ — discrete resonant frequencies of a guide closed by walls a distance d apart

$Q = \omega \dfrac{\text{energy stored}}{\text{power lost}}$ — quality factor; $2\pi/Q$ of the energy is lost each cycle; resonance width ω/Q

Transmission lines

$\mathbf{B} = \dfrac{1}{c}\hat{\mathbf{z}} \times \mathbf{E}$ — the TEM mode of two conductors: speed c (or c/n) at every frequency; the cross-section fields are electrostatic and magnetostatic

$\dfrac{\partial V}{\partial z} = -\mathscr{L}\dfrac{\partial I}{\partial t}, \quad \dfrac{\partial I}{\partial z} = -\mathscr{C}\dfrac{\partial V}{\partial t}$ — telegrapher's equations; $v = 1/\sqrt{\mathscr{L}\mathscr{C}}$

$Z_c = \sqrt{\dfrac{\mathscr{L}}{\mathscr{C}}}, \quad \text{coaxial: } Z_c = \dfrac{Z_0}{2\pi n}\ln\dfrac{b}{a}$ — characteristic impedance; $V/I = \pm Z_c$ for a wave traveling along $\pm z$

$r = \dfrac{R_L - Z_c}{R_L + Z_c}$ — reflection at a resistive load; none from a matched load, $R_L = Z_c$

Exercises

Exercises for 23.2 Waves in a hollow guide

P23.1 The longitudinal-to-transverse equations. Derive the second, third, and fourth of the longitudinal-to-transverse equations (23.7) from the component equations (23.4)–(23.6), as Section 23.2.1 did for E_x. Which two of the six equations does each one use?

P23.2 TE plus TM. A field in a hollow guide has the form (23.3), with both E_z and B_z nonzero.

(a) Use the longitudinal-to-transverse equations to explain why this field is the sum of two fields with the same k and ω: one with the same E_z and with $B_z = 0$, and one with $E_z = 0$ and the same B_z.

(b) A TE mode and a TM mode of the same guide, at the same frequency ω, generally have different wave numbers. Is their sum a solution of Maxwell's equations in the guide? Is it of the form (23.3)? Explain.

P23.3 A filled guide. A hollow guide is filled with a material with a real permittivity ε and a real permeability μ, both constant.

(a) Show that the results of Section 23.2 still hold, with c replaced by $v = 1/\sqrt{\mu\varepsilon}$.

(b) By what factor does filling the guide change the cutoff frequency of each mode, in terms of the material's index of refraction n? Evaluate the factor for $n = 1.5$.

Exercises for 23.3 Rectangular waveguides

P23.4 TM modes of a rectangular guide. Analyze the TM modes of the rectangular guide of Figure 23.6 in the same way as Tutorial 23.1 did for the TE modes.

(a) What field component governs the behavior of the TM modes?

(b) What assumption do you need to make that leads to the Helmholtz equation for this component?

(c) What boundary condition is important for this component?

(d) Show that the function $f(x, y) = \sin(m\pi x/a)\sin(n\pi y/b)$ satisfies the boundary condition at all four walls. What is the complete function when you include the z and t dependence? (It is enough to do this for the walls at $x = 0$ and $x = a$; the walls at $y = 0$ and $y = b$ are

very similar.) This is easier than the boundary condition for the TE modes, because you should not need the longitudinal-to-transverse equations.

(e) Show that the full function, with its x, y, z, and t dependence, satisfies the Helmholtz equation as long as k and ω obey the same dispersion relation as the TE modes, (23.10), with the same ω_{mn} (23.11), and that it behaves the same way for $\omega < \omega_{mn}$.

(f) Section 23.3.1 showed that there is no TE mode with both $m = 0$ and $n = 0$. One major difference between the TE and TM modes is that there is no TM mode with either $m = 0$ or $n = 0$. Why is that? Which TM mode has the lowest cutoff frequency?

P23.5 Charges and currents in the walls.

(a) Use (23.2) and the TE_{10} fields (23.13) to find the surface charge density σ and the surface current density $\mathbf{K}$ on all four walls of a rectangular guide.

(b) Sketch $\mathbf{K}$ on the inside of the top wall, $y = b$, over one wavelength of the guide at $t = 0$, and mark where σ is largest.

(c) Show that your σ and $\mathbf{K}$ on the top wall satisfy the continuity equation for surface charge, $\nabla \cdot \mathbf{K} = -\partial\sigma/\partial t$, where the divergence is taken in the plane of the wall.

(d) Measurements of the fields inside a guide are often made with a small probe pushed in through a narrow slot cut along the middle of the top wall, parallel to z. Use your $\mathbf{K}$ to explain why such a slot hardly disturbs the TE_{10} mode.

P23.6 The magnetic field of the two plane waves. Section 23.3.2 wrote $\tilde{E}_y$ of the TE_{10} mode as the sum of two plane waves.

(a) Find the magnetic field of each plane wave from its electric field with (21.10), and show that their sum is the TE_{10} magnetic field.

(b) Each plane wave is transverse, yet their sum has a magnetic field along z. Explain how.

P23.7 Designing a guide. A radar transmitter works at frequencies from 2.7 to 3.5 GHz, all of which must travel in a single mode in a rectangular guide with $b = a/2$.

(a) What range of widths a allows this?

(b) To stay away from the cutoffs, each frequency should be at least 10% above the lowest cutoff frequency and at least 10% below the next one. What range of widths allows this?

Exercises for 23.4 Phase velocity and group velocity

P23.8 Energy flow in the TE_{10} mode.

(a) Find the time-averaged Poynting vector of the TE_{10} mode (23.13), and the power P that the mode carries, the integral of the average of S_z over the cross section.

(b) Find the time-averaged energy per unit length of the guide, U, the integral of the average energy density over the cross section.

(c) Show that P/U equals the group velocity (23.18): the energy travels down the guide at v_g.

P23.9 The product of the two velocities.

(a) Show that any dispersion relation of the form $\omega^2 = \omega_c^2 + c^2k^2$, with a constant ω_c, gives $v_p v_g = c^2$.

(b) Show the converse: if $v_p v_g = c^2$ at every k, the dispersion relation has this form.

(c) Light in a material with a constant index n has $v_p = v_g = c/n$. Does it have $v_p v_g = c^2$?

P23.10 Arrival times in a waveguide. Two short bursts of microwaves, one at 9.5 GHz and one at 10.5 GHz, are sent into 50 m of the WR-90 guide of Example 23.1 at the same moment, both in the TE_{10} mode.

(a) How long does each take to arrive, and how much later is that than through 50 m of empty space?

(b) How far apart in time do they arrive? Use your result to explain why a pulse in a waveguide spreads out as it travels.

P23.11 Pulses in glass. Many lasers that make very short pulses of light work near a vacuum wavelength of 800 nm. Malitson's formula for the index of fused silica (Section 22.5.2) is

$$n^2 = 1 + \frac{B_1\lambda_0^2}{\lambda_0^2 - C_1^2} + \frac{B_2\lambda_0^2}{\lambda_0^2 - C_2^2} + \frac{B_3\lambda_0^2}{\lambda_0^2 - C_3^2},$$

where

$$B_1 = 0.6961663, \qquad B_2 = 0.4079426, \qquad B_3 = 0.8974794,$$
$$C_1 = 0.0684043\,\mu\text{m}, \qquad C_2 = 0.1162414\,\mu\text{m}, \qquad C_3 = 9.896161\,\mu\text{m}.$$

(a) Show that $\omega\, dn/d\omega = -\lambda_0\, dn/d\lambda_0$.

(b) Find n and the group index at 800 nm.

(c) How much longer does a pulse take to travel through 1.00 cm of fused silica than its crests would?

P23.12 Beats on a computer. Take the dispersion relation $\omega = \sqrt{\omega_c^2 + c^2k^2}$ with $c = 1$ and $\omega_c = 1$, in arbitrary units.

(a) Plot $\cos(k_1 z - \omega_1 t) + \cos(k_2 z - \omega_2 t)$ against z at several times, for $k_1 = 0.9$ and $k_2 = 1.1$ and the corresponding frequencies. Follow one crest and the peak of one envelope, and measure their speeds. Compare them with $\bar{\omega}/\bar{k}$, $\Delta\omega/\Delta k$, and $d\omega/dk$ at $k = 1$.

(b) Repeat with five waves of equal amplitude, with $k = 0.8$, 0.9, 1.0, 1.1, and 1.2. What changes, and what stays the same?

Exercises for 23.5 Cylindrical waveguides

P23.13 TE modes of a round guide. For the TE modes of the guide of Figure 23.17, $E_z = 0$, and B_z obeys the same Helmholtz equation (23.23) as E_z, so it has the same form, $J_\alpha(us/R)\cos\alpha\phi$, with α an integer and $u = R\sqrt{\omega^2/c^2 - k^2}$, as in Section 23.5.1. Its boundary condition is different.

(a) Use (23.22) to show that $B_s = 0$ at the wall requires $\partial B_z/\partial s = 0$ at $s = R$.

(b) Show that u must then be a zero of J'_α, the derivative of J_α, rather than of J_α. Call the mth zero $u'_{\alpha m}$. What is the cutoff frequency of the $\mathrm{TE}_{\alpha m}$ mode?

(c) Find $u'_{\alpha m}$ for $\alpha = 0$ to 3 and $m = 1$ to 4 with a computer. (In Python, `scipy.special.jnp_zeros` gives them; in Mathematica, `FindRoot` applied to `D[BesselJ[a, x], x]`, started near each place where it changes sign.) Plot B_z of each of these sixteen modes over the cross section of a guide with $R = 10$ cm. (To find ϕ from x and y, use the two-argument inverse tangent of Section 1.1.4; $\arctan(y/x)$ alone gives the wrong angle for points with $x < 0$.) Every plot should have zero radial slope at the wall.

(d) Which mode of a round guide, TE or TM, has the lowest cutoff frequency? Over what range of frequencies, in units of c/R, does only that mode travel?

P23.14 The TM_{01} mode of a round guide.

(a) Use (23.22) to find the transverse electric and magnetic fields of the mode $E_z = E_0 J_0(u_{01}s/R)$. (Use $J'_0 = -J_1$.)

(b) Sketch the electric field lines at one instant in a plane through the axis, over one guide wavelength, and the magnetic field lines in a cross section.

(c) Show that the magnetic field is parallel to the wall, as $B^{\perp} = 0$ requires, and find the surface current density $\mathbf{K}$ on the wall from (23.2). Which way does it flow?

Exercises for 23.6 Numerical methods: modes of a guide of any shape

P23.15 A double-ridge guide. Some guides have two ridges, one along the middle of each broad wall, facing each other across a gap. Take the guide of Example 23.3, but with two ridges, each $a/4$ wide and $b/4$ high (Figure 23.30), so that the gap between them is $b/2$, as in the example.

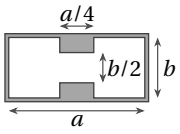


Figure 23.30 The double-ridge guide of Exercise P23.15, in cross section.

(a) Find the cutoff frequencies of its two lowest TE modes, and check that halving the grid spacing hardly changes them.

(b) Compare the range of frequencies in which only one mode travels with that of the single-ridge guide of the example and with that of the plain WR-90 guide.

(c) Plot the electric field lines of the lowest mode in the cross section.

Exercises for 23.7 Resonant cavities

P23.16 A rectangular cavity. A box with perfectly conducting walls has inside dimensions a, b, and d along x, y, and z, with one corner at the origin (Figure 23.31). Try an electric field with the spatial dependence

$$\mathbf{E} = C_1 \cos k_x x \sin k_y y \sin k_z z\,\hat{\mathbf{x}} + C_2 \sin k_x x \cos k_y y \sin k_z z\,\hat{\mathbf{y}} + C_3 \sin k_x x \sin k_y y \cos k_z z\,\hat{\mathbf{z}},$$

times $e^{-i\omega t}$, where C_1, C_2, and C_3 are constants.

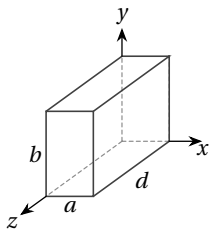


Figure 23.31 The cavity of Exercise P23.16.

(a) What does $\mathbf{E}^{\parallel} = 0$ require of E_x, E_y, and E_z on the walls $x = 0$ and $x = a$? Show that it holds at $x = 0$ automatically, and at $x = a$ if $k_x = m\pi/a$ with m an integer. In the same way, $k_y = n\pi/b$ and $k_z = \ell\pi/d$.

(b) What is $\mathbf{E}$ if two of the three integers m, n, and ℓ are zero?

(c) Show that $\nabla \cdot \mathbf{E} = 0$ requires $C_1 m\pi/a + C_2 n\pi/b + C_3 \ell\pi/d = 0$.

(d) Find $\mathbf{B}$ from Faraday's law. (It satisfies $B^{\perp} = 0$ on all six walls; you need not show this.)

(e) Put E_x into the wave equation, and show that the mode can exist only at the frequency $\omega = c\pi\sqrt{(m/a)^2 + (n/b)^2 + (\ell/d)^2}$. Check that this agrees with (23.27).

(f) Find the twenty lowest resonant frequencies, in hertz, of a cavity 1 cm × 3 cm × 7 cm.

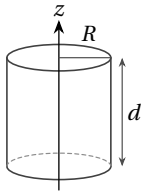


Figure 23.32 The cylindrical cavity of Exercise P23.17.

P23.17 A cavity for electron spin resonance. A cylindrical cavity of radius R and height d is a round guide closed off by conducting caps at $z = 0$ and $z = d$ (Figure 23.32).

(a) Use Exercise P23.13 and Section 23.7.1 to find the resonant frequencies of its $\mathrm{TE}_{\alpha m\ell}$ modes as a function of α, m, and ℓ.

(b) Find the resonant frequency, in hertz, of the TE_{011} mode (Figure 23.25) of a cavity 1.0 cm in radius and 3.0 cm high.

(c) Show that the TM_{111} mode has the same resonant frequency as TE_{011}, whatever the proportions of the cavity.

(d) The silver-plated cavity of Exercise P22.5 was about 1.9 cm across and 3.8 cm tall. Find the resonant frequency of its TE_{011} mode, and compare it with the frequencies near 10 GHz at which the cavity was used. The cavity held pieces of a material with a large dielectric constant, called *dielectric resonators*. Use Exercise P23.3 to explain how they account for the difference, and find the dielectric constant that would bring the TE_{011} mode to 10 GHz if it filled the whole cavity. (The resonators filled only part of the cavity, so their dielectric constant had to be larger than that.)

P23.18 The quality factor of a cavity. Find Q for the TM_{010} mode of the cavity of Example 23.4, whose walls are a good conductor with conductivity σ.

(a) Use (22.15) to show that at normal incidence a good conductor absorbs the average power $\omega\delta B_\parallel^2/4\mu_0$ per unit area, where $B_\parallel$ is the amplitude of the magnetic field just outside its surface and δ is its skin depth. (This holds closely for any field at the surface of a good conductor, so use it for the walls of the cavity.)

(b) Find the energy stored in the cavity. When the electric field is largest, all of the energy is electric. You will need $\int_0^1 J_0^2(u_{01}x)\,x\,dx = \frac{1}{2}J_1^2(u_{01})$.

(c) Find the power lost in the side wall and in the two end walls. You will need $\int_0^1 J_1^2(u_{01}x)\,x\,dx = \frac{1}{2}J_1^2(u_{01})$.

(d) Show that $Q = Rd/[\delta(R+d)]$, and evaluate it for copper walls ($\sigma = 5.96\times10^7$ S/m, as in Section 16.4.1) with $R = d = 5.0$ cm.

Exercises for 23.8 Transmission lines

P23.19 The speed on any line. A transmission line has two conductors of any cross section, with empty space between them, and carries a TEM wave along $+z$.

(a) Use Gauss's law and Ampère's law around one conductor, with (23.29), to show that the current is $I = c\lambda$, where λ is the charge per unit length on that conductor.

(b) Integrate **B** along a path in the cross section from one conductor to the other to show that the magnetic flux per unit length between the conductors is V/c.

(c) Conclude that $\mathscr{L}\mathscr{C} = 1/c^2$ for any line in empty space. With a material of permittivity ε between the conductors, $\mathscr{C}$ is larger by the factor $\varepsilon/\varepsilon_0$, and $\mathscr{L}$ is unchanged. Show that then $v = c/n$.

P23.20 Parallel wires. Two long parallel wires of radius a, with their centers a distance D apart, form a transmission line, with empty space around them (Figure 23.33).

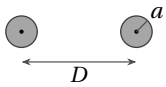


Figure 23.33 The two wires of Exercise P23.20, in cross section.

(a) For $D \gg a$, find the capacitance per unit length from the potential of two opposite line charges, and then $\mathscr{L}$ from Exercise P23.19. Show that $Z_c = (Z_0/\pi)\ln(D/a)$.

(b) Flat cable of this kind, called twin lead, made with $Z_c = 300\,\Omega$, was once common for connecting television antennas. What ratio D/a gives $300\,\Omega$ with empty space around the wires?

(c) The exact capacitance per unit length, without the assumption $D \gg a$, is $\pi\varepsilon_0/\cosh^{-1}(D/2a)$. How much does it change your answer to (b)?

P23.21 Feeding an antenna. A half-wave antenna has a radiation resistance of about $73\,\Omega$ (Section 28.5), and a $50\,\Omega$ cable carries power to it.

(a) What fraction of the power that reaches the antenna is reflected back down the cable?

(b) What characteristic impedance would a cable need in order to reflect none? What ratio b/a would a coaxial cable filled with plastic of $\varepsilon_r = 2.25$ need for it?

(c) A steady sinusoidal signal on the $50\,\Omega$ cable makes a partly standing wave, as on the string of Exercise P21.4. Find the ratio of the largest to the smallest amplitude of the voltage along the cable.

P23.22 Joining two cables. A $50\,\Omega$ cable is joined to a $75\,\Omega$ cable, and a voltage pulse of amplitude V_I travels along the first toward the joint.

(a) At the joint the voltage and the current are each the same on both sides. Find the amplitudes of the reflected and transmitted pulses.

(b) Show that the incident pulse's power equals the sum of the powers of the reflected and transmitted pulses.

(c) Compare your amplitudes with the reflection and transmission coefficients of light at normal incidence, (21.16).

Exercises for 23.9 Summary

P23.23 True or false? Explain your reasoning.

(a) A hollow metal pipe can carry an electromagnetic wave of any frequency.

(b) The fields of every mode of a hollow guide are perpendicular to the direction in which the mode travels.

(c) The crests of a wave in a hollow guide can travel faster than light.

(d) Below its cutoff frequency, a mode sent into a guide with perfectly conducting walls is absorbed by the walls.

(e) In a round guide, the TM modes with $\alpha = 0$ do not depend on ϕ.

(f) A cavity with perfectly conducting walls can hold a standing wave of any frequency above some minimum.

(g) A higher Q means that the energy stored in a cavity dies away more slowly.

(h) A coaxial cable can carry a steady current.

(i) A TEM wave on a coaxial cable filled with plastic travels at c.

(j) The characteristic impedance of a transmission line depends on its length.

(k) A voltage pulse that reaches the open end of a transmission line comes back inverted.

P23.24 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as the fields of a mode of a rectangular guide in three dimensions, a pulse spreading as it travels near the cutoff frequency of a guide, the modes of a guide with an L-shaped or elliptical cross section, the fields of the modes of a cylindrical cavity, a pulse bouncing back and forth on a line with mismatched loads at both ends, or the power lost in the copper walls of a rectangular guide.

Chapter 24

Inhomogeneous Wave Equations

Chapters 15 and 16 let the fields change in time, but their methods worked only for fields that change slowly, or for special cases such as currents that rise at a steady rate. This chapter takes up the general problem: given charges and currents that change in time in any way, find the electric and magnetic fields they produce. The most important case, and the one that drives this chapter and the next three, is a single point charge in motion. Its fields depend on where it is, and also on how fast it is moving and how it is accelerating.

Section 15.4 took the first step. Because $\nabla \cdot \mathbf{B} = 0$ at every instant, the magnetic field is still the curl of a vector potential, and the electric field is $\mathbf{E} = -\nabla V - \partial \mathbf{A}/\partial t$, (15.11). The potentials are not unique: the gauge transformation (15.12) changes them without changing the fields. In the Coulomb gauge, V at each instant is the potential of the charges where they are at that instant, even though the fields respond to a change in the charges only after the news of the change has had time to arrive. That suggests looking for a gauge in which the potentials, too, wait for the news. The delay itself is not a discovery of relativity. It is contained in Maxwell's equations, which predict that changes in the fields travel at the speed of light (Section 16.5).

The plan of the chapter is as follows. Section 24.1 writes Maxwell's equations in terms of the potentials and shows that one choice of gauge, the Lorenz gauge, separates them. Section 24.2 shows that the resulting equations are wave equations with sources. Sections 24.3 and 24.4 solve them: each piece of the source contributes what it was doing at an earlier time, the retarded time. In Section 24.5 the potentials are differentiated once and for all to give the fields themselves, and Section 24.6 shows how to compute retarded times and retarded integrals numerically. The fields of a moving point charge need one more tool, relativity, and are the subject of Chapters 25–27.

24.1 The Lorenz gauge

Two of Maxwell's equations are already built into the potentials, and the other two, the ones that contain the sources, become equations for V and $\mathbf{A}$. This section

writes those two equations down and then chooses a gauge that makes them as simple as possible.

24.1.1 Maxwell's equations for the potentials

With

$$\mathbf{E} = -\nabla V - \frac{\partial \mathbf{A}}{\partial t}, \qquad \mathbf{B} = \nabla \times \mathbf{A},$$

the two equations without sources hold automatically. The divergence of a curl is zero, (1.57), so $\nabla \cdot \mathbf{B} = 0$ for any $\mathbf{A}$; and the curl of a gradient field is zero, (1.56), so

$$\nabla \times \mathbf{E} = -\nabla \times \frac{\partial \mathbf{A}}{\partial t} = -\frac{\partial}{\partial t}(\nabla \times \mathbf{A}) = -\frac{\partial \mathbf{B}}{\partial t},$$

which is Faraday's law. (These two equations are the reason the potentials exist, as Section 15.4 showed.)

The other two contain the sources. Gauss's law becomes

$$\nabla \cdot \mathbf{E} = -\nabla^2 V - \frac{\partial}{\partial t}(\nabla \cdot \mathbf{A}) = \frac{\rho}{\varepsilon_0},$$

or

$$\nabla^2 V + \frac{\partial}{\partial t}(\nabla \cdot \mathbf{A}) = -\frac{\rho}{\varepsilon_0}. \tag{24.1}$$

In the Ampère–Maxwell law, (16.1), the curl of $\mathbf{B}$ is $\nabla \times (\nabla \times \mathbf{A}) = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A}$, by the identity (1.58), so

$$\nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A} = \mu_0 \mathbf{J} + \mu_0 \varepsilon_0 \frac{\partial}{\partial t}\left(-\nabla V - \frac{\partial \mathbf{A}}{\partial t}\right).$$

Collecting terms, and writing $\mu_0 \varepsilon_0 = 1/c^2$,

$$\left(\nabla^2 \mathbf{A} - \frac{1}{c^2}\frac{\partial^2 \mathbf{A}}{\partial t^2}\right) - \nabla\left(\nabla \cdot \mathbf{A} + \frac{1}{c^2}\frac{\partial V}{\partial t}\right) = -\mu_0 \mathbf{J}. \tag{24.2}$$

Equations (24.1) and (24.2) contain all of the information in Maxwell's equations. They are also untidy: V and $\mathbf{A}$ appear in both, and each equation mixes several kinds of derivatives. The freedom to choose a gauge is what lets us tidy them up.

24.1.2 The Coulomb gauge again

In the **Coulomb gauge** of Section 12.1.1, $\nabla \cdot \mathbf{A} = 0$, and Gauss's law (24.1) becomes Poisson's equation,

$$\nabla^2 V = -\frac{\rho}{\varepsilon_0},$$

the same equation as in electrostatics. As Section 15.4 showed, it is solved at each instant by the potential of the charges where they are at that instant, and the fields

still change only after the news of a change has arrived. That much is pleasant. But the equation for **A**, (24.2) with $\nabla\cdot\mathbf{A}=0$,

$$\nabla^2\mathbf{A}-\frac{1}{c^2}\frac{\partial^2\mathbf{A}}{\partial t^2}=-\mu_0\mathbf{J}+\frac{1}{c^2}\nabla\frac{\partial V}{\partial t},$$

has, besides the current, a second source term that depends on how the charges everywhere are changing. The simple equation for V comes at the price of a difficult one for **A**, and when the charges move that price is too high.

24.1.3 The Lorenz condition

The terms that couple the two equations are $\partial(\nabla\cdot\mathbf{A})/\partial t$ in (24.1) and the gradient in (24.2). The second disappears if we choose

$$\boxed{\nabla\cdot\mathbf{A}=-\frac{1}{c^2}\frac{\partial V}{\partial t}}. \tag{24.3}$$

This is the **Lorenz condition**, and potentials that satisfy it are said to be in the **Lorenz gauge**. With it, the term $\partial(\nabla\cdot\mathbf{A})/\partial t$ in (24.1) becomes $-(1/c^2)\,\partial^2V/\partial t^2$, and the two equations separate:

$$\nabla^2V-\frac{1}{c^2}\frac{\partial^2V}{\partial t^2}=-\frac{\rho}{\varepsilon_0},\qquad \nabla^2\mathbf{A}-\frac{1}{c^2}\frac{\partial^2\mathbf{A}}{\partial t^2}=-\mu_0\mathbf{J}.$$

Each potential now has its own equation, with its own source, and the two equations have the same form. Section 24.2 takes them up.

Can the Lorenz condition always be met? Suppose potentials V and **A** do not satisfy it. The gauge transformation (15.12), $\mathbf{A}\to\mathbf{A}+\nabla\psi$ and $V\to V-\partial\psi/\partial t$, changes $\nabla\cdot\mathbf{A}+(1/c^2)\,\partial V/\partial t$ by $\nabla^2\psi-(1/c^2)\,\partial^2\psi/\partial t^2$. The new potentials therefore satisfy the condition if ψ obeys

$$\nabla^2\psi-\frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2}=-\left(\nabla\cdot\mathbf{A}+\frac{1}{c^2}\frac{\partial V}{\partial t}\right),$$

an equation of exactly the kind that this chapter learns to solve, with a known function on the right (Exercise P24.2). The condition does not remove all of the freedom: a gauge function that obeys $\nabla^2\psi-(1/c^2)\,\partial^2\psi/\partial t^2=0$ can still be added without spoiling it (Exercise P24.3), just as a ψ with $\nabla^2\psi=0$ could be added in the Coulomb gauge.

The condition is named for the Danish physicist Ludvig Lorenz, whom we met in Section 9.4.1. He wrote it down in 1867, together with the potentials of Section 24.4. It was long credited to Hendrik Lorentz instead, and many books still call it the "Lorentz gauge."

24.2 The wave equations for V and A

In the Lorenz gauge, then, the potentials obey

$$\boxed{\Box^2 V = -\frac{\rho}{\varepsilon_0}}, \tag{24.4}$$

$$\boxed{\Box^2 \mathbf{A} = -\mu_0 \mathbf{J}}, \tag{24.5}$$

where

$$\Box^2 \equiv \nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2} \tag{24.6}$$

is called the **d'Alembertian** ("box squared"), after Jean le Rond d'Alembert, who found the wave equation of a vibrating string, and its general solution, in 1747. Three things make these equations attractive.



Jean le Rond d'Alembert (1717–1783, French) D'Alembert was abandoned as a newborn on the steps of the church of Saint-Jean-le-Rond in Paris, from which he took his name, and he was raised by a glazier's wife, with whom he lived for nearly fifty years. He studied law and then medicine before teaching himself mathematics, and he was elected to the Academy of Sciences in 1741. His *Traité de dynamique* (1743) stated the principle of mechanics that bears his name, and in 1747 he wrote down the equation of a vibrating string and its general solution, two waves traveling in opposite directions. With Denis Diderot he edited the *Encyclopédie*, for which he wrote the Preliminary Discourse (1751). He became the permanent secretary of the Académie française in 1772 and died in Paris in 1783. (Wikipedia)

- Each is a wave equation, (16.13) with $v = c$, with a source term on the right. Such an equation is called an **inhomogeneous wave equation**, which gives the chapter its title. Away from the sources it is the homogeneous wave equation of Section 16.5, whose solutions are waves that travel at c.
- The same equation governs V and each Cartesian component of $\mathbf{A}$: $\Box^2 A_x = -\mu_0 J_x$, and so on. The strong analogy of Section 12.2.1 (each component of $\mathbf{A}$ depends on the same component of $\mathbf{J}$ as V depends on ρ) holds for fields that change in time too, so one solution serves for all four potentials.
- When nothing depends on time, they become Poisson's equations, $\nabla^2 V = -\rho/\varepsilon_0$, (5.26), and $\nabla^2 \mathbf{A} = -\mu_0 \mathbf{J}$, (12.7). They combine the wave equation with Poisson's equation.

Together with $\mathbf{E} = -\nabla V - \partial \mathbf{A}/\partial t$ and $\mathbf{B} = \nabla \times \mathbf{A}$, the two wave equations contain all of electrodynamics, provided that the potentials also satisfy the Lorenz condition, which was used to derive them. A solution of (24.4) and (24.5) that violated the condition would not satisfy Maxwell's equations. The solutions of Section 24.4 do satisfy it (Exercise P24.8).

The Lorenz gauge has one more advantage, which Section 26.6 uses. Written in terms of space and time together, the d'Alembertian and the Lorenz condition take the same form for every observer moving at constant velocity; the Coulomb condition $\nabla \cdot \mathbf{A} = 0$ does not. That is one reason the Lorenz gauge is the usual choice for radiation (Section 12.1.1).

24.3 Retarded time

Changes in the fields travel at the speed of light. When a source changes, the news of the change reaches a point a distance $\mathscr{r}$ away a time $\mathscr{r}/c$ later. So the fields at a

point $\mathbf{r}$ at time t can depend on what the source at $\mathbf{r}'$ was doing only at the earlier time

$$\boxed{t_r \equiv t - \frac{\mathscr{r}}{c}}, \qquad \mathscr{r} = |\mathbf{r} - \mathbf{r}'|, \tag{24.7}$$

called the **retarded time** (Figure 24.1). Each point of an extended source has its own retarded time, because each is at its own distance from the field point.

For a source that moves, the position that matters is the one it had at the retarded time. I like to call that position the **shadow** of the source. It is not where the source is now; it is where an observer at the field point *sees* the source, since light and the electromagnetic news travel together. In that sense, what you see is exactly what the fields respond to. From here on, for a moving source, $\mathscr{r}$ is measured from its shadow. Finding the shadow is harder than it sounds, even in the simplest case.

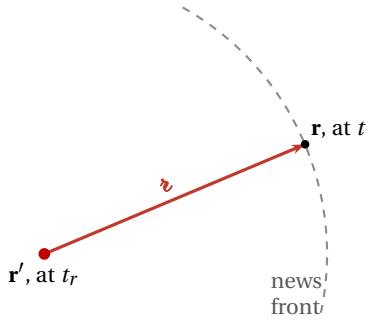


Figure 24.1 The news of what a source at $\mathbf{r}'$ does at the retarded time t_r spreads at the speed of light and reaches the field point $\mathbf{r}$ a distance $\mathscr{r}$ away at the time $t = t_r + \mathscr{r}/c$.

Example 24.1

A point charge moves along the x axis at constant speed v, passing the origin at $t = 0$. For a field point $P = (x, y, 0)$, find the distance $\mathscr{r}$ from the charge's shadow to P at time t. Then specialize to a point on the x axis ahead of the charge, and to one behind it.

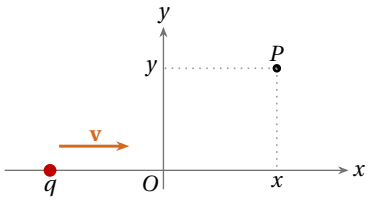
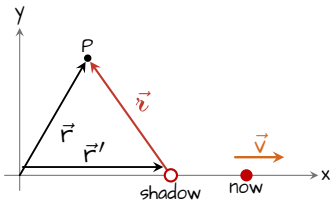


Figure 24.2 The charge of Example 24.1, moving along the x axis, and the field point P.

Solution:



The first step is a sketch of what is given and what is asked: the charge's path along the x axis, the point P, and the charge at two places, where it is now and where it was at the retarded time, its shadow. The vector $\vec{\mathscr{r}}$ runs from the shadow, not from the charge, to P.

Following the script-r method of Section 3.5, with the shadow as the source point,

$$\begin{aligned}
\mathbf{r} &= x\hat{\mathbf{x}} + y\hat{\mathbf{y}}, \\
\mathbf{r}' &= v t_r \hat{\mathbf{x}}, \\
\boldsymbol{\mathscr{r}} &= \mathbf{r} - \mathbf{r}' = (x - v t_r)\hat{\mathbf{x}} + y\hat{\mathbf{y}}, \\
\mathscr{r} &= \sqrt{(x - v t_r)^2 + y^2}.
\end{aligned}$$

The new ingredient is that t_r is not given: it depends on $\mathscr{r}$ itself, through (24.7), $t_r = t - \mathscr{r}/c$. Substituting it, and writing $X \equiv x - vt$ for the x distance from the charge's present position to P, and $\beta \equiv v/c$,

$$x - v t_r = x - vt + \frac{v\mathscr{r}}{c} = X + \beta\mathscr{r},$$

so that

$$\mathscr{r}^2 = (X + \beta\mathscr{r})^2 + y^2, \qquad \text{or} \qquad (1 - \beta^2)\mathscr{r}^2 - 2\beta X\mathscr{r} - \left(X^2 + y^2\right) = 0.$$

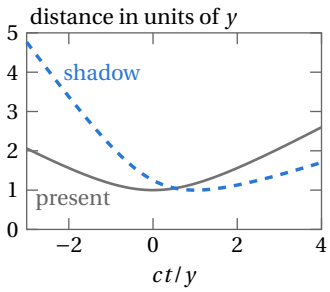


Figure 24.3 The distances in Example 24.1 for $\beta = 0.6$, from the observation point P at $(0, y)$ to the current location of the charge (solid line), and to the shadow ($\mathscr{r}$, dashed line). Animation.

This is a quadratic equation for $\mathscr{r}$. Its roots are

$$\mathscr{r} = \frac{\beta X \pm \sqrt{\beta^2 X^2 + (1-\beta^2)(X^2+y^2)}}{1-\beta^2} = \frac{\beta X \pm \sqrt{X^2 + (1-\beta^2)\,y^2}}{1-\beta^2}.$$

The square root is larger than $|X|$, and so larger than βX, so the minus sign would give a negative $\mathscr{r}$, a retarded time later than t. The shadow is therefore at the distance

$$\mathscr{r} = \frac{\beta X + \sqrt{X^2 + (1-\beta^2)\,y^2}}{1-\beta^2}. \tag{24.8}$$

On the x axis ($y = 0$), ahead of the charge ($X > 0$) and behind it ($X < 0$),

$$\mathscr{r} = \frac{X}{1-\beta} \quad \text{(ahead)}, \qquad \mathscr{r} = \frac{|X|}{1+\beta} \quad \text{(behind)}.$$

Ahead of the charge, the shadow is farther from P than the charge is now, by the factor $1/(1-\beta)$, since the charge has moved on toward P while the news traveled. Behind it, the shadow is closer to P than the charge is now. That was a lot of algebra for just about the simplest situation there is.

For a point off the axis, Figure 24.3 shows $\mathscr{r}$ as time goes by. It shrinks while the charge approaches, reaches its smallest value, y, at $t = y/c$, when the news from the moment of closest approach arrives, and grows again. Compared with the distance to the charge's present position, the curve is shifted later in time and is lopsided.

The shadow does not move at constant speed even though the charge does. On the x axis ahead of the charge, its position is $vt_r = v(t - \mathscr{r}/c)$ with $\mathscr{r} = (x - vt)/(1-\beta)$, and differentiating with respect to t gives the speed at which the shadow appears to move. Ahead of the charge, and behind it,

$$\frac{v}{1-\beta} \quad \text{(approaching)}, \qquad \frac{v}{1+\beta} \quad \text{(receding)}. \tag{24.9}$$

An approaching shadow moves faster than the charge, faster than light if $\beta > 1/2$, and a receding shadow moves more slowly. There is no contradiction with relativity: nothing travels from one position of the shadow to the next. Each position is a separate piece of news, sent at a different time from a different place, and the news of later positions has less distance to travel when the charge approaches. For the same reason, the rate at which the shadow appears to move is not the velocity of the charge at any time: the charge always moves at v. This distinction matters in Section 27.1.1, where the fields of a moving charge turn out to depend on the charge's actual velocity at the retarded time, not on the apparent velocity of its shadow. Factors like $1/(1 \pm \beta)$ look like relativity, but there is no relativity in this section, only the finite speed of the news.

24.3.1 Apparent shapes

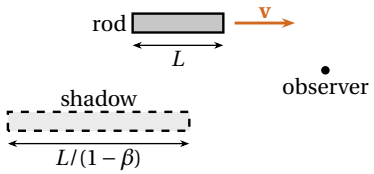


Figure 24.4 A rod moving toward the observer (top) and the shadow the observer sees at one instant (bottom): the news from the far end left earlier, when the rod was farther away, so the shadow is longer than the rod.

An object that is not a point has a shadow of a different shape. Take a rod of length L, moving at speed v along its own length, straight toward an observer (Figure 24.4). The news that arrives from the two ends at the same moment did not leave them at the same moment: the news from the far end left earlier, by a time Δt. During Δt the rod moved a distance $v\,\Delta t$ toward the observer, so the near end's shadow is a distance $L + v\,\Delta t$ ahead of the far end's, and the news from the far end must cover exactly that distance in the time Δt: $c\,\Delta t = L + v\,\Delta t$. The **apparent length** of the rod is therefore

$$c\,\Delta t = \frac{L}{1-\beta} \quad \text{(approaching)}, \qquad \frac{L}{1+\beta} \quad \text{(receding)}, \tag{24.10}$$

the second by the same argument with the rod moving away. An approaching object looks stretched and a receding one squashed, by the same factors as the speeds (24.9).

This is not the Lorentz contraction of Section 25.3.3. The Lorentz contraction makes a moving rod *shorter* by a factor that depends on v^2, and it is a statement about where the two ends are at the same time in a given frame, not about what anyone sees. What an observer actually sees combines the two effects.[1] The factor $1/(1-\beta)$ also has a cousin in Section 27.1, where the shadow of a small moving charge turns out to have a different "volume" from the charge itself.

24.4 Retarded potentials

In electrostatics the potential of a charge distribution is (5.16), and in magnetostatics the vector potential of a current distribution is (12.9). For sources that change in time, the natural guess is to evaluate each piece of the source at its own retarded time:

$$\boxed{V(\mathbf{r},t) = \frac{1}{4\pi\varepsilon_0}\int \frac{\rho(\mathbf{r}',t_r)}{\mathscr{r}}\,d\tau'}, \tag{24.11}$$

$$\boxed{\mathbf{A}(\mathbf{r},t) = \frac{\mu_0}{4\pi}\int \frac{\mathbf{J}(\mathbf{r}',t_r)}{\mathscr{r}}\,d\tau'}, \tag{24.12}$$

where $t_r = t - \mathscr{r}/c$ takes a different value at each source point. These are the **retarded potentials**, and the guess is right: they solve the wave equations (24.4) and (24.5). The proof below is a page of careful differentiation, worth reading once because it shows what the retarded time does. The retarded potentials also satisfy the Lorenz condition (Exercise P24.8), so they solve Maxwell's equations.

[1] The combination has surprising consequences: a moving sphere still looks circular in outline, at any speed, and objects appear rotated rather than contracted. See "Invisibility of the Lorentz contraction," J. Terrell, Phys. Rev. 116, 1041–1045 (1959), https://doi.org/10.1103/PhysRev.116.1041, and "The apparent shape of a relativistically moving sphere," R. Penrose, Proc. Cambridge Philos. Soc. 55, 137–139 (1959), https://doi.org/10.1017/S0305004100033776.

The idea is not new: Lorenz wrote the potentials in this form in the same 1867 paper as his gauge condition.[2]

Proof that the retarded potentials satisfy the wave equations

It is enough to treat V, since each Cartesian component of $\mathbf{A}$ has the same form. The gradient acts on the field point $\mathbf{r}$, and $\mathbf{r}$ appears in the integrand of (24.11) in two places: in $1/\mathscr{r}$, and in t_r. By the product rule,

$$\nabla V = \frac{1}{4\pi\varepsilon_0}\int\left[\frac{1}{\mathscr{r}}\nabla\rho + \rho\,\nabla\!\left(\frac{1}{\mathscr{r}}\right)\right]d\tau'.$$

The density depends on $\mathbf{r}$ only through t_r, so by the chain rule $\nabla\rho = \dot\rho\,\nabla t_r$, where the dot is the derivative with respect to time, evaluated at the retarded time. Since $\nabla\mathscr{r} = \hat{\boldsymbol{\mathscr{r}}}$ (the separation grows fastest straight away from the source point), $\nabla t_r = -\hat{\boldsymbol{\mathscr{r}}}/c$; and $\nabla(1/\mathscr{r}) = -\hat{\boldsymbol{\mathscr{r}}}/\mathscr{r}^2$, as in Chapter 1. So

$$\nabla V = \frac{1}{4\pi\varepsilon_0}\int\left[-\frac{\dot\rho\,\hat{\boldsymbol{\mathscr{r}}}}{c\mathscr{r}} - \frac{\rho\,\hat{\boldsymbol{\mathscr{r}}}}{\mathscr{r}^2}\right]d\tau'.$$

Take the divergence the same way, with $\nabla\dot\rho = \ddot\rho\,\nabla t_r = -\ddot\rho\,\hat{\boldsymbol{\mathscr{r}}}/c$ and $\nabla\rho = -\dot\rho\,\hat{\boldsymbol{\mathscr{r}}}/c$:

$$\nabla^2 V = \frac{1}{4\pi\varepsilon_0}\int\left[-\frac{1}{c}\left(\hat{\boldsymbol{\mathscr{r}}}\cdot\nabla\dot\rho\,\frac{1}{\mathscr{r}} + \dot\rho\,\nabla\cdot\frac{\hat{\boldsymbol{\mathscr{r}}}}{\mathscr{r}}\right) - \left(\frac{\hat{\boldsymbol{\mathscr{r}}}\cdot\nabla\rho}{\mathscr{r}^2} + \rho\,\nabla\cdot\frac{\hat{\boldsymbol{\mathscr{r}}}}{\mathscr{r}^2}\right)\right]d\tau'.$$

Now $\hat{\boldsymbol{\mathscr{r}}}\cdot\hat{\boldsymbol{\mathscr{r}}} = 1$; $\nabla\cdot(\hat{\boldsymbol{\mathscr{r}}}/\mathscr{r}) = 1/\mathscr{r}^2$ (the divergence in spherical coordinates centered on the source point, $(1/\mathscr{r}^2)\,\partial(\mathscr{r}^2\cdot\mathscr{r}^{-1})/\partial\mathscr{r}$); and $\nabla\cdot(\hat{\boldsymbol{\mathscr{r}}}/\mathscr{r}^2) = 4\pi\,\delta^3(\mathbf{r}-\mathbf{r}')$, (1.45). So

$$\nabla^2 V = \frac{1}{4\pi\varepsilon_0}\int\left[\frac{\ddot\rho}{c^2\mathscr{r}} - \frac{\dot\rho}{c\mathscr{r}^2} + \frac{\dot\rho}{c\mathscr{r}^2} - 4\pi\rho\,\delta^3(\mathbf{r}-\mathbf{r}')\right]d\tau'.$$

The middle terms cancel. The last term picks out the source point $\mathbf{r}' = \mathbf{r}$, where $\mathscr{r} = 0$ and $t_r = t$, and the first is $1/c^2$ times the second time derivative of (24.11):

$$\nabla^2 V = \frac{1}{c^2}\frac{\partial^2 V}{\partial t^2} - \frac{\rho(\mathbf{r},t)}{\varepsilon_0},$$

which is (24.4). The same steps, component by component, give (24.5).

The same proof works with $t+\mathscr{r}/c$ in place of $t-\mathscr{r}/c$. The resulting **advanced potentials** also solve the wave equations, but in them each point responds to what the source *will* do, so that an effect would come before its cause. We discard them on physical grounds (Exercise P24.9).

It is tempting to go one step further and delay the fields: to take Coulomb's law and the Biot–Savart law and simply evaluate ρ and $\mathbf{J}$ at the retarded time. That does *not* work. The potentials can be delayed in this simple way, but the fields cannot, and Section 24.5 shows the terms that are missing. The procedure for a problem with changing sources therefore has two steps. Find the potentials from the retarded integrals (24.11) and (24.12); then differentiate, $\mathbf{E} = -\nabla V - \partial\mathbf{A}/\partial t$ and $\mathbf{B} = \nabla\times\mathbf{A}$. (In the Coulomb gauge, by contrast, V would be the integral with ρ at t, not t_r; the delay would then be hidden in $\mathbf{A}$.)

! **Delay the potentials, not the fields.** Coulomb's law and the Biot–Savart law with ρ and $\mathbf{J}$ evaluated at t_r are wrong for changing sources. The retarded potentials are right.

[2] "On the identity of the vibrations of light with electrical currents," L. Lorenz, Phil. Mag. 34, 287–301 (1867), https://doi.org/10.1080/14786446708639882.

Example 24.2

A long, straight, neutral wire carries no current until $t = 0$, when a steady current I_0 is switched on all along it at once. Find the electric and magnetic fields at a distance s from the wire, as functions of time.

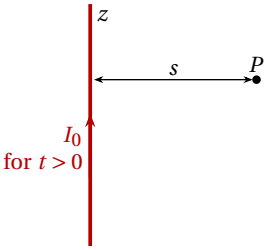
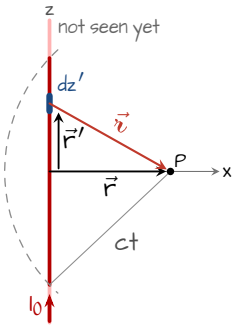


Figure 24.5 The wire of Example 24.2, with the current switched on everywhere at $t = 0$, and a field point a distance s from it.

Solution:



We begin with a sketch of what is given and what is asked: the wire along the z axis, the field point a distance s away, and a piece of the wire at height z', with the vectors $\vec{r}$, $\vec{r}\,'$ and $\vec{\mathscr{r}}$. Only the part of the wire whose news has had time to arrive, inside the circle of radius ct around the field point, can contribute.

The wire is neutral, so $V = 0$. For the vector potential, with the field point on the x axis,

$$\begin{aligned}\mathbf{r} &= s\hat{\mathbf{x}},\\ \mathbf{r}' &= z'\hat{\mathbf{z}},\\ \boldsymbol{\mathscr{r}} &= s\hat{\mathbf{x}} - z'\hat{\mathbf{z}},\\ \mathscr{r} &= \sqrt{s^2 + z'^2}.\end{aligned}$$

The piece at z' contributes only if its current was already on at its retarded time, $t - \mathscr{r}/c > 0$, that is, if $\mathscr{r} < ct$. That never happens for $t < s/c$, so before then $\mathbf{A} = 0$ and there are no fields: the news has not arrived. For $t > s/c$ the condition picks out the part of the wire with $|z'| < \sqrt{c^2t^2 - s^2}$, the segment whose ends are a distance ct from the field point. Over that segment $I(t_r) = I_0$, and

$$\mathbf{A} = \frac{\mu_0 I_0}{4\pi}\hat{\mathbf{z}}\int_{-\sqrt{c^2t^2-s^2}}^{\sqrt{c^2t^2-s^2}} \frac{dz'}{\sqrt{s^2+z'^2}} = \frac{\mu_0 I_0}{2\pi}\ln\left(\frac{ct + \sqrt{c^2t^2 - s^2}}{s}\right)\hat{\mathbf{z}}.$$

The fields follow by differentiation:

$$\mathbf{E} = -\frac{\partial \mathbf{A}}{\partial t} = -\frac{\mu_0 I_0 c}{2\pi}\frac{1}{\sqrt{c^2t^2 - s^2}}\hat{\mathbf{z}}, \tag{24.13}$$

$$\mathbf{B} = \nabla \times \mathbf{A} = -\frac{\partial A_z}{\partial s}\hat{\boldsymbol{\phi}} = \frac{\mu_0 I_0}{2\pi s}\frac{ct}{\sqrt{c^2t^2 - s^2}}\hat{\boldsymbol{\phi}}, \tag{24.14}$$

for $t > s/c$, and zero before.

At long times $ct/\sqrt{c^2t^2 - s^2} \to 1$, so $\mathbf{B}$ approaches the static field of a long wire, $\mu_0 I_0/2\pi s$, and $\mathbf{E}$ goes to zero: at long times the wire looks like any steady current. At $t = s/c$, when the news first arrives, both fields are infinite (Figure 24.6). That

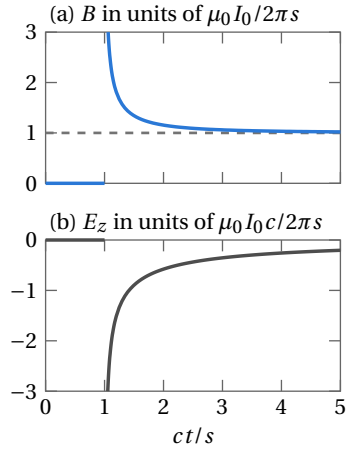


Figure 24.6 The fields of Example 24.2 against time: (a) B, which overshoots and then settles to the static value (dashed); (b) E_z, which is negative and dies away. Animation.

comes from the idealized switch: a current that jumps from zero to I_0 in no time, everywhere along an infinitely long wire. A real switch takes a finite time, and the fields stay finite (Exercise P24.10). In between, B is larger than its final value (at $ct = 2s$, by 15%). Jefimenko's equations, in the next section, show why: besides the steady current in the segment whose news has arrived, the switching itself contributes, from the two ends of that segment (Exercise P24.16). The quasistatic approximation of Section 16.4.1 would put $B = \mu_0 I_0/2\pi s$ from the instant the current is switched on, which is the limit $c \to \infty$ of (24.14).

Checks: $\mu_0 I_0 c$ has the units of $(\mathrm{T\cdot m/A})(\mathrm{A})(\mathrm{m/s}) = \mathrm{T\cdot m^2/s}$, and divided by a length that is $\mathrm{T\cdot m/s} = \mathrm{V/m}$, an electric field. The electric field points along $-\hat{\mathbf{z}}$, opposite to the new current, as Lenz's law requires of an induced field. And $\mathbf{B}$ circles the wire in the direction the right-hand rule gives for a current along $+\hat{\mathbf{z}}$.

24.5 Jefimenko's equations

For most problems the best route to the fields is the one of Example 24.2: retarded potentials first, then derivatives. But the derivatives can also be taken once and for all, inside the integrals. The result expresses the fields directly in terms of the sources.

Start with $\mathbf{E}$. The gradient of V is the one computed in the proof of Section 24.4,

$$\nabla V = \frac{1}{4\pi\varepsilon_0}\int\left[-\frac{\dot\rho\,\hat{\boldsymbol{\imath}}}{c\imath} - \frac{\rho\,\hat{\boldsymbol{\imath}}}{\imath^2}\right]d\tau',$$

and the time derivative of $\mathbf{A}$ acts only on $\mathbf{J}(\mathbf{r}', t_r)$, since $\imath$ does not depend on t:

$$\frac{\partial \mathbf{A}}{\partial t} = \frac{\mu_0}{4\pi}\int\frac{\dot{\mathbf{J}}}{\imath}\,d\tau'.$$

With $\mu_0 = 1/\varepsilon_0 c^2$,

$$\boxed{\mathbf{E}(\mathbf{r},t) = \frac{1}{4\pi\varepsilon_0}\int\left[\frac{\rho(\mathbf{r}',t_r)}{\imath^2}\hat{\boldsymbol{\imath}} + \frac{\dot\rho(\mathbf{r}',t_r)}{c\imath}\hat{\boldsymbol{\imath}} - \frac{\dot{\mathbf{J}}(\mathbf{r}',t_r)}{c^2\imath}\right]d\tau'}. \tag{24.15}$$

The curl of $\mathbf{A}$ can be taken the same way (Exercise P24.14):

$$\boxed{\mathbf{B}(\mathbf{r},t) = \frac{\mu_0}{4\pi}\int\left[\frac{\mathbf{J}(\mathbf{r}',t_r)}{\imath^2} + \frac{\dot{\mathbf{J}}(\mathbf{r}',t_r)}{c\imath}\right]\times\hat{\boldsymbol{\imath}}\,d\tau'}. \tag{24.16}$$

These are **Jefimenko's equations**, named for Oleg Jefimenko, who gave them in his 1966 textbook.[3]

[3] O. D. Jefimenko, *Electricity and Magnetism: An Introduction to the Theory of Electric and Magnetic Fields* (Appleton-Century-Crofts, New York, 1966); 2nd ed. (Electret Scientific, Star City,

The first term of each is Coulomb's law or the Biot–Savart law with the sources at the retarded time: exactly the guess that Section 24.4 warned against. The other terms are the corrections. When nothing changes, $\dot{\rho}$ and $\dot{\mathbf{J}}$ are zero, and Jefimenko's equations are Coulomb's law, (3.13), and the Biot–Savart law, (10.18), exactly. Far from the sources the correction terms matter most, because they fall off as $1/\mathscr{r}$ while the first terms fall off as $1/\mathscr{r}^2$. Fields that fall off as $1/\mathscr{r}$ can carry energy away to arbitrarily large distances; that is radiation, the subject of Chapters 27 and 28.

Notice what Jefimenko's equations do not contain: neither $\partial\mathbf{E}/\partial t$ nor $\partial\mathbf{B}/\partial t$ appears in them. The fields at a point are fixed by the charges and currents, and by their rates of change, at the retarded times alone. The Faraday current and the displacement current of Chapters 15 and 16 relate the two fields to each other, and they are very useful for finding them, but they do not appear here as sources: what the displacement current of an induced electric field adds to the Biot–Savart integral at one instant (Section 16.3.3) is contained in Jefimenko's **B** through the retarded times and the term in $\dot{\mathbf{J}}$.

Jefimenko's equations are rarely the easiest way to solve a problem by hand, but they are the natural starting point for a computer (Section 24.6). They also make some exact results easy to see.

Example 24.3

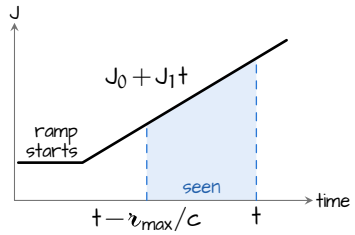
The charge density of a set of sources does not change, and their current density has been changing at a steady rate,

$$\mathbf{J}(\mathbf{r}', t) = \mathbf{J}_0(\mathbf{r}') + \mathbf{J}_1(\mathbf{r}')\, t,$$

for long enough that every source point's retarded time, seen from the field point, falls within the ramp. Show that the magnetic field is exactly the Biot–Savart field of the present current, and that the electric field is exactly the static Coulomb field plus $-(\mu_0/4\pi)\int \dot{\mathbf{J}}/\mathscr{r}\, d\tau'$.

WV, 1989). An equivalent form had appeared earlier in W. K. H. Panofsky and M. Phillips, *Classical Electricity and Magnetism*, 2nd ed. (Addison-Wesley, Reading, MA, 1962), Section 14.3. See "Time-dependent generalizations of the Biot–Savart and Coulomb laws," D. J. Griffiths and M. A. Heald, Am. J. Phys. 59, 111–117 (1991), https://doi.org/10.1119/1.16589, and "The relation between expressions for time-dependent electromagnetic fields given by Jefimenko and by Panofsky and Phillips," K. T. McDonald, Am. J. Phys. 65, 1074–1076 (1997), https://doi.org/10.1119/1.18723.

Solution:



As usual we start with a sketch: a graph of the current density at one source point against time, a straight ramp, with the window of retarded times that the field point sees, from t back to $t - r_{max}/c$, lying entirely on the ramp.

Because $\mathbf{J}$ is linear in time, its value at the retarded time is

$$\mathbf{J}(\mathbf{r}', t_r) = \mathbf{J}_0 + \mathbf{J}_1\left(t - \frac{r}{c}\right) = \mathbf{J}(\mathbf{r}', t) - \mathbf{J}_1\frac{r}{c},$$

exactly, and $\dot{\mathbf{J}} = \mathbf{J}_1$. In Jefimenko's $\mathbf{B}$, (24.16), the two terms combine:

$$\frac{\mathbf{J}(\mathbf{r}', t_r)}{r^2} + \frac{\dot{\mathbf{J}}}{cr} = \frac{\mathbf{J}(\mathbf{r}', t)}{r^2} - \frac{\mathbf{J}_1}{cr} + \frac{\mathbf{J}_1}{cr} = \frac{\mathbf{J}(\mathbf{r}', t)}{r^2}.$$

So

$$\mathbf{B}(\mathbf{r}, t) = \frac{\mu_0}{4\pi}\int \frac{\mathbf{J}(\mathbf{r}', t) \times \hat{\mathbf{r}}}{r^2}\, d\tau',$$

the Biot-Savart law with the present current. The delay in the first term is exactly undone by the second. In Jefimenko's $\mathbf{E}$, (24.15), ρ does not change, so $\rho(\mathbf{r}', t_r) = \rho(\mathbf{r}')$, the $\dot{\rho}$ term is zero, and

$$\mathbf{E}(\mathbf{r}, t) = \frac{1}{4\pi\varepsilon_0}\int \frac{\rho(\mathbf{r}')\hat{\mathbf{r}}}{r^2}\, d\tau' - \frac{\mu_0}{4\pi}\int \frac{\dot{\mathbf{J}}}{r}\, d\tau',$$

the static Coulomb field plus the induced field of Section 15.4.

These are exactly the fields that the quasistatic approximation gives, so for a steady ramp that approximation is not an approximation at all: the first step of Section 16.4.1 is exact (Figure 16.13). The condition "for long enough" is now precise: the news of the start of the ramp must have arrived from every part of the source. For a closed circuit the same result is even quicker in terms of the vector potential: with $I(t) = kt$,

$$\oint \frac{I(t_r)\, d\boldsymbol{\ell}'}{r} = k\oint \frac{(t - r/c)\, d\boldsymbol{\ell}'}{r} = kt\oint \frac{d\boldsymbol{\ell}'}{r} - \frac{k}{c}\oint d\boldsymbol{\ell}' = kt\oint \frac{d\boldsymbol{\ell}'}{r},$$

since $\oint d\boldsymbol{\ell}' = 0$ around any closed path; $\mathbf{A}$ is the static vector potential of the present current (Exercise P24.11). Checks: the units of the two terms in the combination above are the same, those of $\mathbf{J}/r^2$; and for $\mathbf{J}_1 = 0$ the results are those of magnetostatics, as they must be.

24.6 Numerical methods: retarded times and integrals

The retarded integrals can be done by hand only in a few cases. A computer handles the rest, and two things need care: finding retarded times, and integrating over sources whose retarded times differ from point to point.

24.6.1 Retarded times

For a point source that moves along a known path $\mathbf{w}(t')$, the retarded time for the field point $\mathbf{r}$ at time t is the solution t' of

$$f(t') \equiv c\,(t - t') - |\mathbf{r} - \mathbf{w}(t')| = 0, \qquad t' \le t.$$

If the source always moves more slowly than light, there is exactly one solution. The derivative of f is

$$\frac{df}{dt'} = -c + \hat{\boldsymbol{\imath}} \cdot \mathbf{v}(t'),$$

where $\hat{\boldsymbol{\imath}}$ is the unit vector from $\mathbf{w}(t')$ toward $\mathbf{r}$ and $\mathbf{v} = d\mathbf{w}/dt'$, and since $|\hat{\boldsymbol{\imath}} \cdot \mathbf{v}| \le v < c$ it is always negative: f decreases steadily as t' increases. At $t' = t$, $f = -|\mathbf{r} - \mathbf{w}(t)| \le 0$, and as t' goes back into the past, f grows without limit, because $c\,(t - t')$ grows faster than the distance to the source can. So f crosses zero exactly once: every field point sees exactly one shadow at each instant. The **retarded time** is then easy to find numerically. Step back from $t' = t$, doubling the step each time, until $f > 0$; the root is then bracketed (Figure 24.7), and bisection, or a faster bracketing method such as Brent's method, finds it to machine precision in a few dozen evaluations of f.

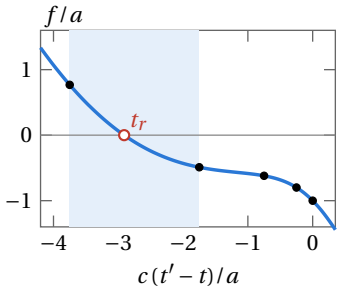


Figure 24.7 The function $f(t')$ for a charge circling at $\beta = 0.9$ on an orbit of radius a, seen at time t from a point in the plane of the orbit a distance $2a$ from its center. The dots are the steps back from $t' = t$, each twice as long as the one before; the first with $f > 0$ brackets the root (shaded), the retarded time t_r.

24.6.2 Retarded integrals

For an extended source, each point of the source has its own retarded time, which is known explicitly, $t_r = t - \imath/c$, so the integrals (24.11) and (24.12), or Jefimenko's (24.15) and (24.16), can be done by ordinary numerical quadrature. Two cautions. First, a source that is switched on or off abruptly makes the integrand jump at the source points whose retarded times fall on the switch, as in Example 24.2; quadrature rules converge slowly across a jump, so split the range of integration there. Second, Jefimenko's form needs $\dot{\rho}$ and $\dot{\mathbf{J}}$, which are usually known exactly from the description of the source, so it avoids the numerical derivatives that differentiating the potentials would require.

24.6.3 Tips for Mathematica and Python

- For the retarded time, give the root finder a bracket rather than a starting guess: `brentq` in Python's `scipy.optimize`, or, in Mathematica, `FindRoot` with the option `Method -> "Brent"` and the two ends of the bracket as its starting values, `{t', a, b}`. A bracketing method cannot wander away from the root.
- Write the path $\mathbf{w}(t')$ as a function, so that the same code works for any motion.
- For retarded integrals, `scipy.integrate.quad` accepts a list of `points` where the integrand has a kink or a jump; in Mathematica, list the break points in the range of `NIntegrate`, `{z, a, z1, z2, b}`.

- Check the code on a case you can do by hand, such as Example 24.2 or a charge at rest, before trusting it on one you cannot.

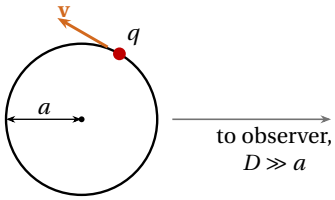


Figure 24.8 The orbit of Example 24.4 and the direction of the distant observer, in the plane of the orbit.

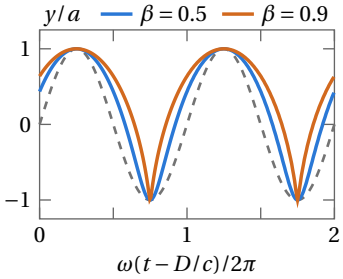
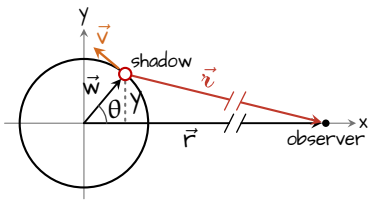


Figure 24.9 The shadow's position across the line of sight in Example 24.4 against the time of observation, for $\beta = 0.5$ and 0.9, with the slow-motion limit (dashed).

Example 24.4

A charge moves around a circle of radius a at constant speed $v = \beta c$. An observer far away, at a distance $D \gg a$ in the plane of the orbit, watches its shadow. Find the shadow's position across the observer's line of sight as a function of the time of observation, for $\beta = 0.5$ and $\beta = 0.9$.

Solution:



The first step is a sketch: the circle in the xy plane with the charge on it, the observer far away along the $+x$ axis, the charge's angle θ at the retarded time, and the vectors $\vec{r}$, $\vec{w}$ and $\vec{v}$. The observer sees the shadow's y coordinate as its position across the line of sight.

The charge is at $\mathbf{w}(t') = a\cos\omega t'\,\hat{\mathbf{x}} + a\sin\omega t'\,\hat{\mathbf{y}}$, with $\omega = v/a$, and the observer at $\mathbf{r} = D\hat{\mathbf{x}}$. The retarded time solves

$$c\,(t - t_r) = |\mathbf{r} - \mathbf{w}(t_r)| = \sqrt{(D - a\cos\omega t_r)^2 + a^2\sin^2\omega t_r},$$

which has exactly one root for each t, by Section 24.6.1. We solve it with a bracketing root finder for $D = 1000a$ and many values of t over two periods. (For $D \gg a$ the distance is $D - a\cos\omega t_r$ to good accuracy, which gives the simpler relation $\omega(t - D/c) = \theta - \beta\cos\theta$, with $\theta = \omega t_r$; the angles from the numerical roots agree with it to within about a/D.) The shadow's position across the line of sight is then $y = a\sin\theta$. Figure 24.9 shows it. At $\beta = 0.5$ the curve is already lopsided, and at $\beta = 0.9$ the shadow races through one part of the orbit and crawls through the rest.

The race happens where the charge moves toward the observer. There, as in Example 24.1, the news from later positions has less far to travel, and the shadow moves faster by the factor $1/(1-\beta)$, which is 10 at $\beta = 0.9$. Where the charge moves away, the shadow moves more slowly, by $1/(1+\beta) = 0.53$. From the simpler relation, the half of the orbit that approaches the observer (θ from $-\pi$ to 0) is seen in the time $(\pi - 2\beta)/\omega$, and the receding half in $(\pi + 2\beta)/\omega$: at $\beta = 0.9$, 21% and 79% of the period. The period itself is the same as the charge's, $2\pi/\omega$, since the pattern repeats once per orbit. Section 27.6 shows that the fields of the charge are concentrated in the short stretch when the shadow races; the observer receives them as a brief pulse once per orbit.

Checks: the largest residual $|f(t_r)|$ in the computation was about 10^{-13} of the orbit's radius, as expected of a root found to machine precision; the period of the computed

curves is $2\pi/\omega$; and as $\beta \to 0$ the curve becomes the sinusoid $a\sin\omega t$, shifted by the travel time D/c.

24.7 Summary: triangle diagram of sources, fields, and potentials

When the sources change in time, the potentials are still the most convenient path from the sources to the fields, provided that they are chosen well. In the Lorenz gauge they obey wave equations with the sources on the right, and their solutions are the static potentials with each piece of the source evaluated at its own retarded time. The fields then follow by differentiation.

Chapters 5 and 12 summarized the relations among the sources, fields and potentials of electrostatics and magnetostatics in two triangle diagrams. The same can now be done for electrodynamics, with ρ and $\mathbf{J}$ together at the top, $\mathbf{E}$ and $\mathbf{B}$ at the lower left, and V and $\mathbf{A}$ at the lower right. The result is the **triangle diagram for electrodynamics** (Figure 24.10).

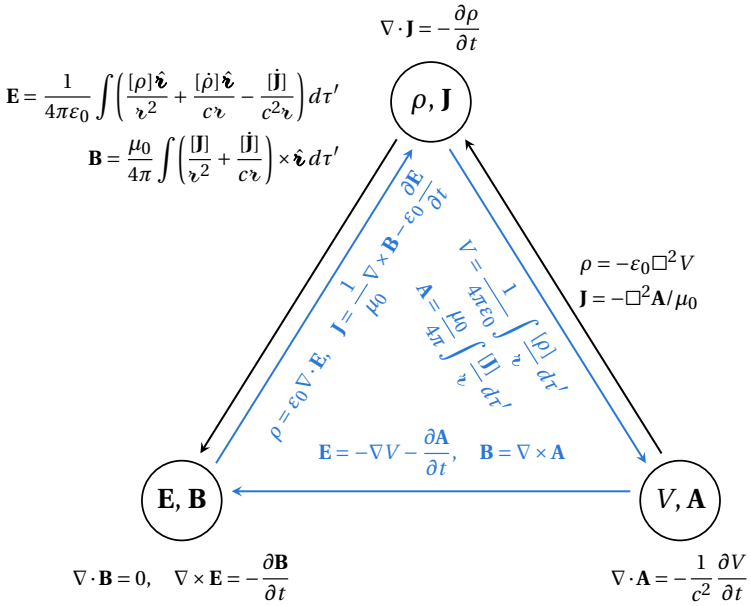


Figure 24.10 The triangle diagram for electrodynamics: one triangle diagram to rule them all. Each corner carries the equations its quantities must satisfy: the continuity equation at the top, the two Maxwell's equations without sources at the lower left, and the Lorenz condition at the lower right. The black arrows go counterclockwise around the outside of the triangle, and the blue arrows go clockwise around the inside. There is no black arrow along the bottom: the fields do not determine the potentials. Square brackets mean that a quantity is evaluated at the retarded time.

Each corner carries the equations that its own quantities must satisfy, and each

arrow gives the quantities at its head in terms of those at its tail. There are five arrows rather than six. The fields do not determine the potentials: a gauge transformation changes V and $\mathbf{A}$ without changing $\mathbf{E}$ and $\mathbf{B}$, so there is no arrow along the bottom from the fields to the potentials. When nothing depends on time, the diagram reverts to the triangle diagrams of Chapters 5 and 12: the retarded times become ordinary times, the d'Alembertians become Laplacians, and Jefimenko's equations become Coulomb's law and the Biot–Savart law. The static diagrams also have the arrow that is missing here, $V = -\int \mathbf{E}\cdot d\boldsymbol{\ell}$ and, with high symmetry, $\oint \mathbf{A}\cdot d\boldsymbol{\ell} = \Phi_B$: when nothing changes, $\mathbf{E}$ fixes V up to the choice of a reference point, and the Coulomb condition fixes the gauge of $\mathbf{A}$.

Potentials and their equations

$\mathbf{E} = -\nabla V - \dfrac{\partial \mathbf{A}}{\partial t}, \quad \mathbf{B} = \nabla \times \mathbf{A}$	the fields from the potentials, in any gauge
$\nabla \cdot \mathbf{A} = -\dfrac{1}{c^2}\dfrac{\partial V}{\partial t}$	the Lorenz condition; it can always be met by a gauge transformation
$\Box^2 \equiv \nabla^2 - \dfrac{1}{c^2}\dfrac{\partial^2}{\partial t^2}$	the d'Alembertian
$\Box^2 V = -\dfrac{\rho}{\varepsilon_0}, \quad \Box^2 \mathbf{A} = -\mu_0 \mathbf{J}$	the inhomogeneous wave equations, in the Lorenz gauge; Poisson's equations when nothing depends on time
$\nabla^2 V = -\dfrac{\rho}{\varepsilon_0}, \quad \nabla \cdot \mathbf{A} = 0$	the Coulomb gauge: V responds instantly to the charges; the fields do not

Retarded time and retarded potentials

$t_r = t - \dfrac{\mathscr{r}}{c}$	the retarded time: news of the source at $\mathbf{r}'$ reaches $\mathbf{r}$ a time $\mathscr{r}/c$ later; for a moving source $\mathscr{r}$ is measured from its shadow, its position at t_r
$V = \dfrac{1}{4\pi\varepsilon_0}\displaystyle\int \frac{\rho(\mathbf{r}', t_r)}{\mathscr{r}}\, d\tau'$	the retarded potentials: the static potentials with each piece of the source at its own retarded time; they satisfy the Lorenz condition
$\mathbf{A} = \dfrac{\mu_0}{4\pi}\displaystyle\int \frac{\mathbf{J}(\mathbf{r}', t_r)}{\mathscr{r}}\, d\tau'$	
$\dfrac{v}{1 \mp \beta}, \quad \dfrac{L}{1 \mp \beta}$	the apparent speed of a shadow and the apparent length of a rod moving along the line of sight, approaching (upper sign) and receding; this is not relativity

The fields directly: Jefimenko's equations

$\mathbf{E} = \dfrac{1}{4\pi\varepsilon_0}\displaystyle\int \left[\frac{\rho\hat{\boldsymbol{\mathscr{r}}}}{\mathscr{r}^2} + \frac{\dot{\rho}\hat{\boldsymbol{\mathscr{r}}}}{c\mathscr{r}} - \frac{\dot{\mathbf{J}}}{c^2\mathscr{r}}\right] d\tau'$	Coulomb's law with retarded sources, plus corrections; every quantity in the brackets is evaluated at t_r
$\mathbf{B} = \dfrac{\mu_0}{4\pi}\displaystyle\int \left[\frac{\mathbf{J}}{\mathscr{r}^2} + \frac{\dot{\mathbf{J}}}{c\mathscr{r}}\right] \times \hat{\boldsymbol{\mathscr{r}}}\, d\tau'$	the Biot–Savart law with retarded sources, plus a correction; the $1/\mathscr{r}$ terms dominate far away and carry radiation
caution	delay the potentials, not the fields: Coulomb's law and the Biot–Savart law with ρ and $\mathbf{J}$ simply evaluated at t_r are wrong

Numerical methods. For a point source moving more slowly than light, $f(t') = c(t - t') - |\mathbf{r} - \mathbf{w}(t')|$ decreases steadily, so it has exactly one root, the retarded time; bracket it and use bisection or Brent's method. Retarded integrals of extended sources can be done by quadrature, splitting the range where the source switches on or off; Jefimenko's form avoids numerical derivatives.

Exercises

Exercises for 24.1 The Lorenz gauge

P24.1 A point charge q sits at rest at the origin, with the usual potentials $V = q/4\pi\varepsilon_0 r$ and $\mathbf{A} = 0$.

(a) Find a gauge function $\psi(\mathbf{r}, t)$ for which the new scalar potential is zero, and find the new vector potential.

(b) Check that the new potentials give the same **E** and **B**.

(c) Do the new potentials satisfy the Coulomb condition? The Lorenz condition? Be careful at the origin.

P24.2 Section 24.1.3 claimed that the Lorenz condition can always be met.

(a) Show that if V and **A** are any potentials for the fields, the gauge transformation (15.12) with a function ψ that satisfies $\Box^2\psi = -(\nabla \cdot \mathbf{A} + (1/c^2)\,\partial V/\partial t)$ gives potentials that satisfy the Lorenz condition.

(b) Write down a solution ψ of that equation, using the methods of Section 24.4.

(c) Show that if V and **A** already satisfy the Lorenz condition, a further gauge transformation keeps them in the Lorenz gauge if and only if $\Box^2\psi = 0$.

P24.3 Exercise P16.17 described a plane wave by the potentials $V = 0$ and $\mathbf{A} = (E_0/\omega)\sin(kz - \omega t)\,\hat{\mathbf{x}}$, with $\omega = ck$.

(a) Show that these potentials satisfy both the Coulomb condition and the Lorenz condition.

(b) Find a gauge function ψ with $\Box^2\psi = 0$ that gives new potentials in the Lorenz gauge with $V \neq 0$. Check that they give the same fields.

(c) Show that if potentials satisfy both the Coulomb and the Lorenz conditions, then V does not change in time.

Exercises for 24.2 The wave equations for V and A

P24.4 Let f be any function of one variable.

(a) Show that $V(r, t) = f(t - r/c)/r$, with r the distance from the origin, satisfies the homogeneous wave equation $\Box^2 V = 0$ everywhere except at the origin.

(b) What point source at the origin would produce this V as its retarded potential, (24.11)?

(c) Such a source, by itself, violates the conservation of charge. Explain why, and say what else would have to be present in a real situation.

Exercises for 24.3 Retarded time

P24.5 A small charged ball is dropped from rest at $t = 0$ from the top of a cliff, where you stand (Figure 24.11). Ignore air resistance.

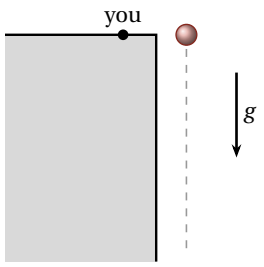


Figure 24.11 The cliff of Exercise P24.5, with the observer at the top and the ball falling from rest.

(a) Find $\mathscr{r}(t)$, the distance between you and the ball's shadow, as a function of time.

(b) Expand your answer for $gt \ll c$, keeping the first two terms, and explain the sign of the second term.

P24.6 A blob of hot gas is ejected from the center of a distant galaxy at speed $v = \beta c$, at an angle θ to our line of sight, toward us (Figure 24.12).

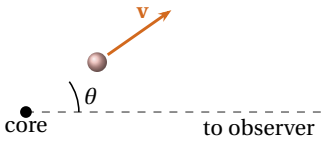


Figure 24.12 The blob of Exercise P24.6, moving at angle θ to the line of sight toward a distant observer.

(a) Show that the speed at which its shadow appears to move across the sky (its transverse displacement divided by the time between our observations) is

$$v_{\text{app}} = \frac{v \sin\theta}{1 - \beta\cos\theta}.$$

(b) Find the angle at which v_{app} is largest, and show that its largest value is $\beta c/\sqrt{1-\beta^2}$, which exceeds c when $\beta > 1/\sqrt{2}$.

(c) Radio astronomers observe blobs in the jets of some quasars that appear to move across the sky at several times the speed of light. What is the smallest speed v that can explain an apparent speed of $5c$? (The explanation was proposed before the observations, in "Appearance of relativistically expanding radio sources," M. J. Rees, Nature 211, 468–470 (1966), https://doi.org/10.1038/211468a0.)

P24.7 A ball of radius R moves upward at speed v past an observer some distance to its side (Figure 24.13).

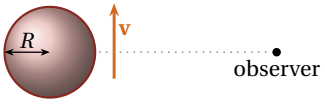


Figure 24.13 The ball of Exercise P24.7, moving upward past an observer to its side.

(a) Sketch the shadow of the ball, as the observer sees it at the moment when the ball's center is level with the observer. Explain the sketch by thinking about the retarded times of the near and far sides of the ball.

(b) A uniformly charged ball has total charge Q. Would the integral of its charge density over its shadow, region by region, give Q? Explain.

Exercises for 24.4 Retarded potentials

P24.8 Show that the retarded potentials (24.11) and (24.12) satisfy the Lorenz condition. Hints: $\mathbf{J}(\mathbf{r}', t_r)$ depends on $\mathbf{r}'$ both directly and through t_r. Show that

$$\nabla \cdot \frac{\mathbf{J}(\mathbf{r}', t_r)}{\mathscr{r}} = -\nabla' \cdot \frac{\mathbf{J}(\mathbf{r}', t_r)}{\mathscr{r}} + \frac{1}{\mathscr{r}} \nabla' \cdot \mathbf{J},$$

where in the last term the divergence acts only on the direct dependence; then use the continuity equation (10.12), and turn the integral of $\nabla' \cdot (\mathbf{J}/\mathscr{r})$ into a surface integral far outside the sources.

P24.9 For the wire of Example 24.2, use the advanced potentials, with $t + \mathscr{r}/c$ in place of $t - \mathscr{r}/c$.

(a) Find the magnetic field at a distance s from the wire for $t > -s/c$.

(b) Find the electric field for $t < -s/c$.

(c) Explain why these results are unacceptable.

P24.10 The current in the wire of Example 24.2 is instead zero for $t < 0$ and $I = kt$ for $t > 0$.

(a) Find the electric field at a distance s from the wire, and show that it is

$$\mathbf{E} = -\frac{\mu_0 k}{2\pi} \ln\left(\frac{ct + \sqrt{c^2 t^2 - s^2}}{s}\right) \hat{\mathbf{z}}$$

for $t > s/c$.

(b) A real switch raises the current from zero to I_0 steadily over a time T and then holds it at I_0. Write this current as the difference of two ramps and find $\mathbf{E}$. Show that the field stays finite, and find how it behaves for $t \gg T$, long after the news of the end of the ramp has arrived. Why does it not drop to zero at once?

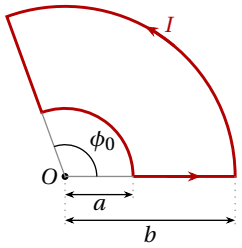


Figure 24.14 The loop of Exercise P24.11, with its arcs of radii a and b and the angle ϕ_0 between its radial sides.

P24.11 A loop of wire consists of two circular arcs, of radii a and b, both centered on the origin and spanning the angle ϕ_0, joined by two radial segments (Figure 24.14). The current in it is $I(t) = kt$.

(a) Find the retarded vector potential at the origin, segment by segment, and show that

$$\mathbf{A}(0, t) = \frac{\mu_0 kt}{4\pi} \ln\left(\frac{b}{a}\right) (\hat{\mathbf{r}}_1 - \hat{\mathbf{r}}_2),$$

where $\hat{\mathbf{r}}_1$ and $\hat{\mathbf{r}}_2$ are the unit vectors along the radial segments, the first traversed outward and the second inward.

(b) Find the electric field at the origin.

(c) The terms that involve c cancel in the sum, whatever ϕ_0 is. Explain why, using the result for closed circuits in Example 24.3.

P24.12 A square loop of wire, of side a and resistance R, lies in a plane with a long straight wire, its near side parallel to the wire and a distance b from it (Figure 24.15). The current in the long wire is I_0 until $t = 0$ and $I_0 e^{-t/\tau}$ afterward.

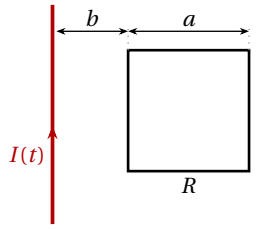


Figure 24.15 The long wire and the square loop of Exercise P24.12.

(a) Assuming the current changes slowly, find the current induced in the loop, and its direction. Ignore the loop's self-inductance.

(b) Now account for the travel time. Set up the electric field at a point a distance s from the long wire as an integral over the wire, with the limits written out. (The retarded vector potential itself diverges for an infinitely long wire, but its time derivative does not.)

(c) When is the answer to part (a) a good approximation?

P24.13 An infinite sheet in the xy plane carries no current until $t = 0$, when a uniform surface current $\mathbf{K} = K_0\,\hat{\mathbf{y}}$ is switched on everywhere at once (Figure 24.16).

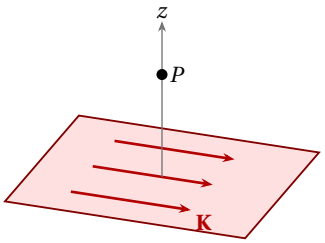


Figure 24.16 The sheet of Exercise P24.13, with its surface current along $+\hat{\mathbf{y}}$, and a field point on the z axis.

(a) Find the retarded potentials at a point $z > 0$ on the z axis.

(b) Find **E** and **B** there, and describe what an observer at that point sees as time goes by.

(c) Repeat for a current that grows steadily, $\mathbf{K} = \alpha t\,\hat{\mathbf{y}}$ for $t > 0$. How do the fields at the front differ from those of part (b)?

Exercises for 24.5 Jefimenko's equations

P24.14 Derive Jefimenko's equation for **B**, (24.16), by taking the curl of the retarded vector potential (24.12) inside the integral.

P24.15 Suppose the current density does not change in time, but the charge density does.

(a) Use the continuity equation to show that ρ changes at a steady rate, $\rho(\mathbf{r}', t) = \rho_0(\mathbf{r}') + \rho_1(\mathbf{r}')\,t$.

(b) Show that the electric field is then given exactly by Coulomb's law with the *present* charge density.

(c) Give a physical example of such a situation.

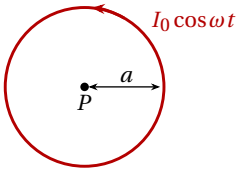


Figure 24.17 The loop of Exercise P24.17, with its alternating current, and the field point at its center.

P24.16 Find the magnetic field of Example 24.2 from Jefimenko's equation (24.16) instead of from the vector potential. The term in $\dot{\mathbf{J}}$ involves a delta function in time; show that it contributes only from the two ends of the segment whose news has arrived, and that the two terms together give (24.14).

P24.17 A circular loop of radius a carries the current $I(t) = I_0 \cos\omega t$ (Figure 24.17).

(a) Use Jefimenko's equation for $\mathbf{B}$ to show that the field at the center of the loop is

$$\mathbf{B} = \frac{\mu_0}{2a}\left[I(t_r) + \frac{a}{c}\dot{I}(t_r)\right]\hat{\mathbf{z}}, \qquad t_r = t - \frac{a}{c}.$$

(b) Write the result as $B_{\text{in}} \cos\omega t + B_{\text{out}} \sin\omega t$, and expand B_{in} and B_{out} in powers of $\omega a/c$. Compare with the estimate of the size of the corrections in Section 16.4.1.

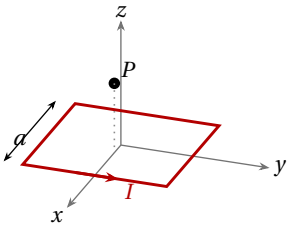


Figure 24.18 The square loop of Exercise P24.18 and the field point P above its plane, off center.

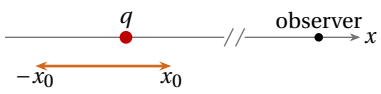


Figure 24.19 The charge of Exercise P24.19, oscillating along the x axis, and the distant observer on that axis.

Exercises for 24.6 Numerical methods: retarded times and integrals

P24.18 A square loop of side a in the xy plane, centered on the origin, carries the current $I(t) = I_0\left(1 - e^{-t/\tau}\right)$ for $t > 0$ and none before (Figure 24.18). Use Jefimenko's equations and numerical quadrature to find $\mathbf{E}$ and $\mathbf{B}$ at the point $(a/4, 0, a/2)$ as functions of time, for $\tau = 10a/c$ and for $\tau = 0.1a/c$. Compare each with the quasistatic fields: the Biot–Savart field of the present current and the induced field $-(\mu_0/4\pi)\oint \dot{I}\, d\boldsymbol{\ell}'/\mathscr{r}$. When do they agree, and why?

P24.19 A charge oscillates along the x axis, $x(t) = x_0 \sin\omega t$, with a peak speed $\omega x_0 = 0.8c$ (Figure 24.19). An observer far away on the positive x axis watches its shadow. Use the method of Section 24.6.1 to find the shadow's position as a function of the time of observation over a few periods, and plot it. Describe how it differs from a sinusoid, and explain the difference.

Exercises for 24.7 Summary: triangle diagram of sources, fields, and potentials

P24.20 True or false? Explain your reasoning.

(a) The retarded scalar potential is the electrostatic potential of the charges, with each piece of charge evaluated at its own retarded time.

(b) The electric field of changing charges is given by Coulomb's law, with each piece of charge evaluated at its own retarded time.

(c) In the Coulomb gauge, the scalar potential at a point responds at once to a change in the charges anywhere, so the electric field does too.

(d) Whatever the potentials, a gauge transformation can always put them in the Lorenz gauge.

(e) The shadow of a moving charge is where the charge is now.

(f) A rod that moves toward you looks longer than it is because of the Lorentz contraction.

(g) Far from changing sources, the terms of Jefimenko's equations that involve $\dot{\rho}$ and $\dot{\mathbf{J}}$ are the important ones.

(h) The advanced potentials do not satisfy the inhomogeneous wave equations.

(i) For a current that has been rising at a steady rate for a long time, the quasistatic magnetic field is exact.

P24.21 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as the retarded times of a charge in an arbitrary motion, the shadows of moving charges and extended objects, the fields of a wire or a loop whose current is switched on, Jefimenko's fields of a changing current distribution, the apparent superluminal motion of a quasar jet, or the relations in the triangle diagram for electrodynamics.

Part VII

Relativity

The introduction of a "luminiferous ether" will prove to be superfluous inasmuch as the view here to be developed will not require an "absolutely stationary space" provided with special properties, nor assign a velocity-vector to a point of the empty space in which electromagnetic processes take place.

— Albert Einstein,
"On the Electrodynamics of Moving Bodies" (1905)*

*"Zur Elektrodynamik bewegter Körper," A. Einstein, Ann. Phys. 17, 891–921 (1905). https://doi.org/10.1002/andp.19053221004. The quotation is from the translation of W. Perrett and G. B. Jeffery, in *The Principle of Relativity* (Methuen, London, 1923), p. 38.

Chapter 25

Basic Relativity

Maxwell's equations predict electromagnetic waves that travel at $c = 1/\sqrt{\mu_0 \varepsilon_0}$ (Section 16.5), and nothing in the equations says what that speed is measured relative to. Every other speed in physics is measured relative to something, so either Maxwell's equations hold only for one special observer, or our ideas about space and time need repair. A second puzzle has already appeared in this book. A magnet moving toward a loop of wire, and a loop moving toward a magnet at the same speed, produce the same EMF (Section 15.2.1), but the explanations are different: in one case a changing magnetic field produces an electric field that pushes the charges in the loop, and in the other there is no electric field at all, only a magnetic force on the moving charges. Is there an electric field or not? The answer, worked out by Albert Einstein in 1905, is that the two descriptions belong to two observers in relative motion, and that each is correct for its observer.

This chapter and the next take up that answer. This chapter is about space and time themselves: how two observers in relative motion assign positions and times to the same events, and what becomes of energy, momentum, and force. Much of it may be familiar from an introductory course, and here it is cast in the form that the following chapters need: spacetime diagrams, the Lorentz transformation as a matrix, and four-vectors. Chapter 26 then answers the question that matters most for electromagnetism: given the fields **E** and **B** that one observer measures, what fields does another observer measure? Chapter 27 uses the answer to find the fields of a moving point charge, which is why this book takes up relativity before radiation.

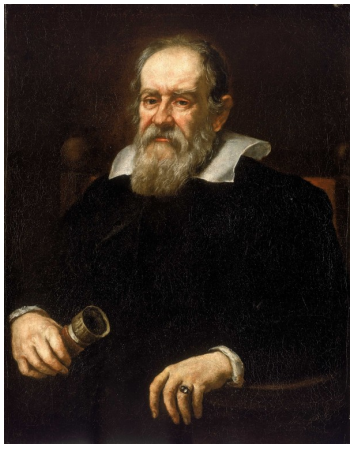
Two observers make most of the measurements in this chapter. Rather than the labels S and S' of many books, they have names: Luke, whom we usually take to be at rest, and Rachel, who moves to the right relative to Luke at speed v. Their names begin with R, for right, and L, for left: Rachel moves to the right as Luke sees it, and Luke moves to the left as Rachel sees it. Quantities that each of them measures carry the subscripts R and L. For example, x_R and t_R are the position and time of an event as Rachel measures them, and x_L and t_L are the position and time of the same event as Luke measures them.

25.1 The postulates

Relativity rests on two **postulates**, statements that are assumed rather than derived and are then tested by their consequences. This section explains where they come from: first the older principle of relativity of Galileo and Newton, then the problem that light posed for it, and finally Einstein's resolution.

25.1.1 Galilean relativity

A **frame of reference** is a way of assigning positions and times to events: a set of axes, with rulers to measure positions and clocks to measure times, all at rest relative to one another. In an **inertial frame**, an object with no net force on it moves in a straight line at constant speed, as Newton's first law says. Any frame that moves at constant velocity relative to an inertial frame is also inertial; a frame that accelerates or rotates is not.



Galileo Galilei (1564–1642, Italian) Galileo was born in Pisa and studied at its university, where he became professor of mathematics in 1589; from 1592 to 1610 he taught at Padua. He studied falling bodies and projectiles with inclined planes and pendulums, and showed that a projectile's path is a parabola. In 1609 he built a telescope and turned it on the sky, where he found mountains on the Moon and four moons circling Jupiter, which he reported in *Sidereus Nuncius* (1610). His *Dialogue Concerning the Two Chief World Systems* (1632), which defended the Copernican system and contains the ship of Section 25.1.1, led to his trial by the Inquisition in 1633, and he spent the rest of his life under house arrest. There he wrote *Two New Sciences* (1638), on the strength of materials and on motion. He died at Arcetri, near Florence, in 1642. (Wikipedia)

Galileo pointed out that no experiment done inside a closed room can tell whether the room is at rest or moving at constant velocity. His example was the cabin of a ship below decks: flies, fish in a bowl, drops falling from a bottle into a jar, and a ball tossed to a friend all behave the same way whether the ship is at anchor or sailing smoothly.[1] This is the **principle of relativity** in its earliest form: the laws of mechanics are the same in every inertial frame. Only relative motion can be detected.

Suppose that Rachel moves in the $+x$ direction at constant speed v relative to Luke, and that she passes him at $t = 0$. If Luke finds an event at position x_L at time t_L, then common sense says that Rachel, who has moved a distance $v t_L$ by then, finds it at $x_L - v t_L$, at the same time:

$$x_R = x_L - v t_L, \qquad y_R = y_L, \qquad z_R = z_L, \qquad t_R = t_L. \tag{25.1}$$

This is the **Galilean transformation**. Differentiating it with respect to time gives the Galilean rule for adding velocities: an object that moves at u_L in Luke's frame moves at $u_R = u_L - v$ in Rachel's, and one that moves at u_R in Rachel's frame moves at

$$u_L = u_R + v \qquad \text{(Galileo)} \tag{25.2}$$

in Luke's. Differentiating again shows that accelerations are the same in both frames. Forces that depend only on the distances between objects are the same too, so Newton's second law, $\mathbf{F} = m\mathbf{a}$, has the same form in every inertial frame: Newtonian mechanics obeys the principle of relativity.

25.1.2 The speed of light

Maxwell's equations do not obey the principle of relativity in this Galilean form. They predict waves that travel at

$$c = \frac{1}{\sqrt{\mu_0 \varepsilon_0}} = 2.998 \times 10^8 \,\text{m/s}$$

[1] G. Galilei, *Dialogue Concerning the Two Chief World Systems* (1632), the Second Day.

with no mention of a frame. If Luke measures a light wave moving to the right at c, then by (25.2) Rachel should measure $c - v$. Exercise P25.8 shows the same thing directly: under the Galilean transformation the wave equation changes its form, and its solutions move at different speeds in the two directions. The magnetic force raises the same question. In $\mathbf{F} = q\mathbf{v} \times \mathbf{B}$, $\mathbf{v}$ is the velocity relative to what? A charge at rest in Luke's frame feels no magnetic force there, but it moves in Rachel's frame, and she would calculate a magnetic force that is not zero.

Before 1905 the usual answer was that light is a wave in a medium that fills all of space, the **ether**, and that Maxwell's equations hold exactly only in the frame in which the ether is at rest. The Earth moves around the Sun at about 30 km/s, so, whatever the ether does, the Earth must move through it at about that speed at some time of the year, and light on Earth should travel at slightly different speeds in different directions. In 1887 Albert Michelson and Edward Morley, in Cleveland, Ohio, looked for the difference.[2] Their interferometer split a beam of light into two beams that traveled at right angles to each other, each about 11 m out and back between mirrors, and then recombined them to form interference fringes. Light going out and back along the direction of the Earth's motion through the ether should take longer than light going out and back across it, by a fraction of about $\beta^2/2$, where β is the Earth's speed through the ether divided by c. For 30 km/s, $\beta = 10^{-4}$, and the fraction is only 5×10^{-9}, but the interferometer could detect it. The whole instrument rested on a stone slab floating on mercury, so that it could be turned slowly while they watched the fringes, and turning it through 90° should have shifted the fringes by about 0.4 of the spacing between them. The shift they saw was "certainly less than the twentieth part of this, and probably less than the fortieth part." As far as they could tell, light traveled at the same speed in every direction. Many later experiments, far more sensitive, have found the same.



Albert A. Michelson (1852–1931, American) Michelson was born in Strelno, Prussia (now Strzelno, Poland), and came to the United States with his family as a small child. He graduated from the U.S. Naval Academy in 1873 and, while an instructor there, measured the speed of light with a rotating mirror (1879). The interferometer he designed for the ether experiment, which he carried out with Morley in 1887, became one of the most precise measuring instruments in physics. From 1892 he led the physics department of the new University of Chicago. In 1907 he became the first American to receive a Nobel Prize in the sciences, "for his optical precision instruments and the spectroscopic and metrological investigations carried out with their aid." In 1920, with Francis Pease, he measured the diameter of the star Betelgeuse with an interferometer on Mount Wilson. He died in Pasadena in 1931. (Wikipedia)

25.1.3 Einstein's postulates

Many explanations of Michelson and Morley's result were proposed. Perhaps the Earth drags the ether along with it; perhaps objects moving through the ether contract along their direction of motion just enough to hide the difference (Section 25.4 returns to that idea). In 1905 Einstein took a different approach.[3] He began with the magnet and the conductor of Section 15.2.1: what can be observed depends only on their relative motion, while the usual description depends on which of them moves. "Examples of this sort," he wrote, "together with the unsuccessful attempts to discover any motion of the earth relatively to the 'light medium,' suggest that the phenomena of electrodynamics as well as of mechanics possess no properties corresponding to the idea of absolute rest." He raised that observation to a postulate and added a second one:

[2] "On the relative motion of the Earth and the luminiferous ether," A. A. Michelson and E. W. Morley, Am. J. Sci. 34, 333–345 (1887). https://doi.org/10.2475/ajs.s3-34.203.333.

[3] "Zur Elektrodynamik bewegter Körper," A. Einstein, Ann. Phys. 17, 891–921 (1905). https://doi.org/10.1002/andp.19053221004. The quotation is from the translation of W. Perrett and G. B. Jeffery (1923).



Edward W. Morley (1838–1923, American) Morley was born in Newark, New Jersey, and studied at Williams College and then at Andover Theological Seminary. He taught chemistry at Western Reserve College, first in Hudson, Ohio, and from 1882 in Cleveland, until he retired in 1906. His best-known chemical work was a precise measurement of the atomic weight of oxygen relative to hydrogen, and he also studied the oxygen content of the atmosphere. Besides the 1887 experiment with Michelson, he later repeated the ether measurement with Dayton Miller. He received the Davy Medal of the Royal Society in 1907 and died in West Hartford, Connecticut, in 1923. (Wikipedia)

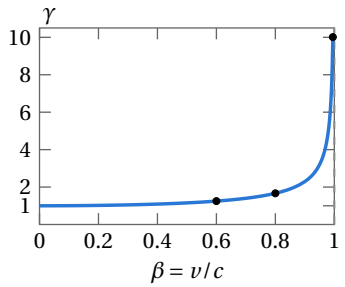


Figure 25.1 The Lorentz factor $\gamma = 1/\sqrt{1-\beta^2}$ against $\beta = v/c$. It stays close to 1 until β is a sizable fraction of 1 and grows without limit as β approaches 1.

Postulate 1 (the principle of relativity). *The laws of physics are the same in every inertial frame.*

Postulate 2 (the constancy of the speed of light). *Light in empty space travels at the same speed c in every inertial frame, whatever the motion of its source.*

The first postulate extends Galileo's principle from mechanics to all of physics, electromagnetism included. The second is then what Maxwell's equations require if they are among the laws of physics, since they predict the speed c without reference to any frame.

Together the two postulates are incompatible with the Galilean transformation. By (25.2), a light pulse that Luke finds moving at c moves at $c - v$ for Rachel, but Postulate 2 says that it moves at c for both. Something in (25.1) has to give, and what gives turns out to be the assumption $t_R = t_L$, that time is the same for every observer. The rest of the chapter works out the consequences. One of them (Section 25.6) is that velocities do not simply add. An object moving at u_R along x in Rachel's frame moves in Luke's frame at

$$u_L = \frac{u_R + v}{1 + u_R v/c^2}.$$

At everyday speeds the denominator is 1 to many decimal places, and (25.2) is an excellent approximation. But if $u_R = c$ then $u_L = c$, as Postulate 2 requires, and if u_R and v are both less than c, so is u_L.

Postulate 1 also settles the question about the magnet and the conductor. Luke, at rest relative to the magnet, says that the charges in the moving loop feel a magnetic force. Rachel, riding with the loop, sees the magnet approach and says that its changing field produces an electric field, which pushes the charges in her loop. If the laws of electromagnetism are the same for both of them, both descriptions are correct, and they must predict the same current. The resolution is that the electric and magnetic fields are not separately the same for all observers: what one observer calls a magnetic field, another calls partly electric. Chapter 26 works out exactly how the fields transform.

Two combinations of v and c appear throughout relativity and have their own symbols,

$$\beta \equiv \frac{v}{c}, \qquad \gamma \equiv \frac{1}{\sqrt{1-\beta^2}}. \tag{25.3}$$

The factor γ is called the **Lorentz factor**. It is 1 when $v = 0$ and grows slowly at first, $\gamma \approx 1 + \beta^2/2$ for small β, so that it is only 1.005 at a tenth of the speed of light. Then it grows rapidly as v approaches c: it is 1.25 at $0.6c$, 5/3 at $0.8c$, and 10 at $0.995c$ (Figure 25.1).

The speed of light is not only the same for all observers: since 1983 it has had a defined value. The meter is now defined as the distance light travels in empty space in 1/299 792 458 of a second,[4] so that $c = 299\,792\,458\,\text{m/s}$ exactly, and

[4] Resolution 1 of the 17th General Conference on Weights and Measures (1983). https://www.bipm.org/committees/cg/cgpm/17-1983/resolution-1.

Postulate 2 is built into the system of units. (Section 17.4 describes the changes of 2019, which also made μ_0 and ε_0 measured quantities.)

25.2 Events, world lines, and spacetime diagrams

An **event** is something that happens at a particular place at a particular time: a firecracker exploding, a flash of light leaving a lamp, two particles colliding, a party at a given house on a given evening. An observer describes an event by four numbers, its position (x, y, z) and its time t.

How does an observer assign those numbers to an event far away? Imagine Luke's frame filled with a lattice of rulers, with a clock at every intersection, all at rest relative to Luke. Each event is recorded by the clock next to it, so the time of the event is when it happened, not when the news of it reached Luke. For this to work, the clocks must agree with one another: they must be **synchronized**. Postulate 2 says how to synchronize them. A flash of light sent from the clock at the origin when it reads t_1 is reflected by a distant clock and returns to the origin at t_2. Since the light travels at c both ways, it reached the distant clock at $(t_1 + t_2)/2$, and that clock is set so that it read $(t_1 + t_2)/2$ when the flash arrived.

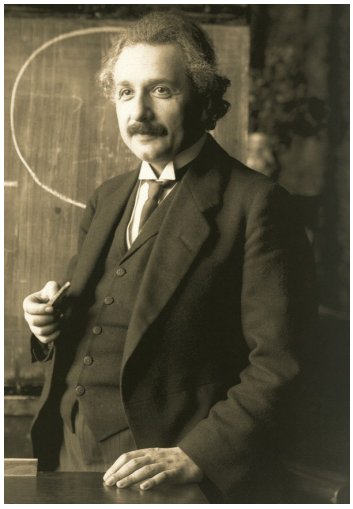
The distinction between when an event happens and when it is seen matters. The shadows of Section 24.3 were about seeing: the news that an observer receives at one instant left different parts of a moving object at different times. The effects of the next sections are different. They remain after every measurement has been corrected for the time light takes to travel, and they are about how events are arranged in space and time, not about how things look.

Events are drawn on a **spacetime diagram**, with time running up the page and one dimension of space, x, running across (Figure 25.2; y and z are left out). The vertical axis is ct rather than t, so that both axes have units of length. In this chapter's examples, distances are usually in light-years (ly) and times in years (yr), so that ct in light-years has the same number as t in years. A light pulse then moves one unit across for every unit up, along a line at 45°.

The history of an object is a line on the diagram, its **world line**. An object at rest relative to the observer has a vertical world line. One that moves at constant speed v has a straight world line, and since it moves a distance $v\Delta t$ while ct increases by $c\Delta t$, the slope of the line is

$$\frac{\Delta(ct)}{\Delta x} = \frac{c}{v} = \frac{1}{\beta}. \tag{25.4}$$

Nothing moves faster than light (Section 25.5.2), so every world line is steeper than 45°. An accelerating object has a curved world line, but its slope is everywhere steeper than 45°. The two 45° lines through an event make its **light cone** (a cone if a second dimension of space is drawn). Events inside the upper half of the cone can be reached from the event by an object or a signal, and events inside the lower half can have influenced it.



Albert Einstein (1879–1955, German-born) Einstein was born in Ulm, grew up in Munich, and graduated from the Swiss Federal Polytechnic in Zürich in 1900. Unable to find an academic post, he worked from 1902 at the Swiss patent office in Bern, where in 1905 he published papers on the light quantum, on Brownian motion, on the electrodynamics of moving bodies, and on the equivalence of mass and energy. He was a professor in Zürich, in Prague, and from 1914 in Berlin, where in 1915 he completed the general theory of relativity. He received the 1921 Nobel Prize in Physics "for his services to Theoretical Physics, and especially for his discovery of the law of the photoelectric effect." In 1933 he left Germany for the Institute for Advanced Study in Princeton, where he died in 1955. (Wikipedia)

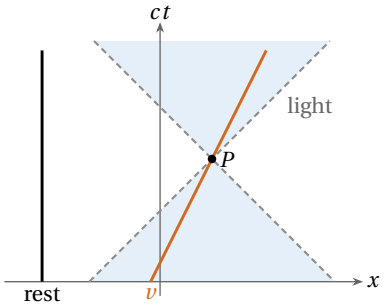


Figure 25.2 A spacetime diagram: the world lines of an object at rest, of an object moving at speed v (slope $1/\beta$), and of light, and the light cone of an event P.

A spacetime diagram is just a graph of position against time with the axes exchanged, and Galileo could have drawn one. What is new in relativity is what happens when two observers draw diagrams of the same events and compare them. The geometric view of relativity, in which space and time together form a four-dimensional **spacetime**, was introduced by Hermann Minkowski in 1908, three years after Einstein's paper.

25.3 Simultaneity, time dilation, and length contraction

Einstein's two postulates have three consequences that run against everyday experience. Each can be found from a thought experiment that uses little more than Postulate 2: light moves at c for every observer. Section 25.4 then shows that all three follow from a single set of equations.

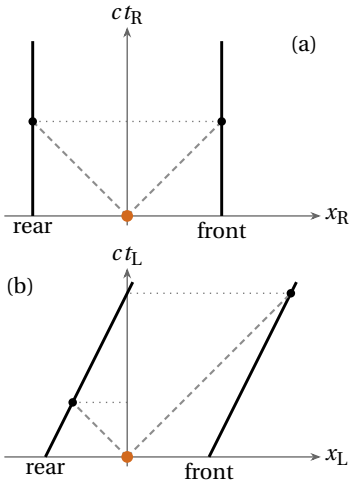


Figure 25.3 A flash at the center of Rachel's train car, drawn for $\beta = 0.5$. (a) In Rachel's frame the car's ends have vertical world lines, and the light (dashed) reaches them at the same time. (b) In Luke's frame the ends move to the right, and the light reaches the rear first.

25.3.1 Simultaneity

Rachel rides in the middle of a train car of length L_0, which moves to the right at speed v relative to Luke, who stands on the platform. She sets off a flash of light at the center of the car. In her frame the car is at rest, the light moves at c toward both ends, and it reaches the two ends at the same moment: the events "the light reaches the rear wall" and "the light reaches the front wall" are simultaneous.

Luke sees the same light moving at c in both directions, by Postulate 2, but in his frame the car moves. The rear wall moves toward the light that is heading for it, and the front wall moves away from its light, so the light reaches the rear wall first (Figure 25.3). In his frame the two events are not simultaneous. If the car's length in Luke's frame is L, the light reaches the rear after $(L/2)/(c+v)$ and the front after $(L/2)/(c-v)$, and with the length $L = L_0/\gamma$ that Section 25.3.3 will find, the difference is

$$\Delta t_L = \frac{L/2}{c-v} - \frac{L/2}{c+v} = \frac{Lv}{c^2-v^2} = \frac{\gamma v L_0}{c^2}. \tag{25.5}$$

This is the **relativity of simultaneity**: two events at different places that are simultaneous for one observer are not simultaneous for another observer moving relative to the first. For Luke, the event at the rear of the car, the end that trails in its motion, happens first.

Which observer is right? Both are. Postulate 1 says that no experiment can single out one inertial frame as the true one, so no frame's judgment of simultaneity is better than another's. And the effect is not about seeing: both observers have corrected for the time light takes to reach them, as Section 25.2 described. Two events at the same place and the same time, on the other hand, such as two particles colliding, are simultaneous for every observer, since everyone agrees that the particles collided.

25.3.2 Time dilation

A **light clock** is a clock built directly on Postulate 2: a pulse of light bounces between two mirrors a distance h apart, and each round trip is one tick (Figure 25.4). Rachel carries one, with the mirrors one above the other, perpendicular to her motion. In her frame one tick takes

$$\Delta\tau = \frac{2h}{c}.$$

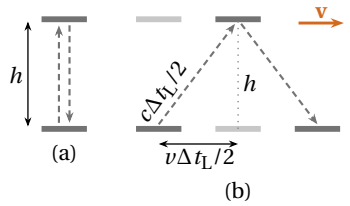


Figure 25.4 A light clock carried by Rachel. (a) In her frame the light goes straight up and down. (b) In Luke's frame the mirrors move to the right while the light is in flight, and the light follows a longer, zigzag path.

In Luke's frame the clock moves to the right at v, and the light follows a zigzag path. During half a tick, $\Delta t_L/2$, the mirrors move a distance $v\,\Delta t_L/2$ while the light crosses from one to the other, so the light travels along the hypotenuse of a right triangle whose sides are h and $v\,\Delta t_L/2$. It travels at c in Luke's frame too, so

$$\left(\frac{c\,\Delta t_L}{2}\right)^2 = h^2 + \left(\frac{v\,\Delta t_L}{2}\right)^2,$$

and solving for Δt_L gives

$$\Delta t_L = \frac{2h/c}{\sqrt{1 - v^2/c^2}} = \gamma\,\Delta\tau.$$

The distance h between the mirrors is the same in both frames, because lengths perpendicular to the relative motion do not change.[5]

Rachel's light clock ticks once in a time $\Delta\tau$ in her frame, but Luke finds each tick lasting $\gamma\,\Delta\tau$: in his frame her clock runs slow. Nothing in the argument depends on how the clock is built. A clock of any other kind that Rachel carries must keep time with her light clock, since otherwise she could detect her own motion, against Postulate 1. So every process in Rachel's frame runs slow in Luke's frame, from the ticking of a wristwatch to her heartbeat to the decay of an unstable particle she carries. This is **time dilation**.

The two ticks of Rachel's clock happen at the same place in her frame, on her clock. The time between two events that happen at the same place in some inertial frame, measured by a clock at rest in that frame, is the **proper time** between them, written τ. Time dilation then reads

$$\boxed{\Delta t = \gamma\,\Delta\tau}, \tag{25.6}$$

where Δt is the time between the same two events in a frame in which the clock moves at speed v, and γ is the Lorentz factor (25.3) of that speed. Since $\gamma \geq 1$, the proper time is the shortest time between the two events that any inertial observer measures.

[5] Suppose they did, and that Luke found Rachel's upright meter stick shorter than his own. By Postulate 1, Rachel would find Luke's stick shorter than hers. Let each stick carry a pen at its top end, and let the sticks pass each other side by side, with their lower ends sliding along the same line. Luke would find that the pen on Rachel's stick marks his stick below its top, and Rachel that the pen on Luke's stick marks hers below its top. But the marks are physical facts that both can inspect afterward, and only one of the two pens can be the lower one.

Time dilation is symmetric: from Rachel's point of view Luke moves at speed v, and his clocks run slow in her frame by the same factor. How can each find the other's clocks slow? Because they measure different things. Luke compares one of Rachel's clocks, which moves past him, with two of his own lattice clocks, the ones beside which her clock starts and ends a measurement. Those two clocks are synchronized in his frame but not in hers (Section 25.3.1). In her frame the second clock was set ahead of the first, which is why it reads more than her own clock when she reaches it. Rachel's measurement of Luke's clock is the mirror image of his, and each is right in her or his own frame. The twin paradox of Section 25.5.4 shows what happens when two clocks that have moved differently are brought back together.

Unstable particles moving close to the speed of light show time dilation as a large effect, as in the following example.

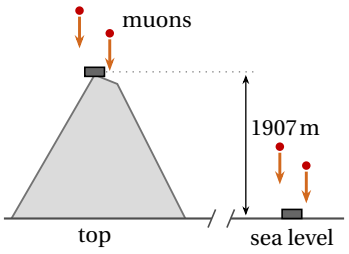
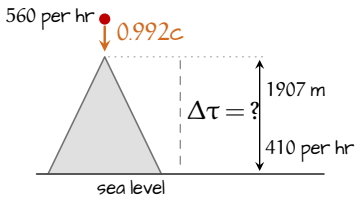


Figure 25.5 The two detectors of Example 25.1, on Mount Washington and at sea level, 1907 m lower, and the muons arriving from above. Not to scale.

Example 25.1

Muons are unstable particles produced high in the atmosphere when cosmic rays strike it. At rest, a muon lives on average $\tau_0 = 2.197\,\mu$s before it decays.[6] David Frisch and James Smith counted the muons that arrived at the top of Mount Washington, New Hampshire, with speeds between $0.9950c$ and $0.9954c$: about 560 per hour. Then, at Cambridge, Massachusetts, 1907 m lower, they counted the muons arriving with the lower speeds that muons of that group have after slowing down in the air: about 410 per hour.[7] The muons slow from about $0.995c$ at the top to about $0.989c$ at sea level, so take their average speed during the descent to be $0.992c$. How many muons per hour would reach sea level if their clocks did not run slow? How much proper time passes for the muons during the descent, and what time-dilation factor does that imply?

Solution:



The first step is a sketch of what is given and what is asked: the muons at the top, the height they fall, their speed, and the rate at which they arrive at each height. The muons are counted at two heights; the question is how many of those at the top survive the trip.

In the Earth's frame the descent takes

$$\Delta t = \frac{1907\,\text{m}}{(0.992)(2.998\times10^8\,\text{m/s})} = 6.41\,\mu\text{s}.$$

Muons decay at random, so the fraction of them that survive a time t, measured on their own clocks, is e^{-t/τ_0}. If their clocks kept the same time as the Earth's, the fraction

[6] "Review of Particle Physics," S. Navas et al. (Particle Data Group), Phys. Rev. D 110, 030001 (2024). https://doi.org/10.1103/PhysRevD.110.030001. The muon's mean life is $(2.1969811 \pm 0.0000022)\,\mu$s.

[7] "Measurement of the relativistic time dilation using μ-mesons," D. H. Frisch and J. H. Smith, Am. J. Phys. 31, 342–355 (1963). https://doi.org/10.1119/1.1969508.

surviving the descent would be

$$e^{-\Delta t/\tau_0} = e^{-6.41/2.197} = e^{-2.92} = 0.054,$$

and of the 560 per hour at the top, about 30 per hour would reach sea level. About 410 per hour arrive, more than ten times as many. The fraction that survives, 410/560, corresponds to a proper time $\Delta\tau$ given by $e^{-\Delta\tau/\tau_0} = 410/560$, or

$$\Delta\tau = \tau_0 \ln\frac{560}{410} = (2.197\,\mu\text{s})(0.312) = 0.69\,\mu\text{s}.$$

The muons' clocks recorded only $0.69\,\mu$s during a descent that took $6.41\,\mu$s in the Earth's frame, a time-dilation factor of about 9. That fits the Lorentz factors of the muons, which fall from $\gamma = 10.2$ at $0.9952c$, at the top, to 6.8 at $0.989c$, at sea level. Frisch and Smith's careful analysis of their data gave 8.8 ± 0.8, compared with 8.4 ± 2 calculated for the speeds of the muons in their apparatus.

25.3.3 Length contraction

Rachel measures the length of her train car with a laser rangefinder: a pulse of light leaves the rear of the car, reflects from a mirror at the front, and returns to the rear (Figure 25.6). In her frame the car has its **rest length** L_0, also called its **proper length**, the light covers a distance $2L_0$, and the round trip takes $2L_0/c$. The events "the pulse leaves" and "the pulse returns" happen at the same place in her frame, at the rear of the car, so $2L_0/c$ is a proper time, and by (25.6) Luke finds that the round trip takes

$$\Delta t_{\text{L}} = \gamma\,\frac{2L_0}{c}.$$

Luke can also work out the round trip from the car's length L in his own frame. On the way out the light chases the front of the car, which moves away at v, so it closes the gap at the rate $c - v$ and takes a time $L/(c-v)$. On the way back the rear moves toward the light, and that leg takes $L/(c+v)$. The total is

$$\Delta t_{\text{L}} = \frac{L}{c-v} + \frac{L}{c+v} = \frac{2Lc}{c^2 - v^2} = \gamma^2\,\frac{2L}{c}.$$

Setting the two expressions equal gives

$$\boxed{L = \frac{L_0}{\gamma}}. \tag{25.7}$$

A moving object is shorter along its direction of motion, by the factor γ, than it is in its own frame. This is **length contraction**, also called the Lorentz contraction. Lengths perpendicular to the motion do not change (Section 25.3.2), so a moving ball is a flattened ellipsoid in Luke's frame.

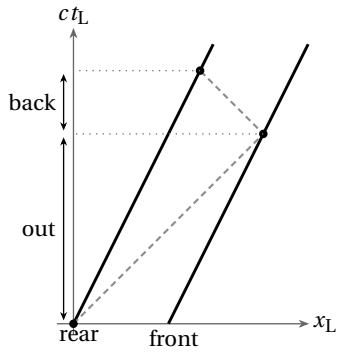


Figure 25.6 Rachel's laser rangefinder in Luke's spacetime diagram, for $\beta = 0.5$: the world lines of the rear and the front of her car, and the pulse of light (dashed) that leaves the rear, reflects from the front, and returns. The trip out takes three times as long as the trip back.

Length contraction is the relativity of simultaneity in another guise. To measure the length of a moving car, Luke must note where its front and its rear are at the same time in his frame. Rachel says that he noted the position of the front first and the position of the rear later, after the rear had moved forward, so that the two marks are closer together than the length of her car. Each observer finds the other's objects contracted, and both are right, because they use different rulers and different meanings of "at the same time." Exercise P25.4 works through a famous example, a pole that does and does not fit in a barn.

In the muons' frame of Example 25.1, the muons are at rest, and the Earth rushes up toward them at $0.992c$. Their clocks run at the normal rate and they live $2.197\,\mu$s on average, but in their frame the distance from the top of Mount Washington down to sea level is not 1907 m. It is contracted by about the factor 9 that the example found, to about 200 m, which goes past in about $0.7\,\mu$s. The two descriptions disagree about why so many muons survive (in one the muons' clocks run slow, in the other the distance they cross is short), but they agree about how many survive, as Postulate 1 requires.

Length contraction, like time dilation, is what remains after the correction for the time light takes to travel. What a camera photographs combines it with the effects of Section 24.3.1: an object moving past at high speed looks rotated rather than contracted (Exercise P25.5), and a sphere still looks round (see the footnote in Section 24.3.1).



Hermann Minkowski (1864–1909, German) Minkowski was born in Aleksotas, near Kaunas, then in the Russian Empire, and grew up in Königsberg, where his family moved when he was about eight. While still a student he shared the 1883 Grand Prize of the Paris Academy of Sciences, for work on writing integers as sums of five squares, with Henry Smith of Oxford. He created the geometry of numbers and taught at Bonn, Königsberg, and the Polytechnic in Zürich, where Einstein was one of his students; from 1902 he was a professor at Göttingen, with his friend David Hilbert. In a lecture in Cologne on 21 September 1908 he presented relativity as the geometry of a four-dimensional spacetime (Section 25.2) and declared that "henceforth space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality" (in the translation of W. Perrett and G. B. Jeffery). He died of appendicitis in Göttingen a few months later, in January 1909. (Wikipedia)

25.4 The Lorentz transformations

Each of the three effects of Section 25.3 came from a separate thought experiment. All three, and much more, follow from the equations that relate the coordinates (ct_L, x_L, y_L, z_L) that Luke assigns to an event to the coordinates (ct_R, x_R, y_R, z_R) that Rachel assigns to the same event. They replace the Galilean transformation (25.1). As before, Rachel moves in the $+x$ direction at speed v relative to Luke, their x axes lie along the same line, and both of them set their clocks to zero at the event where Rachel passes Luke, so that event is at the origin in both frames.

The Lorentz transformation from the two postulates

Distances perpendicular to the motion are the same in both frames (Section 25.3.2), so $y_R = y_L$ and $z_R = z_L$. For the other two coordinates there are six steps.

(1) An object that moves at constant velocity in Luke's frame moves at constant velocity in Rachel's, since Newton's first law holds in both (Postulate 1). Straight world lines must therefore transform into straight world lines, so the transformation is linear, and since the origins coincide it has no constant terms:

$$x_R = Ax_L + Bct_L, \qquad ct_R = Cx_L + Dct_L,$$

with constants A, B, C, D that can depend only on v.

(2) Rachel's own world line, $x_R = 0$, is $x_L = \beta ct_L$ in Luke's frame. Putting these into the first equation gives $0 = (A\beta + B)\,ct_L$ for every t_L, so $B = -\beta A$.

(3) By symmetry, Rachel finds Luke moving to the left at the same speed v, so his world line, $x_L = 0$, is $x_R = -\beta c t_R$ in her frame. With $x_L = 0$ the two equations give $x_R = B c t_L$ and $c t_R = D c t_L$, so $x_R / c t_R = B/D = -\beta$, and $D = A$. So far,

$$x_R = A\left(x_L - \beta c t_L\right), \qquad c t_R = C x_L + A c t_L.$$

(4) The roles of the two observers can be exchanged. Luke moves at speed v in the $-x$ direction relative to Rachel, so by Postulate 1 the inverse transformation has the same form with β replaced by $-\beta$,

$$x_L = A\left(x_R + \beta c t_R\right),$$

with the same A, since A cannot depend on the direction of the motion.

(5) A pulse of light leaves the origin in the $+x$ direction. In Luke's frame $x_L = c t_L$, and by Postulate 2, $x_R = c t_R$ in Rachel's frame. The transformation and its inverse then give

$$c t_R = x_R = A(1-\beta)\, c t_L, \qquad c t_L = x_L = A(1+\beta)\, c t_R,$$

and multiplying the two, $A^2(1-\beta^2) = 1$. Since x_R increases with x_L, A is positive, and $A = 1/\sqrt{1-\beta^2} = \gamma$.

(6) Finally, solve the inverse $x_L = \gamma(x_R + \beta c t_R)$ for $c t_R$, and substitute $x_R = \gamma(x_L - \beta c t_L)$:

$$\beta c t_R = \frac{x_L}{\gamma} - \gamma\left(x_L - \beta c t_L\right) = \gamma \beta c t_L + \left(\frac{1}{\gamma} - \gamma\right) x_L = \gamma \beta c t_L - \gamma \beta^2 x_L,$$

since $1/\gamma - \gamma = (1-\gamma^2)/\gamma = -\gamma\beta^2$. Dividing by β gives $c t_R = \gamma(c t_L - \beta x_L)$.

The result is the **Lorentz transformation**:

$$\boxed{c t_R = \gamma\left(c t_L - \beta x_L\right), \quad x_R = \gamma\left(x_L - \beta c t_L\right), \quad y_R = y_L, \quad z_R = z_L}. \tag{25.8}$$

Written as a matrix equation, it is

$$\begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}_R = \begin{pmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}_L . \tag{25.9}$$

This is why the time coordinate is written ct: with ct in place of t, all four coordinates have units of length, and the matrix is symmetric. (Section 25.7 gives the matrix the name Λ.) The inverse transformation, from Rachel's coordinates to Luke's, is the same with β replaced by $-\beta$:

$$c t_L = \gamma\left(c t_R + \beta x_R\right), \qquad x_L = \gamma\left(x_R + \beta c t_R\right). \tag{25.10}$$

! **The inverse.** To go back from Rachel's coordinates to Luke's, change the sign of β. The Lorentz factor γ stays the same, since it depends only on β^2.

Multiplying the two matrices confirms that one undoes the other, since $\gamma^2(1-\beta^2) = 1$.

Which sign goes with which direction? A rule of thumb: in Luke's spacetime diagram, Rachel's world line leans to the right of vertical (its slope is $1/\beta$), and

in her own diagram it is vertical. Going from Luke's frame to Rachel's therefore turns her world line to the left, counterclockwise, and the matrix that does it has minus signs. Going the other way turns Luke's world line, which leans to the left in Rachel's diagram, to the right, clockwise, and the signs are plus. The order of the two coordinates does not matter: because the matrix is symmetric, the same matrix transforms (x, ct) as it transforms (ct, x).

! **Which sign?** Going to the frame of an observer whose world line leans to the right turns that world line to the left: the matrix has $-\gamma\beta$. Turning a world line to the right takes $+\gamma\beta$.

The transformation looks a lot like a rotation, but not completely. The rotation of axes (8.2) of Section 8.1.1 mixes x and y through $\cos\phi$ and $\sin\phi$, and it preserves $x^2 + y^2$. The Lorentz transformation mixes ct and x, and it preserves $(ct)^2 - x^2$ instead, as Section 25.5.1 shows (Exercise P25.16 makes the comparison exact). At low speeds, $\gamma \approx 1$, and the transformation becomes $x_R \approx x_L - vt_L$ and $t_R \approx t_L - vx_L/c^2$. The last term is v/c times the time light takes to cross the distance x_L, which in everyday situations is tiny, and without it (25.8) becomes the Galilean transformation (25.1).

The three effects of Section 25.3 are now consequences of (25.8) and (25.10), applied to the differences between the coordinates of two events.

- *Simultaneity.* Two events that are simultaneous in Rachel's frame, $\Delta t_R = 0$, a distance Δx_R apart, are separated in Luke's frame by $\Delta t_L = \gamma(\Delta t_R + v\,\Delta x_R/c^2) = \gamma v\,\Delta x_R/c^2$, which is (25.5) with $\Delta x_R = L_0$.
- *Time dilation.* Two ticks of Rachel's clock happen at the same place in her frame, $\Delta x_R = 0$, so $\Delta t_L = \gamma\,\Delta t_R$.
- *Length contraction.* Luke measures the length of Rachel's car by noting where its ends are at the same time in his frame, $\Delta t_L = 0$. Then $\Delta x_R = \gamma(\Delta x_L - \beta c\,\Delta t_L) = \gamma\,\Delta x_L$, and since Δx_R is the car's rest length L_0, its length in Luke's frame is $\Delta x_L = L_0/\gamma$.

In each case the important step is to identify the observer for whom the two events happen at the same place, or at the same time.

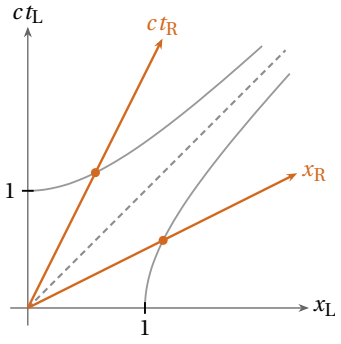


Figure 25.7 Rachel's axes (orange) on Luke's spacetime diagram, for $\beta = 0.5$. Her time axis is her world line, with slope $1/\beta$; her space axis has slope β. The dashed line is light; the gray curves are the hyperbolas $(ct)^2 - x^2 = \pm 1$, which pass through the unit marks on both observers' axes.

It is often useful to draw Rachel's axes on Luke's spacetime diagram (Figure 25.7). Her ct_R axis is her world line, $x_R = 0$, which is the line $x_L = \beta ct_L$, with slope $1/\beta$. Her x_R axis is the set of events with $ct_R = 0$, which by (25.8) is the line $ct_L = \beta x_L$, with slope β. Both axes tilt toward the 45° line of a light pulse, the time axis clockwise and the space axis counterclockwise, by the same angle, and the faster Rachel moves, the more they close in, like the blades of a pair of scissors. The light line itself is the same for both observers. Lines parallel to Rachel's space axis connect events that are simultaneous for her, and lines parallel to her time axis are the world lines of points at rest in her frame. The scales on her axes are not the same as Luke's: the event one unit up her time axis, $(ct_R, x_R) = (1, 0)$, is at $(\gamma, \gamma\beta)$ in Luke's coordinates, farther from the origin than one of his units. Section 25.5.3 explains the hyperbolas in the figure, on which the unit marks lie.

Example 25.2

Rachel moves to the right at $v = 0.5c$ relative to Luke. Four events happen at the same time in Luke's frame, $t_L = 2\,\text{yr}$, at $x_L = 0.5$, 0, -1, and $-2\,\text{ly}$. Where and when do they happen in Rachel's frame? Before calculating, make some guesses. Which of them happens first in her frame? Are they farther apart or closer together than in Luke's?

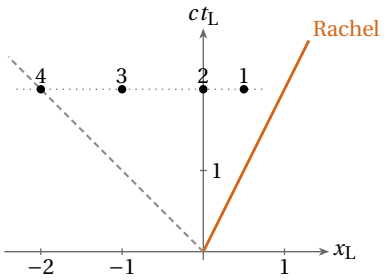
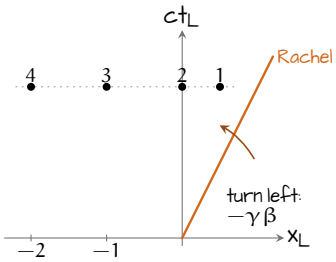


Figure 25.8 The four events of Example 25.2 in Luke's spacetime diagram, and Rachel's world line.

Solution:



The first step is a sketch: Luke's spacetime diagram, with the four events on the line $ct_L = 2\,\text{ly}$, and Rachel's world line, leaning to the right with slope 2. Going to her frame turns her world line to the left, so the matrix has minus signs.

With $\beta = 0.5$, the Lorentz factor is $\gamma = 1/\sqrt{1-0.25} = 1.1547$, and $\gamma\beta = 0.5774$. In light-years, $ct_L = 2$ for all four events, and (25.9) gives, for the first,

$$\begin{pmatrix} ct \\ x \end{pmatrix}_R = \begin{pmatrix} 1.1547 & -0.5774 \\ -0.5774 & 1.1547 \end{pmatrix} \begin{pmatrix} 2 \\ 0.5 \end{pmatrix} = \begin{pmatrix} 2.0207 \\ -0.5774 \end{pmatrix},$$

so it happens at $t_R = 2.0207\,\text{yr}$, at $x_R = -0.5774\,\text{ly}$. The same multiplication for the others gives

$$(ct_R, x_R) = (2.3094, -1.1547), \quad (2.8868, -2.3094), \quad \text{and} \quad (3.4641, -3.4641)\,\text{ly}.$$

The four events are not simultaneous in Rachel's frame (Figure 25.9). The first, the farthest to the right, happens first, and the last happens last. In Rachel's frame Luke and his lattice of clocks move to the left, so the events farther to the right are at the trailing end of his frame, and, as on Rachel's train car in Section 25.3.1, the event at the trailing end happens first.

The events are also farther apart. The second and third are 1 ly apart in Luke's frame, but $2.3094 - 1.1547 = 1.1547\,\text{ly}$ apart in Rachel's. That is not length contraction in reverse: length contraction concerns the two ends of an object located at the same time, and these two events are not simultaneous for Rachel.

The fourth event lies on the light line $x_L = -ct_L$ through the origin in Luke's frame, and on the light line $x_R = -ct_R$ in Rachel's: a light pulse that passes through the origin is a light pulse for both observers, as Postulate 2 requires.

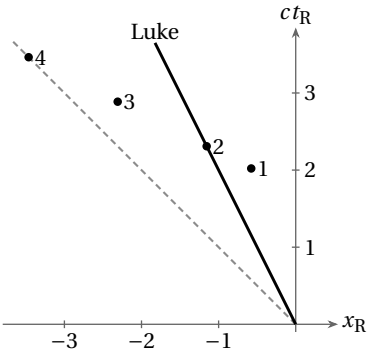


Figure 25.9 The four events of Example 25.2 in Rachel's spacetime diagram, where they are no longer simultaneous, and Luke's world line.

The transformation is named after Hendrik Lorentz, whose portrait is in Section 10.1.2, but it had several authors. Woldemar Voigt wrote down a transformation of this kind in 1887, in a study of the Doppler effect; it kept the wave

equation unchanged and differed from (25.8) by an overall scale factor. To explain Michelson and Morley's result, George FitzGerald (1889) and Lorentz (1892) independently proposed that objects moving through the ether contract along their direction of motion, and Lorentz introduced for moving systems a "local time" (1892–1895) that he regarded as a mathematical aid, not as the time a real clock would show. Joseph Larmor (1900, after an approximate version in 1897) and Lorentz (1904) arrived at transformations equivalent to (25.8) in their efforts to show that Maxwell's equations take the same form for observers in uniform motion through the ether. In 1905 Henri Poincaré completed Lorentz's work, showed that the transformations form a group, and named them after Lorentz; in a longer paper (1906) he showed that $x^2 + y^2 + z^2 - c^2 t^2$ is unchanged by them. Einstein, also in 1905 and independently, derived them from the two postulates alone, with no ether, and read them as statements about the times and lengths that real clocks and rulers measure.



Henri Poincaré (1854–1912, French) Poincaré was born in Nancy, studied at the École polytechnique and the École des mines, and worked briefly as a mining engineer while completing his doctorate in mathematics (1879). From 1881 he taught at the University of Paris. He contributed to nearly every branch of the mathematics and mathematical physics of his time. His prize-winning memoir on the three-body problem of celestial mechanics, written for a competition held by King Oscar II of Sweden, contains the first glimpse of what is now called chaos, and he is regarded as a founder of algebraic topology. In 1904 he stated the principle of relativity as a general law of nature, and in 1905 he gave the Lorentz transformations their modern form, showed that they form a group, and named them. He also wrote widely read books on the philosophy of science. He was elected to the Académie française in 1908 and died in Paris in 1912. (Wikipedia)

25.5 The structure of spacetime

Observers in relative motion disagree about the time between two events and about the distance between them. This section is about what they agree on, and about the consequences: an invariant that plays the role of distance, a division of spacetime into regions that every observer sees the same way, and a rule that keeps causes before their effects.

25.5.1 The invariant interval

Under a rotation of axes the coordinates of a point change, but its distance from the origin does not. The Lorentz transformation has an invariant of its own. For two events separated by $c\Delta t$, Δx, Δy, and Δz in some frame, define the **interval**

$$(\Delta s)^2 \equiv -(c\Delta t)^2 + (\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2. \tag{25.11}$$

Its value is the same in every inertial frame. To see this for a transformation along x, (25.8), note that Δy and Δz do not change, and that

$$\begin{aligned} -(c\Delta t_{\mathrm{R}})^2 + (\Delta x_{\mathrm{R}})^2 &= \gamma^2\left[-\left(c\Delta t_{\mathrm{L}} - \beta\Delta x_{\mathrm{L}}\right)^2 + \left(\Delta x_{\mathrm{L}} - \beta c\Delta t_{\mathrm{L}}\right)^2\right] \\ &= \gamma^2\left(1-\beta^2\right)\left[-(c\Delta t_{\mathrm{L}})^2 + (\Delta x_{\mathrm{L}})^2\right] = -(c\Delta t_{\mathrm{L}})^2 + (\Delta x_{\mathrm{L}})^2, \end{aligned}$$

since the cross terms cancel. The four events of Example 25.2 illustrate it: $-(ct)^2 + x^2$ is -3.75, -4, -3, and $0\,\mathrm{ly}^2$ in both frames. The minus sign in front of the time term is what distinguishes the interval from a distance, and it is why the invariance is not obvious.

The interval can be negative, positive, or zero, so every pair of events falls into one of three classes, the same for every observer:

- **Timelike** separation, $(\Delta s)^2 < 0$. The events are close enough in space, for the time between them, that something moving slower than light can be

present at both. In the frame of such an object both events happen at the same place, and the time between them is the proper time, $\Delta\tau = \sqrt{-(\Delta s)^2}/c$, by (25.11) with $\Delta x = \Delta y = \Delta z = 0$.

- **Spacelike** separation, $(\Delta s)^2 > 0$. Nothing that moves slower than light can be present at both events. There is a frame in which they are simultaneous, and in that frame they are a distance $\sqrt{(\Delta s)^2}$ apart (Exercise P25.10).
- **Lightlike** separation, $(\Delta s)^2 = 0$. Only a pulse of light can be present at both events.

For a single event P, the events with timelike or lightlike separation from it lie inside or on its light cone (Figure 25.10). Those above it are in the **future** of P, and those below it are in the **past** of P. The events with spacelike separation from P lie outside its light cone, in a region sometimes called **elsewhere**. Since the interval is invariant, these regions are the same for every observer.

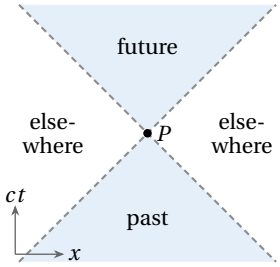


Figure 25.10 The light cone of an event P divides spacetime into its future and its past (timelike separation from P) and elsewhere (spacelike separation). The light cone itself (lightlike separation) is drawn dashed. Every observer agrees on the three regions.

25.5.2 Cause and effect

For timelike and lightlike separations the order of the two events in time is also the same for every observer. By (25.8),

$$\Delta t_{\mathrm{R}} = \gamma\left(\Delta t_{\mathrm{L}} - \frac{v\,\Delta x_{\mathrm{L}}}{c^2}\right).$$

If the separation is timelike or lightlike, $|\Delta x_{\mathrm{L}}| \le c\,|\Delta t_{\mathrm{L}}|$, and since $|v| < c$, the second term in the parentheses is smaller than the first and cannot change its sign. Everyone agrees that one event came before the other. For a spacelike separation, on the other hand, $|\Delta x_{\mathrm{L}}| > c\,|\Delta t_{\mathrm{L}}|$, and an observer moving at $v = c^2\,\Delta t_{\mathrm{L}}/\Delta x_{\mathrm{L}}$, a speed less than c, finds the two events simultaneous. An observer moving faster still finds them in the opposite order.

This is why nothing (no object, no signal, no information) can travel faster than light. If a signal went from event A to event B faster than light, A and B would have spacelike separation, and some observers would find the signal arriving before it was sent. With two such signals, a message could be answered before it was sent (Exercise P25.12). Relativity keeps causes before their effects only if every cause acts within its own light cone. Chapter 24 already obeys the rule: the retarded potentials at a point depend only on what the sources did on the point's past light cone. (Section 25.10 gives a second reason that nothing reaches the speed of light: it would take an infinite amount of energy.)

25.5.3 Reading a spacetime diagram

The interval explains the hyperbolas of Figure 25.7. The events with $-(ct)^2 + x^2 = -1$, in some unit, lie on a hyperbola that opens up and down, and those with $-(ct)^2 + x^2 = +1$ on one that opens left and right. Since the interval is invariant, a Lorentz transformation moves every event along its hyperbola, as a rotation

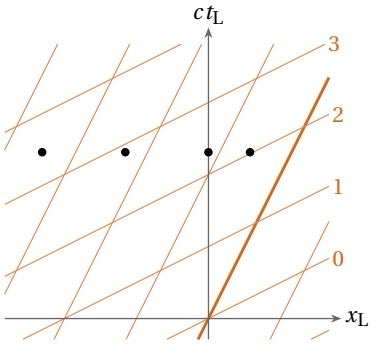


Figure 25.11 Rachel's coordinate grid (orange) on Luke's spacetime diagram, for $\beta = 0.5$: lines of constant x_R parallel to her world line, and lines of constant ct_R, her lines of simultaneity. The four events of Example 25.2 (dots) can be read on either grid. Animation.

moves every point around a circle. In particular, the event one unit up Rachel's time axis, $(ct_R, x_R) = (1,0)$, has $-(ct)^2 + x^2 = -1$, so it lies where her time axis crosses that hyperbola in Luke's diagram, at $(\gamma, \gamma\beta)$. Her unit of length lies where her space axis crosses the other hyperbola. The hyperbolas let one find Rachel's scales directly on Luke's diagram.

With her axes and her scales, Rachel's whole coordinate grid can be drawn on Luke's diagram: lines of constant x_R, parallel to her world line, and lines of constant ct_R, her lines of simultaneity, parallel to her space axis (Figure 25.11). The grid is skewed rather than square, but every event on the diagram can be read in either set of coordinates. The events of Example 25.2, for example, lie on Luke's line $ct_L = 2\,\text{ly}$, which cuts across Rachel's lines of simultaneity, so on her grid they happen at different times. The grid lines are straight because the Lorentz transformation takes straight lines into straight lines (Exercise P25.9). A diagram with two such grids is a graphical Lorentz transformation, and Exercise P25.14 uses one to add velocities. A short program that applies (25.8) to a whole grid of events, and redraws the grid as β changes, is a good way to build intuition for these pictures.

25.5.4 The twin paradox

Because time dilation is symmetric (Section 25.3.2), it raises a famous puzzle, the **twin paradox**. Rachel flies to a distant planet and back while Luke stays home. Luke says that Rachel's clocks run slow, so she should be younger than he is when she returns; Rachel says that Luke's clocks run slow, so he should be younger than she is. When they meet again, standing side by side, they can compare clocks directly, and only one answer can be right.

The resolution is that the situation is not symmetric. Luke stays in one inertial frame the whole time, while Rachel changes frames when she turns around. The time dilation formula (25.6) holds in an inertial frame, and only Luke's frame is inertial throughout. The following example works out the numbers, and shows what Rachel's turnaround does to her judgment of Luke's time.

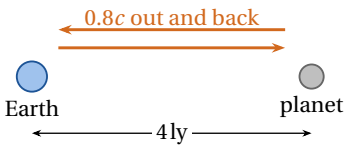
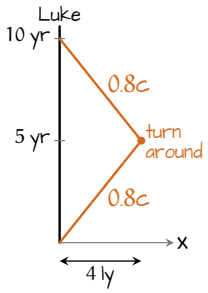


Figure 25.12 The trip of Example 25.3: Rachel travels from Earth to a planet 4 ly away at $0.8c$, and back.

Example 25.3

Rachel leaves Earth, where Luke stays, and travels at $0.8c$ to a planet 4 ly away in Earth's frame. She turns around at once and returns at $0.8c$. (a) How long does the trip take for Luke, and how much does Rachel age? (b) According to Rachel, how much does Luke age during her trip out, and during her trip back? Use her lines of simultaneity to account for the rest of his aging. (c) Luke watches Rachel through a telescope. At what times on his clock does he see her celebrate the birthdays she has during the trip, one each year of her time, the first a year after she leaves?

Solution:



The first step is a sketch of the two world lines: Luke's, straight up, and Rachel's, out to the planet and back, with the turnaround at $x_L = 4\,\text{ly}$. Her two legs are straight lines, but in two different inertial frames.

(a) In Luke's frame each leg is $4\,\text{ly}$ at $0.8c$, which takes $5\,\text{yr}$, so the trip takes $10\,\text{yr}$. Rachel's clocks run slow by $\gamma = 1/\sqrt{1-0.64} = 5/3$, so each leg takes $5\,\text{yr}/(5/3) = 3\,\text{yr}$ of her time, and she ages $6\,\text{yr}$. Rachel agrees: in her frame the distance to the planet is contracted to $4\,\text{ly}/(5/3) = 2.4\,\text{ly}$, which takes $2.4/0.8 = 3\,\text{yr}$ each way. She returns $4\,\text{yr}$ younger than Luke.

(b) On the trip out Rachel moves at $0.8c$ relative to Luke, and in her frame his clocks run slow by the same factor $5/3$: during her $3\,\text{yr}$, Luke ages $3\,\text{yr}/(5/3) = 1.8\,\text{yr}$. On the trip back, likewise, Luke ages $1.8\,\text{yr}$. That accounts for only 3.6 of Luke's $10\,\text{yr}$, and the rest belongs to the turnaround. Just before it, Rachel's line of simultaneity through the turnaround event, $(ct_L, x_L) = (5,4)\,\text{ly}$, has slope $\beta = 0.8$ on Luke's diagram (Figure 25.13a),

$$ct_L = 5 + 0.8\,(x_L - 4)\,,$$

and it crosses Luke's world line, $x_L = 0$, at $ct_L = 1.8\,\text{ly}$. Just before she turns around, Rachel says that Luke's clock reads $1.8\,\text{yr}$. Just after, she moves at $-0.8c$ relative to Luke, her line of simultaneity has slope -0.8,

$$ct_L = 5 - 0.8\,(x_L - 4)\,,$$

and it crosses Luke's world line at $ct_L = 8.2\,\text{ly}$. When she changes frames, her judgment of what Luke's clock reads "now" jumps from 1.8 to $8.2\,\text{yr}$, which accounts for the remaining $6.4\,\text{yr}$: $1.8 + 6.4 + 1.8 = 10$. Rachel's change of frame is not a small detail. It is where the asymmetry lies.

(c) On the trip out, Rachel's birthdays happen at her proper times $\tau = 1$, 2, and $3\,\text{yr}$. In Luke's frame these events are at $t_L = \gamma\tau = 5\tau/3$, at $x_L = 0.8\,ct_L = 4\tau/3$ (in years and light-years). Light from each one takes a further x_L/c to reach Luke, so he sees it at

$$t = \frac{5\tau}{3} + \frac{4\tau}{3} = 3\tau,$$

at 3, 6, and $9\,\text{yr}$. On the trip back her birthdays are at $\tau = 3 + \tau'$, with $\tau' = 1$, 2, and $3\,\text{yr}$, which are the events $t_L = 5 + 5\tau'/3$, $x_L = 4 - 4\tau'/3$, and he sees them at

$$t = 5 + \frac{5\tau'}{3} + 4 - \frac{4\tau'}{3} = 9 + \frac{\tau'}{3},$$

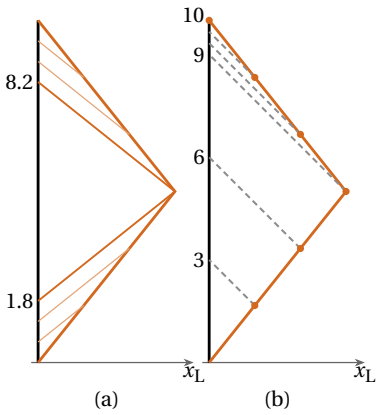


Figure 25.13 The twin paradox of Example 25.3 in Luke's spacetime diagram. (a) Rachel's lines of simultaneity (orange) at her birthdays and just before and after her turnaround: they cross Luke's world line at 1.8 and $8.2\,\text{yr}$. (b) Light (dashed) from her six birthdays, which reaches Luke at 3, 6, 9, $9\frac{1}{3}$, $9\frac{2}{3}$, and $10\,\text{yr}$.

at $9\frac{1}{3}$, $9\frac{2}{3}$, and 10 yr, the last as she arrives (Figure 25.13b). Through the telescope, Luke sees Rachel's clock run at a third of his rate for 9 yr and at three times his rate for the last year. These factors combine time dilation with the changing distance that the light must travel; they are the relativistic Doppler factors of Chapter 26. Rachel, watching Luke, sees the same factors, but she sees his clock run at a third of her rate only during her 3 yr out, and at three times her rate during her 3 yr back: she sees $1+9=10$ yr of his life go by, in agreement with (a).

25.6 Velocity addition

Rachel throws a ball, which moves at velocity u_R along x in her frame. If she throws it from the origin, its world line in her frame is $x_R = u_R t_R$. By (25.10), the ball's time and position in Luke's frame are

$$ct_L = \gamma\left(ct_R + \beta x_R\right) = \gamma c t_R\left(1+\frac{\beta u_R}{c}\right), \qquad x_L = \gamma\left(x_R + \beta c t_R\right) = \gamma c t_R\left(\frac{u_R}{c}+\beta\right),$$

so its velocity in Luke's frame, $u_L = x_L/t_L$, is

$$\boxed{u_L = \frac{u_R + v}{1 + u_R v/c^2}}. \tag{25.12}$$

This is Einstein's rule for **velocity addition**, stated in Section 25.1.3. The inverse, found by replacing v with $-v$, is $u_R = (u_L - v)/(1 - u_L v/c^2)$. Three properties are worth noting:

- For u_R and v small compared with c, the denominator is close to 1, and (25.12) becomes the Galilean rule $u_L = u_R + v$.
- If $u_R = c$, then $u_L = (c+v)/(1+v/c) = c$: light moves at c in both frames.
- If u_R and v are both less than c in magnitude, so is u_L, since

$$c - u_L = \frac{(c-u_R)(c-v)}{c + u_R v/c} > 0,$$

and similarly $c + u_L > 0$. No combination of speeds below c adds up to c or more (Figure 25.14).

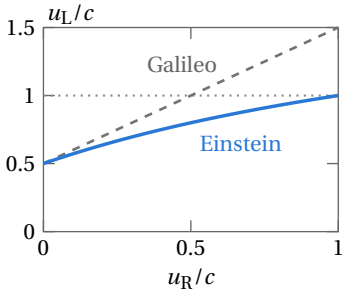


Figure 25.14 A ball thrown at u_R by Rachel, who moves at $v = 0.5c$ relative to Luke: its speed u_L in Luke's frame by Einstein's rule (solid) and by Galileo's (dashed). Einstein's never reaches c for $u_R < c$.

A component of velocity perpendicular to the motion of the frames transforms differently. Distances in the y direction are the same in both frames, but times are not, so for a ball with velocity components $u_{R,x}$ and $u_{R,y}$ in Rachel's frame,

$$u_{L,y} = \frac{u_{R,y}}{\gamma\left(1 + u_{R,x} v/c^2\right)} \tag{25.13}$$

(Exercise P25.13). The x component still obeys (25.12).

Light emitted at an angle θ_R to the x axis in Rachel's frame has $u_{R,x} = c\cos\theta_R$ and $u_{R,y} = c\sin\theta_R$. By (25.12) and (25.13), it travels in Luke's frame at the angle θ_L given by

$$\cos\theta_L = \frac{\cos\theta_R + \beta}{1 + \beta\cos\theta_R}, \qquad \sin\theta_L = \frac{\sin\theta_R}{\gamma\left(1 + \beta\cos\theta_R\right)}. \tag{25.14}$$

This change of direction is the **aberration** of light. Light that a source emits equally in all directions in its own frame is crowded forward in a frame in which the source moves fast. The half of it that goes forward in the source's frame, $\theta_R < 90°$, goes into a cone around the direction of motion with $\cos\theta_L = \beta$, or $\sin\theta_L = 1/\gamma$. At $\beta = 0.9$ the half-angle of the cone is 26°, and at $\gamma = 1000$ it is about a milliradian. This **headlight effect** makes a fast source much brighter when seen from in front, and it explains why the jets of some galaxies are bright only when they point nearly at us (Exercise P25.15). Section 27.6 meets it again in the radiation of fast charges.

Example 25.4

Rachel passes Luke at $0.7c$, and as she passes she throws a ball ahead of her at $0.6c$ in her frame. The ball is caught at a marker that Luke has set up 1 ly away, at rest in his frame. (a) How fast does the ball move in Luke's frame, and how long does it take to reach the marker? (b) How much time passes on a clock carried by the ball? (c) Where and when is the ball caught in Rachel's frame? (d) Rachel reasons that Luke's 1 ly is contracted in her frame to $1\,\text{ly}/\gamma = 1\,\text{ly}/1.400 = 0.714\,\text{ly}$, so the ball travels 0.714 ly in her frame. What is wrong with that reasoning?

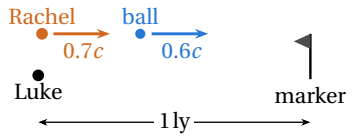
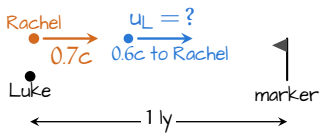


Figure 25.15 Example 25.4: Rachel passes Luke at $0.7c$ and throws a ball ahead at $0.6c$ in her frame; it is caught at a marker 1 ly from Luke, at rest in his frame.

Solution:



The first step is a sketch of what is given and what is asked: Luke and his marker 1 ly apart, Rachel passing Luke at $0.7c$, and the ball leaving her at $0.6c$ relative to her. The ball's speed in Luke's frame is not $1.3c$.

(a) By (25.12),

$$u_L = \frac{0.6c + 0.7c}{1 + (0.6)(0.7)} = \frac{1.3c}{1.42} = 0.9155c,$$

less than c, as it must be, and the trip takes $t_L = 1\,\text{ly}/0.9155c = 1.092\,\text{yr}$.

(b) A clock carried by the ball measures the proper time of the trip. With $\gamma_{\text{ball}} = 1/\sqrt{1 - 0.9155^2} = 2.486$, it records $1.092\,\text{yr}/2.486 = 0.439\,\text{yr}$.

(c) The catch is the event $(c\,t_L, x_L) = (1.092, 1)\,\text{ly}$. With $\beta = 0.7$ and $\gamma = 1/\sqrt{1 - 0.49} = 1.400$, (25.8) gives

$$c\,t_R = 1.400\,[1.092 - (0.7)(1)] = 0.549\,\text{ly},$$

$$x_R = 1.400\,[1 - (0.7)(1.092)] = 0.330\,\text{ly}.$$

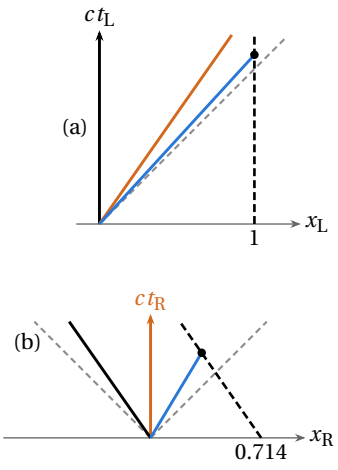


Figure 25.16 The ball of Example 25.4 (blue) in (a) Luke's spacetime diagram and (b) Rachel's. The marker (dashed) is at rest 1 ly from Luke in his frame; in Rachel's frame it starts 0.714 ly away and moves toward her at $0.7c$. Light lines are gray.

In Rachel's frame the ball goes 0.330 ly in 0.549 yr, at $0.330/0.549 = 0.600c$, the speed at which she threw it, which checks the arithmetic. The proper time of the ball's trip, found in her frame with $\gamma = 1.25$ for $0.6c$, is $0.549\,\text{yr}/1.25 = 0.439\,\text{yr}$, the same as in (b): observers agree about proper times.

(d) The 1 ly is the distance between Luke and his marker in Luke's frame, where both are at rest, and in Rachel's frame that distance is indeed contracted to 0.714 ly. But in her frame the marker moves toward her at $0.7c$, so it comes partway to meet the ball (Figure 25.16). The marker is 0.714 ly away at $t_R = 0$, and at $t_R = 0.549\,\text{yr}$ it is at $0.714 - (0.7)(0.549) = 0.330\,\text{ly}$, where the ball reaches it. Length contraction gives the length of an object, located at one time in a given frame. The distance the ball travels in Rachel's frame is a different quantity: it is the distance between the start and the catch, which happen at different times.

25.7 Four-vectors

The Lorentz transformation acts on the coordinates (ct, x, y, z) of an event much as a rotation of the axes acts on the coordinates (x, y, z) of a point. Section 8.1 defined a vector by that behavior: three quantities that change under a rotation of the axes exactly as the coordinates do. The same idea, carried over to spacetime, gives the four-vectors of relativity. They are the natural language for the rest of this chapter and for Chapter 26, because an equation between four-vectors has the same form in every inertial frame, as Postulate 1 requires.

The coordinates of an event are written with an index,

$$x^\mu = \begin{pmatrix} x^0 \\ x^1 \\ x^2 \\ x^3 \end{pmatrix} = \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}, \tag{25.15}$$

where the superscript μ is a label, not a power, and runs from 0 to 3. Greek indices run over all four components, and Latin indices $(i, j, \ldots)$ over the three components of space, 1 to 3, as in Section 8.2. The matrix of (25.9) is called Λ, and its entry in row μ and column ν, with the rows and columns numbered from 0, is written $\Lambda^\mu{}_\nu$. For Rachel moving along x, $\Lambda^0{}_0 = \Lambda^1{}_1 = \gamma$, $\Lambda^0{}_1 = \Lambda^1{}_0 = -\gamma\beta$, $\Lambda^2{}_2 = \Lambda^3{}_3 = 1$, and the other entries are zero. The Lorentz transformation is then

$$x_R^\mu = \sum_{\nu=0}^{3} \Lambda^\mu{}_\nu \, x_L^\nu .$$

A **four-vector** is any set of four quantities a^μ that transforms in the same way,

$$a_R^\mu = \sum_{\nu=0}^{3} \Lambda^\mu{}_\nu \, a_L^\nu . \tag{25.16}$$

Since the transformation is linear, the difference Δx^μ of the coordinates of two events is a four-vector, and so are the sum of two four-vectors and a four-vector multiplied by a **Lorentz scalar**, a quantity that has the same value in every inertial frame. (A Lorentz scalar is also called an invariant.) The following sections find several more four-vectors.

25.7.1 The scalar product

A rotation leaves the squared length $x^2 + y^2 + z^2$ of a vector unchanged. Its counterpart for the Lorentz transformation is the interval (25.11), in which the time term has the opposite sign. To build that sign into the notation, a four-vector is also written with its index down, by changing the sign of its time component:

$$a_\mu = \left(-a^0,\ a^1,\ a^2,\ a^3\right). \tag{25.17}$$

It helps to think of a^μ as a column, as in (25.15), and of a_μ as a row. The two forms are often called the *contravariant* and *covariant* components of the four-vector, names this book will not need again.[8] The **scalar product** of two four-vectors is the row times the column,

$$\sum_{\mu=0}^{3} a_\mu b^\mu = -a^0 b^0 + a^1 b^1 + a^2 b^2 + a^3 b^3. \tag{25.18}$$

For $a^\mu = b^\mu = \Delta x^\mu$ it is the interval, $(\Delta s)^2 = -(c\,\Delta t)^2 + (\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2$, which Section 25.5.1 showed has the same value in every inertial frame. The same algebra shows that the scalar product of any four-vector with itself is invariant, since every four-vector transforms as Δx^μ does, and Exercise P25.17 extends the proof to the scalar product of two different four-vectors: it too is a Lorentz scalar. As for intervals, a four-vector is called timelike, spacelike, or lightlike as its scalar product with itself is negative, positive, or zero.

25.7.2 The summation convention

In (25.18) the index μ appears twice, once up and once down, and is summed over. Sums like this come up so often that the summation sign is usually left out. In the **summation convention**, introduced by Einstein, an index that appears twice in one term, once up and once down, is summed from 0 to 3. (Earlier chapters, such as Chapter 8, wrote their sums out.) So

$$a_\mu b^\mu \equiv \sum_{\mu=0}^{3} a_\mu b^\mu \qquad \text{and} \qquad a_{\mathrm{R}}^\mu = \Lambda^\mu{}_\nu\, a_{\mathrm{L}}^\nu,$$

[8] More advanced treatments write the lowering of an index as a matrix multiplication, $a_\mu = \sum_\nu g_{\mu\nu} a^\nu$, with the *metric* $g_{\mu\nu} = \mathrm{diag}(-1,1,1,1)$; changing the sign of the time component is the same thing. Some books use the opposite signature, $g_{\mu\nu} = \mathrm{diag}(1,-1,-1,-1)$, in which the time component keeps its sign and the space components change theirs, so that their scalar products are the negatives of this book's. None of these choices changes any physics as long as you are consistent within a convention, so just be sure to know what convention is being used by a given reference.

the second with a sum over ν. A summed index is a dummy index and can be renamed, $a_\mu b^\mu = a_\nu b^\nu$. An index that is not summed, like μ in $\Lambda^\mu{}_\nu a^\nu$, must appear on both sides of an equation, in the same position, up or down. These two rules catch many mistakes.

! **Up and down.** A lower index means that the time component has changed sign: $a_\mu = (-a^0, a^1, a^2, a^3)$. A repeated index, once up and once down, is summed from 0 to 3.

25.7.3 Four-tensors

The idea extends to quantities with more indices, as it did for rotations in Section 8.2.2. A **four-tensor** of rank 2 is a set of 16 quantities $t^{\mu\nu}$ that transforms with one factor of Λ for each index,

$$t_{\mathrm{R}}^{\mu\nu} = \Lambda^\mu{}_\lambda \Lambda^\nu{}_\sigma \, t_{\mathrm{L}}^{\lambda\sigma}, \tag{25.19}$$

with sums over λ and σ, the spacetime version of (8.4). The products $a^\mu b^\nu$ of the components of two four-vectors form such a tensor (Exercise P25.18). An index of a tensor is lowered as an index of a four-vector is: every component with a 0 in that index changes sign. Chapter 26 shows that the electric and magnetic fields together form a rank-2 four-tensor, which is why neither of them can be part of a four-vector.

One more property of Λ is needed there. Its determinant, for a transformation along x, is $\gamma^2 - \gamma^2\beta^2 = 1$. The determinant of a linear transformation is the factor by which it changes volumes, so a small region of spacetime has the same four-dimensional volume in every inertial frame:

$$d^4x \equiv dx^0\,dx^1\,dx^2\,dx^3 = c\,dt\,dx\,dy\,dz \qquad \text{(invariant)} \tag{25.20}$$

25.8 Proper time and four-velocity

From here on, $\mathbf{u}$ is the velocity of an object, and v is the speed of one frame relative to another, as before. The Lorentz factors of the two speeds are written $\gamma_u = 1/\sqrt{1-u^2/c^2}$ and $\gamma_v = 1/\sqrt{1-v^2/c^2}$, to keep them apart.

The ordinary velocity $\mathbf{u} = d\mathbf{r}/dt$ of an object is not part of a four-vector. The change dx^μ along the object's world line is a four-vector, but dt is not a Lorentz scalar, so their ratio does not transform by Λ: its time component is $dx^0/dt = c$ in every frame, and its space components transform by the velocity-addition rule (25.12). The points of a world line transform by the Lorentz transformation, but its slope does not. The remedy is to divide by a time that is the same for every observer: the time $d\tau$ that passes on a clock carried by the object, its **proper time** (Section 25.3.2). By time dilation (25.6) it is

$$d\tau = \frac{dt}{\gamma_u} = \sqrt{1-\frac{u^2}{c^2}}\,dt, \tag{25.21}$$

and since $c^2\,d\tau^2 = c^2\,dt^2 - dx^2 - dy^2 - dz^2 = -dx_\mu dx^\mu$, it is invariant. The **four-**

velocity is

$$\boxed{\eta^\mu \equiv \frac{dx^\mu}{d\tau} = \gamma_u \begin{pmatrix} c \\ u_x \\ u_y \\ u_z \end{pmatrix}}, \tag{25.22}$$

since $dx^\mu/d\tau = \gamma_u\, dx^\mu/dt$. It is a four-vector, because dx^μ is one and $d\tau$ is a scalar, so it transforms as $\eta^\mu_{\rm R} = \Lambda^\mu{}_\nu \eta^\nu_{\rm L}$.

The space part of η^μ, $\gamma_u \mathbf{u}$, is not the velocity of the object. It is the distance the object covers, measured in the observer's frame, per unit of the object's own time. Both rates matter to a traveler. To know whether she will arrive on time by the clocks at her destination, she needs u; to know whether she will be hungry when she gets there, she needs $\gamma_u u$, which counts the time her own body experiences. At low speeds the two are the same, and $\eta^\mu \approx (c, \mathbf{u})$.

The scalar product of the four-velocity with itself is

$$\eta_\mu \eta^\mu = \gamma_u^2 \left(-c^2 + u^2\right) = -c^2 \tag{25.23}$$

for every object, in every frame. That had to be so: in the object's own frame $\eta^\mu = (c, 0, 0, 0)$, and scalar products are invariant. Exercise P25.19 checks it with numbers in three frames.

Transforming the four-velocity is a quick way to transform velocities. Suppose that an object moves along x at $u_{\rm R}$ in Rachel's frame. The time component of the inverse transformation, $\eta^0_{\rm L} = \gamma_v(\eta^0_{\rm R} + \beta\eta^1_{\rm R})$, gives

$$\gamma_{u,{\rm L}} = \gamma_v \gamma_{u,{\rm R}} \left(1 + \frac{u_{\rm R} v}{c^2}\right), \tag{25.24}$$

where $\gamma_{u,{\rm L}}$ and $\gamma_{u,{\rm R}}$ are the object's Lorentz factors in the two frames, and the ratio of the space component to the time component, $\eta^1_{\rm L}/\eta^0_{\rm L} = u_{\rm L}/c$, gives the velocity-addition rule (25.12) again.

25.9 Four-acceleration and the instantaneous rest frame

The **four-acceleration** is the derivative of the four-velocity with respect to proper time,

$$\alpha^\mu \equiv \frac{d\eta^\mu}{d\tau} = \frac{d^2 x^\mu}{d\tau^2},$$

a four-vector for the same reason that η^μ is one. To find its components in terms of the ordinary velocity $\mathbf{u}$ and acceleration $\mathbf{a} = d\mathbf{u}/dt$, use $d/d\tau = \gamma_u\, d/dt$ and the derivative of the Lorentz factor,

$$\frac{d\gamma_u}{dt} = \frac{d}{dt}\left(1 - \frac{\mathbf{u}\cdot\mathbf{u}}{c^2}\right)^{-1/2} = \gamma_u^3 \frac{\mathbf{u}\cdot\mathbf{a}}{c^2}.$$

Then $\alpha^0 = \gamma_u\, d(\gamma_u c)/dt$ and the space part is $\gamma_u\, d(\gamma_u \mathbf{u})/dt$, or

$$\alpha^0 = \gamma_u^4 \frac{\mathbf{u}\cdot\mathbf{a}}{c}, \qquad \alpha^{1,2,3} = \gamma_u^2 \mathbf{a} + \gamma_u^4 \frac{(\mathbf{u}\cdot\mathbf{a})\,\mathbf{u}}{c^2}. \tag{25.25}$$

At low speeds $\alpha^\mu \approx (0, \mathbf{a})$. Differentiating (25.23) with respect to τ gives $2\eta_\mu \alpha^\mu = 0$: the scalar product of the four-acceleration with the four-velocity is always zero.

An accelerating object is not at rest in any one inertial frame, but at each instant there is an inertial frame in which it is momentarily at rest, the frame that moves with the object's velocity at that instant. This frame is the object's **instantaneous rest frame**, or IRF. A moment later the object has a different velocity, and a different instantaneous rest frame. In the IRF $\mathbf{u} = 0$ and $\gamma_u = 1$, so by (25.25) $\alpha^\mu = (0, \mathbf{a}_{\text{proper}})$, where $\mathbf{a}_{\text{proper}}$, the acceleration measured in the IRF, is called the **proper acceleration**. It is the acceleration that the object itself feels: what an accelerometer it carries reads, or the number of g's an astronaut reports. Since $\alpha_\mu \alpha^\mu$ is invariant, in any frame it is the square of the proper acceleration,

$$\alpha_\mu \alpha^\mu = a^2_{\text{proper}} = \gamma_u^4 a^2 + \gamma_u^6 \frac{(\mathbf{u}\cdot\mathbf{a})^2}{c^2} = \gamma_u^6 \left(a^2 - \frac{|\mathbf{u}\times\mathbf{a}|^2}{c^2} \right), \tag{25.26}$$

where the last form uses $|\mathbf{u}\times\mathbf{a}|^2 = u^2a^2 - (\mathbf{u}\cdot\mathbf{a})^2$.

The four-acceleration also tells how accelerations transform. Let an object move along x in Luke's frame, and at some instant let its speed be v, so that Rachel's frame, which moves at v, is the object's IRF at that instant, and $\gamma_u = \gamma_v = \gamma$. Suppose first that the acceleration is along the motion, in the x direction. In Luke's frame, by (25.25), $\alpha^\mu_{\text{L}} = (\gamma^4 \beta a_{\text{L},x},\ \gamma^4 a_{\text{L},x},\ 0,\ 0)$, since $\gamma^2 + \gamma^4\beta^2 = \gamma^4$; in Rachel's frame, $\alpha^\mu_{\text{R}} = (0,\ a_{\text{proper},x},\ 0,\ 0)$. The x component of the inverse transformation, $\alpha^1_{\text{L}} = \gamma(\alpha^1_{\text{R}} + \beta\alpha^0_{\text{R}})$, then gives $\gamma^4 a_{\text{L},x} = \gamma a_{\text{proper},x}$. If instead the acceleration is across the motion, in the y direction, then $\mathbf{u}\cdot\mathbf{a} = 0$, $\alpha^\mu_{\text{L}} = (0, 0, \gamma^2 a_{\text{L},y}, 0)$, and $\alpha^\mu_{\text{R}} = (0, 0, a_{\text{proper},y}, 0)$, and since the y component of a four-vector does not change, $\gamma^2 a_{\text{L},y} = a_{\text{proper},y}$. Together,

$$a_{\text{L},x} = \frac{a_{\text{proper},x}}{\gamma^3}, \qquad a_{\text{L},y} = \frac{a_{\text{proper},y}}{\gamma^2}, \tag{25.27}$$

for an object moving along x with Lorentz factor γ in Luke's frame. An object with a constant proper acceleration has a smaller and smaller acceleration in Luke's frame as it speeds up, as the following example shows. Section 27.5 uses both rules to find the power radiated by a fast charge.

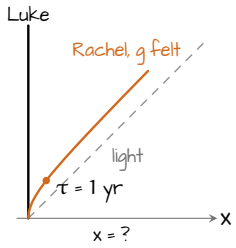


Figure 25.17 Example 25.5: Rachel's ship starts at rest at Earth and accelerates away with the proper acceleration g, felt on board. Not to scale.

Example 25.5

Rachel's spaceship leaves Earth, where Luke stays, starting from rest and accelerating in a straight line with a constant proper acceleration $g = 9.80\,\text{m/s}^2$, so that life on board feels normal. (a) Find her speed in Luke's frame as a function of his time t. (b) Find Luke's time t and the distance x she has traveled in his frame when her own clock reads τ. (c) How far has she gone after 1 yr and after 10 yr of her own time, and how much time has passed on Earth?

Solution:



The first step is a sketch of Luke's spacetime diagram: his world line straight up, and Rachel's leaving it vertically, since she starts at rest, and bending toward the light line as she speeds up, without reaching it.

(a) At each instant Rachel's ship moves along x, and her proper acceleration is along the motion, so by (25.27) its acceleration in Luke's frame is g/γ^3. Writing her speed as $u = \beta c$,

$$c\frac{d\beta}{dt} = \frac{g}{\gamma^3} = g\left(1-\beta^2\right)^{3/2}.$$

Separating the variables and integrating from the start, where $\beta = 0$ at $t = 0$,

$$\int_0^\beta \frac{d\beta'}{\left(1-\beta'^2\right)^{3/2}} = \frac{\beta}{\sqrt{1-\beta^2}} = \frac{gt}{c},$$

so $\gamma\beta = gt/c$, and

$$\beta = \frac{gt/c}{\sqrt{1+(gt/c)^2}}, \qquad \gamma = \sqrt{1+\left(\frac{gt}{c}\right)^2}.$$

At first $\beta \approx gt/c$, as in Newtonian mechanics, but β never reaches 1, while Newtonian mechanics, with $u = gt$, would have Rachel passing the speed of light in less than a year.

(b) Her own time runs slow by γ in Luke's frame, $d\tau = dt/\gamma$, so

$$\tau = \int_0^t \frac{dt'}{\sqrt{1+(gt'/c)^2}} = \frac{c}{g}\sinh^{-1}\frac{gt}{c}, \qquad \text{or} \qquad t = \frac{c}{g}\sinh\frac{g\tau}{c}.$$

The distance is the integral of her speed:

$$x = \int_0^t \beta c\,dt' = \frac{c^2}{g}\left[\sqrt{1+\left(\frac{gt}{c}\right)^2} - 1\right] = \frac{c^2}{g}\left(\cosh\frac{g\tau}{c} - 1\right).$$

Since $(x + c^2/g)^2 - (ct)^2 = (c^2/g)^2$, her world line is a hyperbola in Luke's spacetime diagram (Figure 25.18), and motion with a constant proper acceleration is called hyperbolic motion.

(c) In light-years and years, $c = 1\,\text{ly/yr}$ and

$$g = (9.80\,\text{m/s}^2)\,\frac{(3.156\times10^7\,\text{s/yr})^2}{9.461\times10^{15}\,\text{m/ly}} = 1.032\,\text{ly/yr}^2,$$

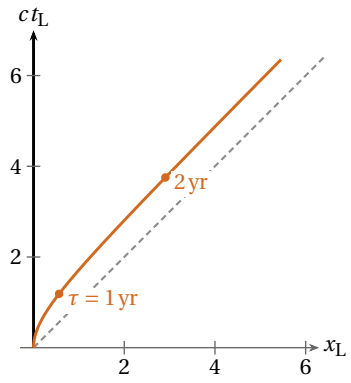


Figure 25.18 Rachel's world line (orange) in Luke's spacetime diagram, for a constant proper acceleration $g = 9.80\,\text{m/s}^2$ from rest (Example 25.5), in light-years and years. Dots mark her proper times $\tau = 1$ and 2 yr. The dashed line is light that leaves Earth with her.

so $c/g = 0.969\,\text{yr}$ and $c^2/g = 0.969\,\text{ly}$. After $\tau = 1\,\text{yr}$ of Rachel's time, $g\tau/c = 1.032$, and

$$t = (0.969\,\text{yr})\sinh 1.032 = 1.19\,\text{yr}, \qquad x = (0.969\,\text{ly})\,(\cosh 1.032 - 1) = 0.563\,\text{ly},$$

and she is moving at $0.77c$. After $\tau = 10\,\text{yr}$, $g\tau/c = 10.3$, and t and x are both about $(c/2g)e^{g\tau/c}$, or $1.5\times10^4\,\text{yr}$ and $1.5\times10^4\,\text{ly}$. Once her speed is close to c, the distance she has covered grows by a factor of e with every further $0.97\,\text{yr}$ of her time. Ten years on board take her far beyond the nearest stars, but nearly fifteen thousand years pass on Earth. Two questions are left aside here: how a ship could keep up a $1g$ acceleration for years, and how it would stop at the other end. Exercise P25.21 takes up the second.

25.10 Energy and momentum

Multiplying the four-velocity of an object by its mass m gives its **four-momentum**,

$$p^\mu \equiv m\eta^\mu = \begin{pmatrix}\gamma_u mc\\ \gamma_u m\mathbf{u}\end{pmatrix}, \tag{25.28}$$

a four-vector, since m is a scalar. Here, as everywhere in this book, m is the mass of the object measured in its own rest frame, which is the same in every frame. (Some older books call $\gamma_u m$ the "relativistic mass" and m the "rest mass"; most physicists today, and this book, use only the one mass m.) The space part of p^μ is the **momentum** of the object,

$$\mathbf{p} = \gamma_u m\mathbf{u},$$

which becomes $m\mathbf{u}$ at low speeds. The time component is E/c, where

$$\boxed{E = \gamma_u mc^2} \tag{25.29}$$

is the **energy** of the object. So $p^\mu = (E/c, \mathbf{p})$. To see why E deserves the name, expand γ_u for small speeds:

$$E = mc^2\left(1 + \frac{u^2}{2c^2} + \frac{3u^4}{8c^4} + \cdots\right) = mc^2 + \frac{1}{2}mu^2 + \cdots.$$

The second term is the kinetic energy of Newtonian mechanics. The first, mc^2, is the **rest energy**, which an object has even when it is at rest. The kinetic energy is the rest of E,

$$K = E - mc^2 = \left(\gamma_u - 1\right)mc^2. \tag{25.30}$$

The famous equation $E = mc^2$ is the special case of (25.29) for an object at rest; in general $E = \gamma_u mc^2$. Einstein found the equivalence of mass and energy a few months after his first paper on relativity: a body that gives off an energy

L as radiation loses an amount L/c^2 of mass.[9] Rest energies are usually given in electron volts: 0.511 MeV for the electron and 938.3 MeV for the proton.[10] In particle physics, masses are often quoted in MeV/c^2 and momenta in MeV/c, so that mc^2 and pc come out in MeV.

By (25.30), the kinetic energy grows without limit as u approaches c, since γ_u does. This is the second reason, promised in Section 25.5.2, that no object reaches the speed of light: it would take an infinite amount of energy.

That energy and momentum belong together in a four-vector matters most for their conservation. Experiments show that the total four-momentum of an isolated system is the same before and after any collision or decay, so its total energy and its total momentum are both conserved. If the total four-momentum is conserved in one inertial frame, it is conserved in all of them: the change in the total, $p^\mu_{\text{after}} - p^\mu_{\text{before}}$, is a four-vector, and a four-vector that is zero in one frame is zero in every frame. Mass, on the other hand, is *not* conserved. In a collision, kinetic energy can become rest energy, and the reverse.

The scalar product of the four-momentum with itself is, by (25.23),

$$p_\mu p^\mu = m^2 \eta_\mu \eta^\mu = -m^2 c^2.$$

With $p^\mu = (E/c, \mathbf{p})$, the left side is $-E^2/c^2 + p^2$, so

$$\boxed{E^2 = p^2 c^2 + m^2 c^4}. \tag{25.31}$$

This relation between the energy and the momentum of an object holds in every frame. A right triangle with sides mc^2 and pc and hypotenuse E is a handy way to remember it (Figure 25.19). Since $\mathbf{p}/E = \mathbf{u}/c^2$ by (25.28) and (25.29), the velocity of an object is

$$\mathbf{u} = \frac{\mathbf{p}c^2}{E}. \tag{25.32}$$

Equation (25.31) also allows particles with no mass. For $m = 0$ it gives $E = pc$, and then (25.32) gives $u = c$: a massless particle always travels at the speed of light. (Its energy and momentum cannot be found from (25.28) and (25.29), which give 0/0 for $m = 0$ and $u = c$.) The photon, the particle of light, is one, with $E = h\nu$ and $p = E/c = h/\lambda$, where h is Planck's constant, ν the frequency, and λ the wavelength.

Two properties are easily confused. A quantity is **invariant** if it has the same value in every inertial frame, and **conserved** if its total for an isolated system is the same before and after any collision, in any one frame. Mass is invariant but not conserved; energy and momentum are conserved but not invariant. Charge is both (Table 25.1). Its conservation and its invariance are two of the experimental properties of charge listed in Section 3.1, and Chapter 26 relies on both.

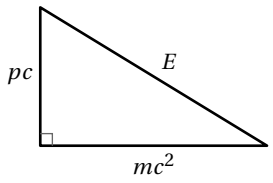


Figure 25.19 A way to remember (25.31): a right triangle with the rest energy mc^2 and pc as its sides and the energy E as its hypotenuse.

	Invariant	Conserved
Mass	yes	no
Energy	no	yes
Momentum	no	yes
Charge	yes	yes

Table 25.1 Invariant (the same in every inertial frame) and conserved (the same total before and after a collision).

[9] "Ist die Trägheit eines Körpers von seinem Energieinhalt abhängig?" A. Einstein, Ann. Phys. 18, 639–641 (1905). https://doi.org/10.1002/andp.19053231314.

[10] From the CODATA 2022 values of the masses (the reference of the Physical Constants page): $m_e c^2 = 0.510\,998\,950\,69$ MeV and $m_p c^2 = 938.272\,089\,43$ MeV.

25.11 Collisions

In a **collision**, or a decay, the total four-momentum of the particles is the same before and after, and that is often all one needs to know. Kinetic energy need not be conserved. A collision is elastic if the kinetic energy is conserved, which happens when every particle keeps its mass, and inelastic otherwise; but energy, rest energy included, is conserved in every collision. The following example, which in Newtonian mechanics would be a typical inelastic collision, shows what becomes of the kinetic energy.

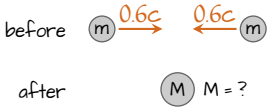


Figure 25.20 The two lumps of Example 25.6, each moving toward the other at $0.6c$ in Luke's frame.

Example 25.6

Two lumps of clay, each of mass m, move toward each other in Luke's frame, each at $0.6c$. They collide head on and stick together. (a) What is the mass M of the lump that results? (b) Rachel moves with one of the lumps before the collision. Check that energy and momentum are conserved in her frame.

Solution:



The first step is a sketch of the situation before and after: the two lumps moving toward each other, and the single lump of unknown mass that results.

(a) In Luke's frame the two momenta are equal and opposite, so the total momentum is zero, and the new lump is at rest. Conservation of energy, with $\gamma_u = 1/\sqrt{1-0.36} = 1.25$ for each lump, gives

$$1.25\,mc^2 + 1.25\,mc^2 = Mc^2, \qquad \text{so} \qquad M = 2.5\,m.$$

The new lump has more mass than the two lumps together. Their kinetic energy, $2(1.25-1)mc^2 = 0.5\,mc^2$, has become rest energy: the lump is hotter, and its internal energy is part of its rest energy.

(b) Rachel moves with the lump on the left, at $0.6c$ to the right in Luke's frame. By (25.12), the other lump moves toward her at

$$u = \frac{0.6c + 0.6c}{1 + (0.6)(0.6)} = 0.882c,$$

with $\gamma_u = 2.125$, and the new lump, at rest in Luke's frame, moves to the left at $0.6c$ in hers. Before the collision, with the momentum to the right counted positive,

$$E = mc^2 + 2.125\,mc^2 = 3.125\,mc^2, \qquad p = -(2.125)\,m\,(0.882c) = -1.875\,mc,$$

and after it,

$$E = (1.25)(2.5\,m)c^2 = 3.125\,mc^2, \qquad p = -(1.25)(2.5\,m)(0.6c) = -1.875\,mc.$$

Both are conserved in her frame too, as they must be. The kinetic energy is not: before the collision it is $1.125\,mc^2$ in her frame, and after it $0.625\,mc^2$.

An important collision involving a massless particle is **Compton scattering**: an X-ray photon scattering from an electron at rest. A photon of energy E_0, moving along x, strikes an electron of mass m; the photon leaves at the angle θ with energy E, and the electron recoils with momentum p_e at the angle ϕ on the other side (Figure 25.21). Conservation of momentum, along x and along y, gives

$$\frac{E_0}{c} = \frac{E}{c}\cos\theta + p_e\cos\phi, \qquad 0 = \frac{E}{c}\sin\theta - p_e\sin\phi,$$

and conservation of energy, with (25.31) for the electron after the collision,

$$E_0 + mc^2 = E + \sqrt{m^2c^4 + p_e^2c^2}.$$

The electron's angle is not needed. Moving the photon terms of the momentum equations to the left, squaring, and adding eliminates it:

$$p_e^2c^2 = (E_0 - E\cos\theta)^2 + (E\sin\theta)^2 = E_0^2 - 2E_0E\cos\theta + E^2.$$

Then the energy equation, rearranged and squared, is

$$\left(E_0 + mc^2 - E\right)^2 = m^2c^4 + E_0^2 - 2E_0E\cos\theta + E^2.$$

The squares of E_0, E, and mc^2 cancel, leaving $E_0mc^2 - E_0E - Emc^2 = -E_0E\cos\theta$, which, divided by E_0Emc^2, is

$$\frac{1}{E} - \frac{1}{E_0} = \frac{1-\cos\theta}{mc^2}. \tag{25.33}$$

The scattered photon has less energy than the incident one, and so, by $E = hc/\lambda$, a longer wavelength:

$$\boxed{\lambda - \lambda_0 = \frac{h}{mc}(1-\cos\theta)}. \tag{25.34}$$

The length $h/mc = 2.43\,\text{pm}$ is called the Compton wavelength of the electron. Arthur Compton found the shift in 1922 and explained it in this way in 1923. In his paper, X-rays from molybdenum, of wavelength 0.0708 nm, scattered at 90° by graphite, had their wavelength increased by 0.0022 nm, where (25.34) gives 0.0024 nm.[11] Classical electromagnetism predicts scattered X-rays of the same wavelength as the incident ones, and the shift was strong evidence that light carries momentum in quanta, as particles do.

When a collision makes new particles, its energy is best judged in the **center-of-momentum frame**, the frame in which the total momentum of the particles is zero (Exercise P25.24 finds its velocity). Only the energy in that frame can go into making mass, since the total momentum in any other frame is not zero and must still be there after the collision, as kinetic energy of the products. Exercises P25.27 and P25.28 show why particle physicists collide two beams head on rather than aim one beam at a target at rest. The following tutorial uses the conservation of four-momentum to answer three questions about decays and collisions.

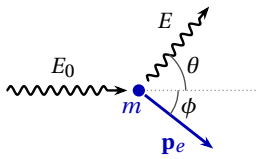


Figure 25.21 Compton scattering: a photon of energy E_0 strikes an electron at rest, and leaves at the angle θ with energy E, while the electron recoils with momentum $\mathbf{p}_e$.

[11] "A quantum theory of the scattering of X-rays by light elements," A. H. Compton, Phys. Rev. 21, 483–502 (1923). https://doi.org/10.1103/PhysRev.21.483.



Arthur H. Compton (1892–1962, American) Arthur Holly Compton was born in Wooster, Ohio, the son of a professor at the College of Wooster, from which he graduated in 1913. He received his PhD from Princeton in 1916, taught for a year at the University of Minnesota, and worked as a research engineer for the Westinghouse Lamp Company in Pittsburgh. After a year at Cambridge, he became head of the physics department of Washington University in St. Louis in 1920. There he found, in 1922, that X-rays scattered by electrons have longer wavelengths than the incident X-rays, and explained the shift with light quanta that carry momentum. He moved to the University of Chicago in 1923 and shared the 1927 Nobel Prize in Physics with C. T. R. Wilson, Compton's "for his discovery of the effect named after him." In the 1930s he led a worldwide survey of cosmic rays, which showed that their intensity depends on the geomagnetic latitude, so that most of them are charged particles. During the Second World War he directed the Metallurgical Laboratory in Chicago, part of the Manhattan Project, where the first nuclear reactor was built. After the war he returned to Washington University as its chancellor. He died in Berkeley in 1962. (Wikipedia)

Tutorial 25.1 Relativistic particle decay

Printable version

Background

In a decay, one particle, called the parent, turns into two or more daughter particles; in a collision, two parents turn into daughters. The four-momentum of a group of particles is the sum of the four-momenta of the particles, and the center-of-momentum frame of the group is the one in which its total momentum is zero. Just as energy and momentum are conserved in introductory mechanics, four-momentum is conserved here: it is an experimental law, which holds in every frame of reference. The energy E, the momentum p, and the mass m of each particle are related by (25.31), $E^2 = p^2c^2 + m^2c^4$. This tutorial uses these facts to answer three questions:

(a) Since mass and energy can be transformed into each other, could we make more massive decay products by giving an unstable particle more kinetic energy?

(b) Why does a particle need to accelerate in order to emit photons?

(c) Can we make more massive particles by colliding two highly energetic parent particles?

Part 1 — One parent, two daughters

A parent particle of mass M decays into two daughter particles with masses m_1 and m_2. The parent travels at speed v in the lab frame. In an attempt to answer question 1, we have given the parent a lot of kinetic energy: v is close to c in the lab frame.

(a) There are four reference frames that could be used to analyze this situation: the lab frame, the parent's frame, the frame of daughter 1, and the frame of daughter 2. Which is likely to be the simplest? Choose that frame. (If you have any doubts about whether you are making the best choice, please check.) Draw the situation in that frame, and write down the initial four-momentum, the four-momentum of the parent particle.

(b) For the sake of consistent notation, write the three-momenta of the daughter particles in your chosen frame as $\mathbf{p}_1 = p_{1x}\hat{\mathbf{x}} + p_{1y}\hat{\mathbf{y}} + p_{1z}\hat{\mathbf{z}}$ and $\mathbf{p}_2 = p_{2x}\hat{\mathbf{x}} + p_{2y}\hat{\mathbf{y}} + p_{2z}\hat{\mathbf{z}}$. Write down the four-momenta of the two daughters, using (25.31) to express their time components in terms of m_1, m_2, p_1, and p_2.

(c) Now apply conservation of four-momentum: set your answer to 1(a) equal to the sum of your answers to 1(b). The time component represents conservation of energy, and the three space components can be combined into one vector equation representing conservation of momentum.

(d) Prove that the conservation of energy equation has the consequence that $m_1 + m_2 \le M$. To do this, think about what you need to do with p_1 and p_2 to make m_1 and m_2 as large as possible. This means that the total mass of the daughter particles can never be more than the parent's mass, even if the parent has a lot of kinetic energy in the lab frame. We have the answer to question 1: the decay of a single particle cannot be used to convert kinetic energy into mass.

(e) Does that statement generalize to more than two daughters? In other words, will it be true that $m_1 + m_2 + m_3 + \cdots \leq M$? Why or why not?

(f) The conservation of momentum equation also has an important physical consequence: in this frame the momenta are equal and opposite, $\mathbf{p}_1 = -\mathbf{p}_2$, whatever the daughters' masses, so that $p_1 = p_2$ in magnitude. Two daughters of different masses still have the same magnitude of momentum. Does that statement generalize to more than two daughters? In other words, if there are several daughters, will it be true that $p_1 = p_2 = p_3 = \cdots$ in this frame, whatever the values of m_1, m_2, m_3, and so on? Why or why not?

(g) Define $m_{\text{tot}} = m_1 + m_2$ and $\Delta m = m_1 - m_2$. Can Δm ever be larger than m_{tot}? Why or why not? Can Δm ever be larger than M? Why or why not?

(h) Rewrite your conservation of energy equation in terms of m_{tot} and Δm.

(i) Use Mathematica's `Solve` command, or `sympy.solve` in Python, to solve the conservation of energy equation for p_1 (which also equals p_2) in terms of M, m_{tot}, and Δm. You will likely get two possible answers. Are either or both of them physically reasonable? How do you know? Can the constraints on m_{tot} and Δm help you decide?

(j) Show that the physically reasonable solution can be put in the form

$$p_1 = p_2 = \frac{Mc}{2}\sqrt{\left[1 - \left(\frac{m_{\text{tot}}}{M}\right)^2\right]\left[1 - \left(\frac{\Delta m}{M}\right)^2\right]}. \tag{25.35}$$

One nice thing about this form is that everything inside the square root is dimensionless, so both sides of the equation explicitly have the dimensions of momentum.

Part 2 — Reflection questions

Use (25.35) to address the following questions.

(a) What happens to the magnitudes of the momenta, p_1 and p_2, as the daughters' total mass approaches M?

(b) What happens to p_1 and p_2 if both daughters are massless?

(c) Now consider a charged particle of mass M that emits a gamma ray (a photon). What are m_1 and m_2 in this scenario? What would the momentum of the photon need to be? Is that really a photon?

We now have the answer to question 2. Your equations should show that the decay of a particle cannot convert its kinetic energy into other particles, even massless ones, however much kinetic energy it has in the lab frame. So a charge cannot emit a photon and remain otherwise unchanged unless it is accelerating: as long as the charge's rest frame is an inertial frame, special relativity forbids it.

Part 3 — Two parents, two daughters

An important extension of this problem is to $2 \to 2$ scattering. In an attempt to answer question 3, let two incoming (parent) particles, with a lot of kinetic energy, collide and turn into two daughter particles. Let the parents' total energies be E_1 and E_2, and their momenta $\mathbf{P}_1$ and $\mathbf{P}_2$; the daughters' masses and momenta

are m_1, m_2, $\mathbf{p}_1$, and $\mathbf{p}_2$, as in Part 1. Again analyze the situation in the center-of-momentum frame.

(a) What is the relationship between $\mathbf{P}_1$ and $\mathbf{P}_2$ in this frame? Define $E = E_1 + E_2$, and write down the total four-momentum of the incoming particles.

(b) Show that this four-momentum is mathematically equivalent to the four-momentum you found in 1(a), with the substitution $E = Mc^2$.

(c) Since the situation is mathematically equivalent to Part 1 with that substitution, the solution for p_1 and p_2 is (25.35) with M replaced by E/c^2:

$$p_1 = p_2 = \frac{E}{2c}\sqrt{\left[1-\left(\frac{m_{\text{tot}}c^2}{E}\right)^2\right]\left[1-\left(\frac{\Delta m\, c^2}{E}\right)^2\right]}.$$

What is the new constraint on m_{tot}? What about the constraint on Δm?

(d) We are finally ready to answer question 3. Based on your results, if you increased the kinetic energy of the incoming particles, would it be possible to make more massive outgoing particles?

Part 4 — Wrap-up

(a) Are there ways you can think of to check (25.35) and your answers in Part 2? (Limiting cases, units, special cases you can work out another way, and so forth.)

25.12 Relativistic dynamics

How does a particle respond to a force when it moves near the speed of light? In **relativistic dynamics**, Newton's second law keeps the form in which Newton himself wrote it, as the rate of change of momentum, but with the momentum of Section 25.10:

$$\boxed{\mathbf{F} = \frac{d\mathbf{p}}{dt}, \qquad \mathbf{p} = \gamma_u m\mathbf{u}}. \tag{25.36}$$

Like the conservation of four-momentum, this is a law to be tested by experiment, and it passes: the electric and magnetic forces of the Lorentz force law (10.2) obey it, as the motion of fast charged particles confirms. The force is no longer the mass times the acceleration. Differentiating $\gamma_u m\mathbf{u}$ with the help of the derivative of γ_u in Section 25.9 gives

$$\mathbf{F} = \gamma_u m\mathbf{a} + \gamma_u^3 m\frac{(\mathbf{u}\cdot\mathbf{a})\,\mathbf{u}}{c^2},$$

so a force along the motion produces the acceleration $F/\gamma_u^3 m$, and a force across the motion the acceleration $F/\gamma_u m$. The same force accelerates a fast particle much less along its motion than across it.

The work done by the force still changes the energy. Since $E = \gamma_u mc^2$,

$$\frac{dE}{dt} = mc^2 \frac{d\gamma_u}{dt} = \gamma_u^3 m\,\mathbf{u}\cdot\mathbf{a},$$

and the dot product of the equation above with $\mathbf{u}$ is

$$\mathbf{F}\cdot\mathbf{u} = \gamma_u m\,\mathbf{u}\cdot\mathbf{a}\left(1 + \frac{\gamma_u^2 u^2}{c^2}\right) = \gamma_u^3 m\,\mathbf{u}\cdot\mathbf{a},$$

since $1 + \gamma_u^2 u^2/c^2 = \gamma_u^2$. So

$$\frac{dE}{dt} = \mathbf{F}\cdot\mathbf{u}. \tag{25.37}$$

A magnetic force is always perpendicular to $\mathbf{u}$, so it does no work (Section 10.1.3): it leaves E, and with it γ_u and the speed, unchanged.

The force $\mathbf{F}$ is not part of a four-vector. Its four-vector relative is the derivative of the four-momentum with respect to proper time,

$$K^\mu \equiv \frac{dp^\mu}{d\tau} = m\alpha^\mu, \qquad \text{with} \quad K^0 = \frac{\gamma_u}{c}\frac{dE}{dt}, \qquad K^{1,2,3} = \gamma_u \mathbf{F}, \tag{25.38}$$

called the **Minkowski force**. It transforms simply, as every four-vector does, but it is not the force that a spring scale measures or that appears in the Lorentz force law; that force is $\mathbf{F}$.

A constant force F along x, acting on a particle that starts from rest, gives $d(\gamma_u mu)/dt = F$, or $\gamma_u u = Ft/m$. That is the equation of part (a) of Example 25.5 with $g = F/m$, so the particle's motion is the hyperbolic motion found there:

$$u = \frac{Ft/m}{\sqrt{1 + (Ft/mc)^2}}, \qquad x = \frac{mc^2}{F}\left[\sqrt{1 + \left(\frac{Ft}{mc}\right)^2} - 1\right].$$

This is no coincidence. A force along the motion of a particle has the same value in the particle's instantaneous rest frame, as the force transformation below shows, and there it produces the proper acceleration F/m. Exercise P25.30 compares the motion with Newton's.

A particle of charge q moving perpendicular to a uniform magnetic field $\mathbf{B}$ keeps its speed, so γ_u is constant, and $d\mathbf{p}/dt = \gamma_u m\, d\mathbf{u}/dt = q\mathbf{u}\times\mathbf{B}$. This is the Newtonian equation of Section 10.1.4 with m replaced by $\gamma_u m$, so the particle moves on a circle, with radius and angular frequency

$$R = \frac{p}{qB} = \frac{\gamma_u mu}{qB}, \qquad \omega = \frac{qB}{\gamma_u m}. \tag{25.39}$$

The radius is still the momentum divided by qB, as in (10.5), with the relativistic momentum. The centripetal acceleration is still u^2/R, but the force that produces it is $\gamma_u mu^2/R$, not mu^2/R. And the angular frequency is the cyclotron frequency (10.6) divided by γ_u: the period grows with the energy. Example 25.7 shows what that does to a cyclotron.

Chapter 26 finds the fields that one observer measures from those that another measures by comparing the forces they measure on the same charge. For that, it needs the rule that relates the two forces. Let a particle move with velocity $\mathbf{u}_\mathrm{L}$ in Luke's frame, where the force on it is $\mathbf{F}_\mathrm{L}$, and let Rachel move along x at v, as usual. In her frame the force is $\mathbf{F}_\mathrm{R} = d\mathbf{p}_\mathrm{R}/dt_\mathrm{R}$. Since $(E/c, \mathbf{p})$ and $(ct, \mathbf{r})$ are four-vectors, (25.8) gives

$$dp_{\mathrm{R},x} = \gamma_v\left(dp_{\mathrm{L},x} - \beta_v\frac{dE_\mathrm{L}}{c}\right), \qquad dp_{\mathrm{R},y} = dp_{\mathrm{L},y}, \qquad dt_\mathrm{R} = \gamma_v\left(dt_\mathrm{L} - \frac{\beta_v}{c}dx_\mathrm{L}\right).$$

Dividing, and then dividing the numerator and the denominator by dt_L, with $dE_\mathrm{L}/dt_\mathrm{L} = \mathbf{F}_\mathrm{L}\cdot\mathbf{u}_\mathrm{L}$ by (25.37) and $dx_\mathrm{L}/dt_\mathrm{L} = u_{\mathrm{L},x}$,

$$\begin{aligned} F_{\mathrm{R},x} &= \frac{F_{\mathrm{L},x} - \beta_v\,\mathbf{F}_\mathrm{L}\cdot\mathbf{u}_\mathrm{L}/c}{1-\beta_v u_{\mathrm{L},x}/c}, \\ F_{\mathrm{R},y} = \frac{F_{\mathrm{L},y}}{\gamma_v\left(1-\beta_v u_{\mathrm{L},x}/c\right)}, &\qquad F_{\mathrm{R},z} = \frac{F_{\mathrm{L},z}}{\gamma_v\left(1-\beta_v u_{\mathrm{L},x}/c\right)}. \end{aligned} \tag{25.40}$$

These are more complicated than the transformation of a four-vector, and they resemble the rules for velocities, (25.12) and (25.13), which come about in the same way. The case that Chapter 26 needs most is a particle at rest in Rachel's frame, so that $\mathbf{u}_\mathrm{L} = v\hat{\mathbf{x}}$ and $\mathbf{F}_\mathrm{L}\cdot\mathbf{u}_\mathrm{L} = F_{\mathrm{L},x}v$. Then

$$F_{\mathrm{R},x} = F_{\mathrm{L},x}, \qquad F_{\mathrm{R},y} = \gamma_v F_{\mathrm{L},y}, \qquad F_{\mathrm{R},z} = \gamma_v F_{\mathrm{L},z}$$
(particle at rest in Rachel's frame) (25.41)

The component of the force along the relative motion is the same in both frames, and the components across it are larger, by γ_v, in the particle's own rest frame.

One law of Newtonian mechanics does not survive. Newton's third law says that two particles exert equal and opposite forces on each other at the same time, but two observers in relative motion disagree about which events are at the same time, so for particles some distance apart, whose forces change with time, the law cannot hold in every frame. For moving charges it can fail even in a single frame: two charges moving at right angles to each other exert magnetic forces on each other that are in general not equal and opposite. Momentum is still conserved, because the electric and magnetic fields themselves carry momentum (Section 20.3.1). For forces of contact, which act at a single place, there is no such difficulty.

25.13 Numerical method: relativistic trajectories

Section 10.1.5 followed charged particles through electric and magnetic fields with the Boris method. At speeds close to c the method needs one change: the quantity it steps forward must be the momentum, not the velocity. With the momentum per unit mass,

$$\mathbf{w} \equiv \gamma_u\mathbf{u} = \frac{\mathbf{p}}{m},$$

the equations of motion (25.36) with the Lorentz force are

$$\frac{d\mathbf{w}}{dt} = \frac{q}{m}(\mathbf{E} + \mathbf{u} \times \mathbf{B}), \qquad \frac{d\mathbf{r}}{dt} = \mathbf{u} = \frac{\mathbf{w}}{\gamma_u}, \qquad \gamma_u = \sqrt{1 + \frac{w^2}{c^2}},$$

where the last relation follows from $\gamma_u^2 - \gamma_u^2 u^2/c^2 = 1$. The **relativistic Boris method** is the method of Section 10.1.5 applied to $\mathbf{w}$. With $\mathbf{a} = (q\Delta t/2m)\mathbf{E}$, each step is

$$\mathbf{w}^- = \mathbf{w}_n + \mathbf{a}, \qquad \gamma^- = \sqrt{1 + (w^-/c)^2}, \qquad \mathbf{t} = \frac{q\Delta t}{2m\gamma^-}\mathbf{B},$$

$$\mathbf{w}' = \mathbf{w}^- + \mathbf{w}^- \times \mathbf{t}, \qquad \mathbf{w}^+ = \mathbf{w}^- + \frac{2\mathbf{w}' \times \mathbf{t}}{1 + t^2}, \qquad \mathbf{w}_{n+1} = \mathbf{w}^+ + \mathbf{a},$$

$$\gamma_{n+1} = \sqrt{1 + (w_{n+1}/c)^2}, \qquad \mathbf{r}_{n+1} = \mathbf{r}_n + \frac{\mathbf{w}_{n+1}}{\gamma_{n+1}}\Delta t.$$

The differences from Section 10.1.5 are the variable $\mathbf{w}$, the Lorentz factor in $\mathbf{t}$, and the division by γ_{n+1} in the step for the position. As before, the step from $\mathbf{w}^-$ to $\mathbf{w}^+$ is a rotation, so in a pure magnetic field w, and with it γ_u and the speed, stay exactly constant, and the particle turns at the angular frequency $qB/\gamma_u m$ of (25.39), with an error that shrinks as the square of the time step. Since w can grow without limit while $u = w/\gamma_u$ cannot exceed c, no step can carry the particle faster than light.

25.13.1 Tips for Mathematica and Python

- When you use `NDSolve` in Mathematica or `solve_ivp` in Python (from `scipy.integrate`), write the equations for $\mathbf{r}$ and $\mathbf{w}$, not for $\mathbf{u}$, so that no numerical error can carry the speed past c.
- For a slow particle, $\gamma_u - 1$ is the difference of two nearly equal numbers. Compute the kinetic energy as $K = (\gamma_u - 1)mc^2 = mw^2/(\gamma_u + 1)$, which has no such cancellation.
- To catch an event, such as the crossing of a surface, look for a change of sign between two steps and interpolate linearly for its time; `solve_ivp` has an `events` option for this, and Mathematica has `WhenEvent`.
- Check a simulation against cases you can solve by hand: a circle in a uniform $\mathbf{B}$, with (25.39), and the hyperbolic motion of Example 25.5 in a uniform $\mathbf{E}$, with $g = qE/m$.

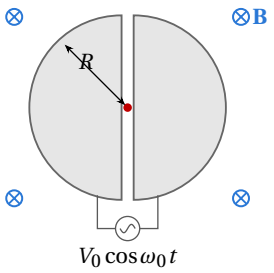


Figure 25.22 The cyclotron of Example 25.7: two dees of radius R in a uniform field **B**, with the voltage $V_0 \cos\omega_0 t$ across the gap between them.

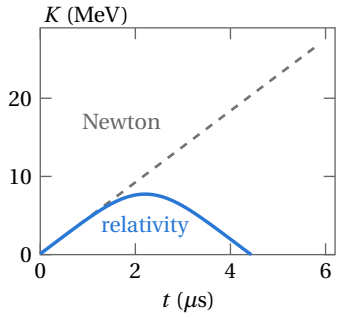
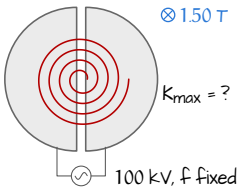


Figure 25.23 The kinetic energy of the proton of Example 25.7 against time, with relativity (blue) and with Newtonian mechanics (dashed), for a voltage of fixed frequency.

Example 25.7

The cyclotron of Example 10.3 accelerates protons in a uniform field $B = 1.50\,\text{T}$, between dees of radius 0.500 m, with a voltage that alternates at the fixed angular frequency $\omega_0 = eB/m$, the cyclotron frequency of a slow proton. Treat the gap as thin: each time a proton crosses it from left to right, at time t, its kinetic energy increases by $eV_0 \cos\omega_0 t$, and each time it crosses from right to left, its kinetic energy decreases by that amount, with $V_0 = 100\,\text{kV}$. At $t = 0$ a proton at rest at the center of the gap crosses it from left to right, gaining 100 keV. Follow the proton with the relativistic Boris method. What largest kinetic energy does it reach, and on an orbit of what radius? Compare with the 26.9 MeV that Example 10.3 found at the edge of the dees.

Solution:



The first step is a sketch of the cyclotron: the dees, the field, the voltage of fixed frequency across the gap, and the proton spiraling outward, with the unknown largest energy.

In Newtonian mechanics a proton takes half a period, $\pi m/eB$, to go around one dee, whatever its speed, and so it reaches the gap each time just as the voltage has reversed. It gains 100 keV at every crossing, and after 269 crossings it reaches the edge of the dees with 26.9 MeV. By (25.39), half a turn actually takes $\pi\gamma_u m/eB$, longer by the factor γ_u, so the proton arrives at the gap a little later each time.

The program starts with the proton at the center of the gap, moving to the right with **w** for a kinetic energy of 100 keV. Inside the dees $\mathbf{E} = 0$, and each step is the rotation of the relativistic Boris method, with 800 steps per period of the voltage. When the proton's x coordinate changes sign between two steps, it has crossed the gap: the program finds the time of the crossing by linear interpolation, adds $\pm eV_0 \cos\omega_0 t$ to the kinetic energy, and scales **w** to the magnitude that gives the new kinetic energy, keeping its direction. Halving the time step changes the results below by less than 0.1%.

The kinetic energy rises at first by nearly 100 keV per crossing, but more and more slowly (Figure 25.23). It reaches its largest value, 7.74 MeV, at the 101st crossing, about 50 turns and 2.20 μs after the start. Then $pc = \sqrt{K^2 + 2Kmc^2} = 120.8\,\text{MeV}$, and by (25.39) the orbit has the radius

$$R = \frac{p}{eB} = \frac{120.8 \times 10^6\,\text{V}}{(1.50\,\text{T})(2.998 \times 10^8\,\text{m/s})} = 0.268\,\text{m}.$$

After that the proton arrives at the gap when the voltage slows it down, and it loses energy at each crossing, spiraling back toward the center: it never gets near the edge of the dees.

A short calculation explains the number. Let φ be the phase of the voltage when the proton crosses the gap, measured from the crest, so that the energy gained at the crossing is $\Delta K = eV_0 \cos\varphi$. Each half turn takes longer than half a period of the voltage by the factor γ_u, so it adds $\pi(\gamma_u - 1) = \pi K/mc^2$ to φ. Treating the number of crossings as continuous,

$$\frac{d\varphi}{dK} = \frac{\pi K/mc^2}{eV_0\cos\varphi}, \qquad \text{so} \qquad \sin\varphi = \frac{\pi K^2}{2eV_0 mc^2},$$

starting from $\varphi = 0$ at $K = 0$. The energy stops growing when the proton reaches the gap a quarter of a cycle late, at $\varphi = 90°$, so

$$K_{\text{max}} = \sqrt{\frac{2eV_0 mc^2}{\pi}} = \sqrt{\frac{2(0.100\,\text{MeV})(938.3\,\text{MeV})}{\pi}} = 7.73\,\text{MeV},$$

in agreement with the simulation. Example 10.3 predicted 26.9 MeV at the edge of the dees, but with a voltage of fixed frequency the protons fall out of step long before they get there. Even with the voltage in step, the edge would stop them at 26.6 MeV, not 26.9: the momentum at the edge is still $p = eBR$, with $pc = 224.8\,\text{MeV}$, but the kinetic energy that goes with it is $\sqrt{p^2c^2 + m^2c^4} - mc^2$, not $p^2/2m$. Larger machines change the frequency or the field as the protons speed up.

25.14 Summary

Einstein's two postulates, that the laws of physics and the speed of light are the same in every inertial frame, lead to the Lorentz transformation, which mixes space and time. Observers in relative motion disagree about lengths, times, and simultaneity, but agree about the interval between two events, about proper times, and about the order of events that could be cause and effect. Four-vectors and their scalar products keep track of what changes from frame to frame and what does not.

The four-vectors of mechanics all come from the position four-vector $x^\mu = (ct, \mathbf{r})$ by steps that keep a four-vector a four-vector. Dividing dx^μ by the proper time $d\tau$, on which every observer agrees, gives the four-velocity η^μ; multiplying by the mass gives the four-momentum $p^\mu = (E/c, \mathbf{p})$; and differentiating with respect to τ gives the four-acceleration α^μ and the Minkowski force K^μ. The scalar product of each with itself is a Lorentz scalar: $-c^2$ for η^μ, $-m^2c^2$ for p^μ, and a^2_{proper} for α^μ. The conservation of four-momentum governs collisions and decays, in which kinetic energy and rest energy can turn into each other, and Newton's second law survives in the form $\mathbf{F} = d\mathbf{p}/dt$. The rule for transforming forces is the bridge to Chapter 26, which uses it to find how the electric and magnetic fields themselves transform.

The Lorentz transformation and its consequences

$\beta = \dfrac{v}{c}, \quad \gamma = \dfrac{1}{\sqrt{1-\beta^2}}$	Rachel moves along x at v relative to Luke
$ct_R = \gamma\left(ct_L - \beta x_L\right),$ $x_R = \gamma\left(x_L - \beta c t_L\right)$	the Lorentz transformation; y and z unchanged; the inverse by $\beta \to -\beta$; turning a world line to the left takes $-\gamma\beta$
$\Delta t = \gamma\,\Delta\tau, \quad L = \dfrac{L_0}{\gamma}$	time dilation and length contraction; $\Delta\tau$ and L_0 are measured in the rest frame of the clock or the object
$\Delta t_L = \dfrac{\gamma v \Delta x_R}{c^2}$	events simultaneous for Rachel, Δx_R apart, are not simultaneous for Luke
$(\Delta s)^2 = -(c\,\Delta t)^2 + \Delta x^2 + \Delta y^2 + \Delta z^2$	the interval, the same in every frame; timelike (< 0), spacelike (> 0), or lightlike
$u_L = \dfrac{u_R + v}{1 + u_R v/c^2}$	velocity addition along x; never more than c

Four-vectors

$a_R^\mu = \Lambda^\mu{}_\nu\, a_L^\nu$	the definition of a four-vector; a repeated index, once up and once down, is summed
$a_\mu = \left(-a^0, a^1, a^2, a^3\right),$ $a_\mu b^\mu = -a^0 b^0 + \mathbf{a}\cdot\mathbf{b}$	lowering an index; the scalar product is the same in every frame
$x^\mu = (ct, \mathbf{r}), \quad d\tau = \dfrac{dt}{\gamma_u}$	position; proper time, invariant
$\eta^\mu = \dfrac{dx^\mu}{d\tau} = \gamma_u(c, \mathbf{u}), \quad \eta_\mu\eta^\mu = -c^2$	four-velocity
$\alpha^\mu = \dfrac{d\eta^\mu}{d\tau}, \quad \alpha_\mu\alpha^\mu = a_{\text{proper}}^2$	four-acceleration; a_{proper} is the acceleration in the instantaneous rest frame
$a_\parallel = \dfrac{a_{\text{proper},\parallel}}{\gamma^3}, \quad a_\perp = \dfrac{a_{\text{proper},\perp}}{\gamma^2}$	accelerations along and across the motion, in a frame where the object moves with Lorentz factor γ
$p^\mu = m\eta^\mu = (E/c, \mathbf{p}), \quad K^\mu = \dfrac{dp^\mu}{d\tau}$	four-momentum; Minkowski force

Energy, momentum, and force

$E = \gamma_u mc^2, \quad \mathbf{p} = \gamma_u m\mathbf{u},$ $K = (\gamma_u - 1)mc^2$	energy, momentum, and kinetic energy; m is the mass, the same in every frame
$E^2 = p^2c^2 + m^2c^4, \quad \mathbf{u} = \dfrac{\mathbf{p}c^2}{E}$	in every frame; for a massless particle $E = pc$ and $u = c$
$\sum p^\mu_{\text{before}} = \sum p^\mu_{\text{after}}$	conservation of four-momentum: energy and momentum are conserved, mass is not (Table 25.1)
$\lambda - \lambda_0 = \dfrac{h}{mc}(1 - \cos\theta)$	Compton scattering of a photon by an electron at rest
$\mathbf{F} = \dfrac{d\mathbf{p}}{dt}, \quad \dfrac{dE}{dt} = \mathbf{F}\cdot\mathbf{u}$	Newton's second law and the work done by a force
$R = \dfrac{p}{qB}, \quad \omega = \dfrac{qB}{\gamma_u m}$	circular motion in a uniform magnetic field
$F_{\text{R},x} = F_{\text{L},x}, \quad F_{\text{R},\perp} = \gamma_v F_{\text{L},\perp}$	the force on a particle at rest in Rachel's frame

Exercises

Exercises for 25.1 The postulates

P25.1 A smuggler's ship flees from a space station at $0.8c$. A patrol ship chasing it at $0.6c$ (both speeds relative to the station) fires a probe forward at $0.25c$ relative to the patrol ship.

(a) According to the Galilean rule (25.2), does the probe catch the smuggler?

(b) According to Einstein's rule (25.12)?

(c) Find the velocities of the patrol ship and of the probe in the smuggler's frame, and show that the answer to (b) is the same in that frame. (Whether the probe reaches the smuggler is a single event, or the absence of one, and every observer must agree about it.)

Exercises for 25.3 Simultaneity, time dilation, and length contraction

P25.2 Charged pions at rest decay with a mean life of 26.03 ns (Particle Data Group, the reference of Example 25.1). In a beam of pions from an accelerator, the number of pions falls by a factor of e every 30.0 m along the beam line. A student divides 30.0 m by 26.03 ns and finds a speed of 1.15×10^9 m/s, almost four times the speed of light.

(a) What is wrong with the student's reasoning?

(b) Find the pions' speed and their Lorentz factor γ.

(c) Describe the same decay in the pions' frame. In that frame, how long is the 30.0 m stretch of beam line, and how long does it take to go past a pion?

P25.3 Rachel travels at $0.9c$ from Earth to the star system α Centauri, about 4.4 ly away in Earth's frame. (Its parallax is 743 mas: "Parallax and masses of α Centauri revisited," D. Pourbaix and H. M. J. Boffin, Astron. Astrophys. 586, A90 (2016), https://doi.org/10.1051/0004-6361/201527859.) Ignore the times needed to speed up and slow down.

(a) How long does the trip take in Earth's frame?

(b) How much does Rachel age during the trip? Find the answer from time dilation, in Earth's frame.

(c) Find the answer to (b) a second way, in Rachel's frame, from length contraction.

(d) How fast must Rachel travel to make the trip in 1.0 yr of her own time?

P25.4 A pole 10.0 m long in its own frame is carried at high speed, along its length, through a barn 8.0 m long in its own frame, with doors at both ends (Figure 25.24).

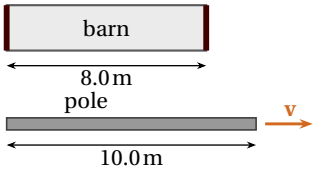


Figure 25.24 The barn and the pole of Exercise P25.4, each drawn in its own frame.

(a) At what speed is the pole 8.0 m long in the barn's frame?

(b) At that speed, the farmer closes both doors at the same moment in the barn's frame, just as the pole is entirely inside, and then opens them again at once. In the pole's frame, how long is the barn? How can a pole 10.0 m long be inside a barn that is shorter than that?

(c) Use the Lorentz transformation (Section 25.4) to find when the two doors close in the pole's frame, and show that in that frame the pole is never inside a closed barn.

P25.5 A cube of side a moves at speed $v = \beta c$ along the x axis. A camera far away on the $-y$ side photographs it when it is straight in front of the camera, so that the light that reaches the camera from every part of the cube travels in the $-y$ direction (Figure 25.25).

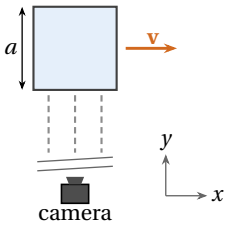


Figure 25.25 The cube of Exercise P25.5, seen from above, moving along x past a distant camera on the $-y$ side.

(a) In the cube's own frame, its face toward the camera is a square of side a. How wide is that face, along x, in the camera's frame?

(b) The face at the back of the cube, which faces the $-x$ direction, is edge-on to the camera and might be expected to be invisible. Show that the camera sees it anyway, as a strip of width βa. (Light from the far edge of that face must leave earlier than light from its near edge, to arrive at the camera at the same time. How much earlier, and how far back was the cube then?)

(c) Show that the photograph shows the two faces with the widths that a cube at rest would have if it were turned by an angle θ with $\sin\theta = \beta$. The moving cube looks rotated, not contracted.

Exercises for 25.4 The Lorentz transformations

P25.6 The matrix

$$M = \begin{pmatrix} 2 & -1.732 \\ -1.732 & 2 \end{pmatrix}$$

takes the coordinates (ct, x) of any event in Luke's frame to its coordinates in Rachel's frame.

(a) Which way does Rachel move relative to Luke, and how fast?

(b) Write the matrix that takes Rachel's coordinates to Luke's, and check that it is the inverse of M.

(c) If the coordinates were written in the order (x, ct) instead, what would M be?

(d) An event happens at $ct_L = 3\,\text{ly}$, $x_L = 1\,\text{ly}$. Where and when does it happen in Rachel's frame?

P25.7 Rachel's spaceship passes Earth, where Luke is, at $0.8c$. One year later by the ship's clocks, she sends a light signal back to Earth.

(a) Draw two spacetime diagrams of the situation, one in Earth's frame and one in the ship's. Label the world lines of the ship and of Earth, the events "signal sent" and "signal received," and the signal itself.

(b) Find the coordinates of the two events in both frames, in years and light-years.

(c) How far does the signal travel, and how long does it take, in each frame?

P25.8 A light wave traveling along x obeys the one-dimensional wave equation

$$\frac{\partial^2 f}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 f}{\partial t^2} = 0,$$

whose solutions $f(x - ct)$ and $f(x + ct)$ move at c in the two directions.

(a) Rewrite the equation in Rachel's coordinates, using the Galilean transformation (25.1). (By the chain rule, $\partial/\partial x_L = \partial/\partial x_R$ and $\partial/\partial t_L = \partial/\partial t_R - v\,\partial/\partial x_R$.) Show that the new equation has solutions of the form $f(x_R - ut_R)$ only for $u = c - v$ and $u = -c - v$.

(b) Repeat (a) with the Lorentz transformation (25.8), and show that the wave equation keeps its form, so that light moves at c in both directions in Rachel's frame too.

Exercises for 25.5 The structure of spacetime

P25.9 Show that the Lorentz transformation always takes a straight line in Luke's spacetime diagram to a straight line in Rachel's. Specifically, if a line in Luke's diagram is $ct_L = mx_L + b$, find its slope and its intercept on the ct_R axis in Rachel's diagram, where Rachel moves to the right at β relative to Luke. Check that a light line, $m = \pm 1$, keeps its slope.

P25.10 In Luke's frame, event A happens at the origin and event B at $x_L = 2.0\,\text{ly}$, $t_L = 1.9\,\text{yr}$.

(a) Show that the separation of A and B is spacelike.

(b) Find the velocity, speed and direction, of an observer for whom the two events are simultaneous.

(c) How far apart are the events for that observer? Find the answer from the interval (25.11), without transforming, and then check it with the Lorentz transformation.

P25.11 Rachel and Luke are twins. On their birthday Rachel leaves Earth at $0.6c$ for a planet 3.0 ly away in Earth's frame, turns around as she arrives, and returns at the same speed.

(a) How much has Rachel aged when she returns? How much has Luke?

(b) On her sixth birthday after leaving, by her own clock, Rachel sends Luke a "happy birthday" message by radio. When does it reach him, by his clock?

(c) Rachel watches Luke's clock through a telescope. What does it read, as she sees it, at the moment she turns around? Explain how what she sees during the whole trip agrees with (a).

P25.12 Suppose that Luke could send a signal at $2c$ in his frame, from the origin to a station at rest 1.0 ly away on his $+x$ axis.

(a) When does the signal arrive, in Luke's frame?

(b) Rachel passes Luke, moving to the right at $0.9c$, at the moment the signal is sent. Show that in her frame the signal arrives before it was sent.

(c) A relay that moves with Rachel's velocity is passing the station when the signal arrives, and at once sends a reply back toward Luke, at $2c$ in its own frame (which is Rachel's frame). When does the reply reach Luke, by his clock? Explain why this rules out signals faster than light.

Exercises for 25.6 Velocity addition

P25.13 Section 25.6 derived Einstein's rule (25.12) for velocities along x. A component of velocity perpendicular to the motion transforms differently.

(a) Rachel throws a ball with velocity components $u_{\mathrm{R},x}$ and $u_{\mathrm{R},y}$ in her frame. For small changes along the ball's world line, the inverse transformation (25.10) gives $dx_{\mathrm{L}} = \gamma(dx_{\mathrm{R}} + \beta c\,dt_{\mathrm{R}})$, $c\,dt_{\mathrm{L}} = \gamma(c\,dt_{\mathrm{R}} + \beta\,dx_{\mathrm{R}})$, and $dy_{\mathrm{L}} = dy_{\mathrm{R}}$. Divide to find $u_{\mathrm{L},x} = dx_{\mathrm{L}}/dt_{\mathrm{L}}$ and $u_{\mathrm{L},y} = dy_{\mathrm{L}}/dt_{\mathrm{L}}$. Show that the first is (25.12) and the second is (25.13). Why is $u_{\mathrm{L},y}$ different from $u_{\mathrm{R},y}$, even though $dy_{\mathrm{L}} = dy_{\mathrm{R}}$?

(b) As Rachel passes Luke at $0.9c$, she throws a ball at $0.9c$, at $45°$ to her x axis as she measures it. Find the ball's speed and direction in Luke's frame.

(c) Sketch spacetime diagrams of the throw in both frames, with two dimensions of space, x and y, and time.

P25.14 This exercise adds velocities graphically, by drawing one observer's coordinate grid on the other's spacetime diagram, as in Figure 25.11. Rachel moves to the right at $3c/5$ relative to Luke, so that $\gamma = 5/4$. On graph paper, draw Luke's spacetime diagram, with ct_L up and x_L across, from -5 to 5 ly in each.

(a) Draw Rachel's world line, $x_R = 0$, with slope $5/3$, and the world lines of the points at rest in her frame at $x_R = -3$, -2, -1, 1, 2, and 3 ly. Because of length contraction, they are 0.8 ly apart horizontally, not 1 ly. Label them.

(b) Draw Rachel's lines of simultaneity for $ct_R = -3, -2, \ldots, 3$ ly. They have slope $3/5$, and by symmetry they are 0.8 ly apart vertically. Label them.

(c) Check your grid by reading two events on it: $(x_L, ct_L) = (1,2)$ ly should be close to $(x_R, ct_R) = (-0.25, 1.75)$ ly, and $(x_R, ct_R) = (1,2)$ ly close to $(x_L, ct_L) = (2.75, 3.25)$ ly.

(d) At the event $(x_R, ct_R) = (-2,-2)$ ly, a ball is thrown in the $+x$ direction at $3c/5$ in Rachel's frame, and it stops after 5.0 yr of Rachel's time. Where does it stop, in her coordinates? Mark the start and the end of its world line on your diagram, using Rachel's grid, and connect them.

(e) From the slope of the ball's world line in Luke's coordinates, find the ball's speed in Luke's frame, and compare it with (25.12).

P25.15 A jet of ionized gas moves at $0.95c$ in the $+x$ direction relative to us (Luke's frame). In the jet's own frame (Rachel's), the gas emits light equally in all directions.

(a) Light leaves the jet at $60°$ to the x axis in the jet's frame. At what angle does it travel in Luke's frame? (One way: in the jet's frame the light reaches the point $(\cos 60°, \sin 60°)$ ly after 1 yr; transform that event.)

(b) With a computer, find the angle in Luke's frame for 36 directions in the jet's frame, $0°$, $10°$, $\ldots$, $350°$, and draw them as arrows of equal length, one plot for each frame. (Use a two-argument arctangent, such as `atan2` or Mathematica's `ArcTan[x, y]`, so that each angle lands in the right quadrant.)

(c) How many of the 36 arrows point within 5° of the $+x$ axis in Luke's frame? In the jet's frame only one does. Jets that point nearly at Earth are seen as *blazars*, among the brightest objects in the sky.

(d) Repeat (c) for $0.99c$.

P25.16 The *rapidity* w of a transformation is defined by $\tanh w = \beta$.

(a) Show that $\gamma = \cosh w$ and $\gamma\beta = \sinh w$, so that the 2×2 part of the matrix (25.9) has the entries $\cosh w$ and $-\sinh w$.

(b) Show that two transformations along x, with rapidities w_1 and w_2, make a single transformation with rapidity $w_1 + w_2$, and use the result to derive (25.12).

(c) Compare with the rotation matrix (8.2). What plays the role of the angle, and what is preserved in place of $x^2 + y^2$?

Exercises for 25.7 Four-vectors

P25.17 Section 25.7.1 showed that the scalar product of a four-vector with itself has the same value in every inertial frame. This exercise shows the same for the scalar product of two different four-vectors, a^μ and b^μ.

(a) Write the components of a^μ and b^μ in Rachel's frame in terms of their components in Luke's, with the Lorentz transformation (25.9), and show that

$$-a^0_{\mathrm{R}}b^0_{\mathrm{R}} + a^1_{\mathrm{R}}b^1_{\mathrm{R}} + a^2_{\mathrm{R}}b^2_{\mathrm{R}} + a^3_{\mathrm{R}}b^3_{\mathrm{R}} = -a^0_{\mathrm{L}}b^0_{\mathrm{L}} + a^1_{\mathrm{L}}b^1_{\mathrm{L}} + a^2_{\mathrm{L}}b^2_{\mathrm{L}} + a^3_{\mathrm{L}}b^3_{\mathrm{L}}.$$

(b) Show the same result another way. Show that

$$a_\mu b^\mu = \frac{1}{4}\left[(a+b)_\mu (a+b)^\mu - (a-b)_\mu (a-b)^\mu\right],$$

where $(a+b)^\mu = a^\mu + b^\mu$ and $(a-b)^\mu = a^\mu - b^\mu$, and explain why this proves that $a_\mu b^\mu$ is a Lorentz scalar.

P25.18 Let a^μ and b^μ be four-vectors.

(a) Show that the 16 products $t^{\mu\nu} = a^\mu b^\nu$ form a four-tensor of rank 2: that is, that they transform by (25.19).

(b) The *trace* of a rank-2 four-tensor is formed by lowering one index and summing, $t^\mu{}_\mu = -t^{00} + t^{11} + t^{22} + t^{33}$. Show that for $t^{\mu\nu} = a^\mu b^\nu$ the trace is the scalar product $a_\mu b^\mu$, so that it has the same value in every inertial frame.

(c) Let $a^\mu = b^\mu = \eta^\mu$, the four-velocity of a ball that moves along x at $0.6c$ in Luke's frame. Find the components of $t^{\mu\nu}$ that are not zero, in the ball's own frame and in Luke's, and find the trace in each frame. Find also the plain sum of the diagonal components, $t^{00} + t^{11} + t^{22} + t^{33}$, in each frame. Which of the two sums is a Lorentz scalar?

Exercises for 25.8 Proper time and four-velocity

P25.19 In Example 25.4, Rachel passes Luke at $0.7c$ and throws a ball ahead at $0.6c$ in her frame, so that it moves at $0.915c$ in Luke's.

(a) Find the ball's four-velocity η^μ in its own frame, in Rachel's frame, and in Luke's, in units of c. (Only η^0 and η^1 are not zero.)

(b) Check that $\eta_\mu \eta^\mu = -c^2$ in all three frames.

(c) Check that the Lorentz transformation from Luke's frame to Rachel's, with $\beta = 0.7$, takes the ball's four-velocity in Luke's frame to its four-velocity in Rachel's.

(d) The ball covers 1 ly in Luke's frame. Use η^1 in Luke's frame to find the time that passes on a clock carried by the ball, and compare with Example 25.4.

Exercises for 25.9 Four-acceleration and the instantaneous rest frame

P25.20 Storage rings that make synchrotron light, the radiation of fast electrons on curved paths (Section 27.6), keep electrons with energies of a few GeV on circles with radii of several meters in their bending magnets. Take an electron with the energy $E = 3.0\,\text{GeV}$, moving at constant speed on a circle of radius $R = 8.0\,\text{m}$.

(a) Find its Lorentz factor γ_u, and write its speed as $(1-\delta)c$.

(b) Find its acceleration in the laboratory, in m/s^2 and in units of $g = 9.80\,\text{m/s}^2$.

(c) Find its proper acceleration in two ways: from (25.26), and from (25.27), in the frame that moves with the electron's velocity at one instant. Show that $a_{\text{proper}} = \gamma_u^2 u^2/R$.

P25.21 Rachel plans a trip from Earth to the center of our galaxy, 8178 pc away ("A geometric distance measurement to the Galactic center black hole with 0.3% uncertainty," GRAVITY Collaboration, Astron. Astrophys. 625, L10 (2019), https://doi.org/10.1051/0004-6361/201935656; 1 pc = 3.262 ly). Her ship accelerates with the proper acceleration $g = 9.80\,\text{m/s}^2$ for the

first half of the distance, as in Example 25.5, and then decelerates with the same proper acceleration for the second half, so that it arrives at rest. Ignore the motions of Earth and of the galactic center.

(a) How long does the trip take by Rachel's clock?

(b) How long does it take in Earth's frame?

(c) What is her greatest speed? Write it as $(1-\delta)c$, and find her greatest Lorentz factor.

P25.22 Rachel's ship in Example 25.5 keeps accelerating with the proper acceleration g forever, while Luke stays on Earth at $x = 0$. Her world line is the hyperbola $(x + c^2/g)^2 - (ct)^2 = (c^2/g)^2$.

(a) Luke sends a light signal after her at the time t_s. Show that it reaches her only if $t_s < c/g$.

(b) Draw Rachel's world line and the light line $x = c(t - c/g)$ on one spacetime diagram, and show that her world line comes closer and closer to the line but never reaches it. No event on or above the line can ever send her a signal: the line is a *horizon* for her.

(c) Rachel watches Luke's clock through a telescope. Show that when her own clock reads τ, she sees his clock read $(c/g)\left(1 - e^{-g\tau/c}\right)$. What does she see it read after 1, 2, and 10 yr of her time?

(d) In Luke's frame, what does Rachel's clock read at the moment his clock reads c/g?

Exercises for 25.10 Energy and momentum

P25.23 The Sun radiates energy at the rate $L = 3.828 \times 10^{26}$ W, and its mass is $M = 1.989 \times 10^{30}$ kg. (L is the nominal value of "Nominal values for selected solar and planetary quantities: IAU 2015 Resolution B3," A. Prša et al., Astron. J. 152, 41 (2016), https://doi.org/10.3847/0004-6256/152/2/41; M is the resolution's nominal value of GM, divided by the CODATA 2022 value of G.)

(a) How much mass does the Sun lose each second by radiating? (It also loses mass in other ways, which this exercise leaves out.)

(b) At this rate, how long would it take the Sun to lose 0.1% of its mass?

(c) A sealed bottle holding 0.50 kg of hot water cools by 60 K. By how much does its mass decrease? (The specific heat of water is 4.18 kJ/(kg·K).) Could the change be found by weighing the bottle?

Exercises for 25.11 Collisions

P25.24 A group of particles has the total energy E_{tot} and the total momentum $\mathbf{p}_{\text{tot}}$ in Luke's frame.

(a) Show that the center-of-momentum frame of the group moves with the velocity

$$\mathbf{v} = \frac{\mathbf{p}_{\text{tot}} c^2}{E_{\text{tot}}}$$

relative to Luke. (The total four-momentum is a four-vector. Take $\mathbf{p}_{\text{tot}}$ along x, transform the total four-momentum to the frame of an observer who moves along x at v, and find the v for which its space part is zero.)

(b) Check the result with the lumps of Example 25.6, using the totals in Rachel's frame.

(c) Compare the result with (25.32). What does it give for a single photon? Does a pair of photons, moving in opposite directions with energies E_1 and E_2, have a center-of-momentum frame?

P25.25 In Luke's frame, a particle of mass m with the energy $3mc^2$ hits an identical particle at rest, and the two stick together.

(a) Find the mass and the speed of the new particle, from the conservation of energy and momentum in Luke's frame.

(b) Find the mass again from the scalar product $P_\mu P^\mu$ of the total four-momentum P^μ, evaluated before the collision in Luke's frame and after it in the rest frame of the new particle.

(c) How much kinetic energy becomes rest energy in the collision, in Luke's frame? Show that the answer is the same in every inertial frame.

P25.26 The linear accelerator at SLAC, in California, is two miles (3.2 km) long. After upgrades in the 1980s, it could accelerate electrons and positrons to energies of 50 GeV ("The Stanford Linear Accelerator Center," https://www.slac.stanford.edu/gen/grad/GradHandbook/slac.html).

(a) How fast does a 50 GeV electron move? Write its speed as $(1-\delta)c$ and find δ. (A calculator may round β to 1; use $1/\gamma^2 = (1-\beta)(1+\beta)$ instead.)

(b) Suppose that a 50 GeV electron and a 50 GeV positron, moving in opposite directions, collide and make a tau lepton and its antiparticle, each of mass 1776.93 MeV/c^2 (Particle Data Group, the reference of Example 25.1). Explain why the two have equal energies, and find their speed. (The tau was discovered at SLAC

in 1974–1977, in electron–positron collisions at much lower energies, about 5 GeV in all: https://www.nobelprize.org/prizes/physics/1995/press-release/.)

P25.27 Two particles of mass m, each with the energy E in Luke's frame, approach each other head on.

(a) Show that in the rest frame of one of them, the other has the energy

$$\frac{2E^2}{mc^2} - mc^2.$$

(One way is to transform the second particle's four-momentum to the frame of the first. Another is to evaluate the scalar product $p_{1\mu}p_2^{\mu}$ of their four-momenta in both frames.)

(b) At CERN's Large Hadron Collider, protons with energies of 6.8 TeV collide head on (CERN, "LHC Run 3 media kit," https://home.cern/press/media-kits/run-3/). What energy would a proton need to make the same collision with a proton at rest?

(c) Use the result of (b) to explain why particle physicists collide two beams head on rather than aim one beam at a target at rest.

P25.28 Antiprotons were first observed in 1955 at the Bevatron, in Berkeley, California, made by a beam of protons that struck a copper target ("Observation of antiprotons," O. Chamberlain, E. Segrè, C. Wiegand, and T. Ypsilantis, Phys. Rev. 100, 947–950 (1955), https://doi.org/10.1103/PhysRev.100.947). The simplest reaction that makes one is $p + p \rightarrow p + p + p + \bar{p}$: a proton of the beam hits a proton at rest, and an antiproton, which has the mass of a proton and the opposite charge, is made together with a third proton. Treat the protons of the target as free and at rest.

(a) What is the least kinetic energy that the beam proton must have? (At this threshold, the four particles after the collision are at rest together in the center-of-momentum frame. Evaluate the scalar product $P_{\mu}P^{\mu}$ of the total four-momentum before the collision in the laboratory and after it in the center-of-momentum frame.)

(b) If instead two protons with equal energies collide head on, what is the least kinetic energy that each must have?

(c) In (a), what fraction of the beam proton's kinetic energy becomes the rest energy of the new particles?

Exercises for 25.12 Relativistic dynamics

P25.29 An electron moves on a circle perpendicular to a uniform magnetic field $B = 1.00\,\text{T}$.

(a) Find the radius of the circle and the frequency $\nu = \omega/2\pi$ of the motion for kinetic energies of 1.00 MeV and 1.00 GeV.

(b) Repeat (a) with Newtonian mechanics, using the same kinetic energies. Which of the Newtonian results are impossible?

(c) Up to what kinetic energy is the Newtonian frequency within 1% of the right one?

P25.30 An electron starts from rest in a uniform electric field of magnitude 1.00 MV/m.

(a) How long does it take to reach $0.99c$, and how far does it travel in that time? (Use the results of Section 25.12 for a constant force.)

(b) Answer the same questions with Newtonian mechanics.

(c) Check that the work done on the electron by the field equals its kinetic energy at $0.99c$.

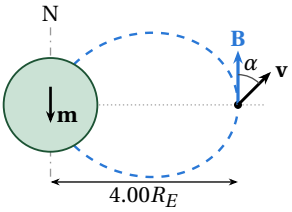


Figure 25.26 Exercise P25.31: Earth's dipole, the field line through $4R_E$ on the equator (dashed), and the electron's starting point and velocity.

Exercises for 25.13 Numerical method: relativistic trajectories

P25.31 Electrons are trapped in Earth's field in the outer Van Allen belt, where many of them move fast enough that relativity matters. Repeat the simulation of Exercise P10.5 for one of them (Figure 25.26). Use the same dipole model of Earth's field, but start the electron, of charge $-e$, at $\mathbf{r}_0 = 4.00R_E\hat{\mathbf{x}}$, on the equator, with a kinetic energy of 1.00 MeV and the velocity $\mathbf{u} = u\sin\alpha\,\hat{\mathbf{x}} + u\cos\alpha\,\hat{\mathbf{z}}$, with $\alpha = 45.0°$.

(a) Find the electron's Lorentz factor and speed, $\mathbf{B}(\mathbf{r}_0)$, and the radius and period of its circular motion there.

(b) Integrate the motion with the relativistic Boris method of Section 25.13, using 50 steps per period of the circular motion, for two complete bounces between the northern and southern hemispheres. Find the greatest magnetic latitude the electron reaches and the bounce period, the time between successive northward crossings of the equatorial plane. Check that the speed stays constant.

(c) Run the simulation again with the Newtonian Boris method of Section 10.1.5, starting with the same speed. Which of the results of (a) and (b) change, and which do not? Explain why, with (25.39).

(d) Why could the Newtonian simulation not start with the same kinetic energy instead?

Exercises for 25.14 Summary

P25.32 True or false? Explain your reasoning.

(a) On a spacetime diagram with ct up and x across, the world line of a particle that has mass can have any slope.

(b) Two events that are simultaneous in Luke's frame are simultaneous in Rachel's frame too.

(c) If one event can cause another, every inertial observer agrees about which of the two happened first.

(d) Rachel and Luke each find that the other's clocks run slow, and both are right.

(e) The time between two events on a clock that is present at both is the same for every inertial observer.

(f) The space part of the four-velocity of an object can be larger than c.

(g) Mass is conserved in collisions.

(h) The energy of a particle is the same in every inertial frame.

(i) A photon, which has no mass, has no momentum.

(j) A force along the motion of a fast particle gives it the same acceleration as an equal force across its motion.

P25.33 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as the Lorentz transformation of a grid of events, the twin paradox with the traveler's lines of simultaneity, relativistic beaming, hyperbolic motion and its horizon, relativistic collisions and decays, or the trajectories of fast charged particles in a cyclotron.

Chapter 26

Relativity and Electromagnetic Fields

Chapter 25 showed that two observers in relative motion disagree about lengths, times, energies, and forces, and it found the rules that translate one observer's measurements into the other's. This chapter does the same for the electric and magnetic fields. A charge that Luke holds at rest produces only an electric field in his frame. Rachel, moving past, sees a moving charge, which is a current, and so she finds a magnetic field as well. What one observer calls an electric field, another calls partly magnetic, and the main task of this chapter is to find the exact rule that relates the fields measured in two frames.

The chapter begins with a calculation that shows how much relativity has to say about magnetism: starting with nothing but the electric field of a line of charge and the transformation of forces, it predicts the magnetic field of a current and the magnetic force on a moving charge. It then finds how **E** and **B** transform in general, and shows that they are not two separate vectors, neither of which could be part of a four-vector, but together form a single four-tensor. In four-vector notation Maxwell's four equations become two, the potentials V and **A** become the components of one four-vector, and the equations of electromagnetism take the same form for every inertial observer, as Einstein's first postulate requires. The rest of the chapter applies these tools to the fields of a uniformly moving charge, the Doppler effect for light, the energy and momentum carried by the fields, and the motion of a charge in crossed electric and magnetic fields.

26.1 Magnetism from relativity

Suppose that magnetic fields and their forces had never been discovered, and that we knew only the electric field of a line of charge, the invariance of charge, and the relativity of Chapter 25. This section shows that relativity would then have forced the discovery of magnetism: an electric field in one frame, together with the rule for transforming forces from one frame to another, predicts in a second

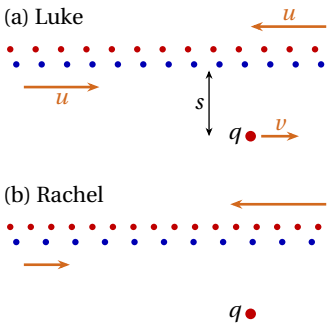


Figure 26.1 Two lines of charge, drawn apart for clarity. (a) In Luke's frame the negative charges (blue) move to the right at u and the positive charges (red) to the left at u, equally spaced, and Rachel's charge q moves to the right at v. (b) In Rachel's frame q is at rest, and the positive charges are closer together than the negative ones. Animation.

frame a new kind of force, with exactly the properties of the magnetic force.[1]

Two long, straight lines of charge lie along the x axis, so close together that they act as a single line (Figure 26.1). In its own rest frame one line has the charge per unit length $-\lambda_0$, and the other $+\lambda_0$. In Luke's frame the negative line moves to the right at speed u, and the positive line moves to the left at the same speed. Rachel moves to the right at v relative to Luke, as usual.

Luke's view. In Luke's frame each line is contracted along its length by the factor $\gamma_u = 1/\sqrt{1-u^2/c^2}$, and its charge is unchanged (Table 25.1), so each has a charge per unit length of magnitude

$$\lambda_\mathrm{L} = \gamma_u \lambda_0.$$

The two densities are equal and opposite, so the lines produce no electric field in Luke's frame. They do carry a current: the negative charges move to the right and the positive charges to the left, so each line carries a current $\lambda_\mathrm{L} u$ to the left, and the total current in the $+x$ direction is

$$I_\mathrm{L} = -2\lambda_\mathrm{L} u. \tag{26.1}$$

Rachel's view. In Rachel's frame the two lines move at different speeds, so they are contracted by different factors, and their charges no longer cancel. The Lorentz factor of an object in her frame follows from its Lorentz factor in Luke's frame by the time component of $\eta^\mu_\mathrm{R} = \Lambda^\mu{}_\nu \eta^\nu_\mathrm{L}$ (Section 25.8), $\eta^0_\mathrm{R} = \gamma_v(\eta^0_\mathrm{L} - \beta\eta^1_\mathrm{L})$, which gives

$$\gamma_{u,\mathrm{R}} = \gamma_v \gamma_{u,\mathrm{L}} \left(1 - \frac{u_{\mathrm{L},x} v}{c^2}\right), \tag{26.2}$$

the inverse of (25.24). For the negative line $u_{\mathrm{L},x} = u$, and for the positive line $u_{\mathrm{L},x} = -u$, so their charges per unit length in Rachel's frame are $-\gamma_v\gamma_u\lambda_0(1 - uv/c^2)$ and $+\gamma_v\gamma_u\lambda_0(1 + uv/c^2)$. Their sum is the net charge per unit length that she measures,

$$\lambda_\mathrm{R} = \frac{2\gamma_v \lambda_\mathrm{L} u v}{c^2}. \tag{26.3}$$

The positive line, which moves faster in her frame, is the more contracted, and the pair carries a net positive charge.

Now let Rachel hold a charge q at rest, a distance s from the lines. In her frame q feels only the electric field of a line charge λ_R, (3.17), so the force on it is

$$\mathbf{F}_\mathrm{R} = \frac{q\lambda_\mathrm{R}}{2\pi\varepsilon_0 s}\hat{\mathbf{s}} = \frac{q}{2\pi\varepsilon_0 s}\frac{2\gamma_v\lambda_\mathrm{L} u v}{c^2}\hat{\mathbf{s}},$$

where s and $\hat{\mathbf{s}}$ are cylindrical coordinates about the x axis, the direction of the relative motion.

[1] Richard Feynman makes the same argument for a charge moving beside a wire that carries a current: *The Feynman Lectures on Physics*, Vol. II, Section 13-6. https://feynmanlectures.caltech.edu/II_13.html.

What force does Luke measure on the same charge? The charge is at rest in Rachel's frame, and the force is perpendicular to the relative motion, so by (25.41) the force in Luke's frame is smaller by γ_v:

$$\mathbf{F}_L = \frac{\mathbf{F}_R}{\gamma_v} = \frac{q}{2\pi\varepsilon_0 s}\frac{2\lambda_L u v}{c^2}\hat{\mathbf{s}}. \tag{26.4}$$

There is no electric field in Luke's frame, and yet the charge feels a force. Three facts about this force stand out. It is proportional to the charge q. It is proportional to the charge's speed v: a charge at rest in Luke's frame would feel no force at all. And it is perpendicular to the charge's velocity. Something other than an electric field must produce it. To see what, write the force in terms of Luke's current (26.1), with $1/\varepsilon_0 c^2 = \mu_0$:

$$\mathbf{F}_L = -qv\frac{\mu_0 I_L}{2\pi s}\hat{\mathbf{s}}.$$

For cylindrical coordinates about the x axis, $\hat{\mathbf{x}}\times\hat{\boldsymbol{\phi}} = -\hat{\mathbf{s}}$, so this is

$$\mathbf{F}_L = q\mathbf{v}\times\mathbf{B}, \qquad \text{with} \qquad \mathbf{B} = \frac{\mu_0 I_L}{2\pi s}\hat{\boldsymbol{\phi}},$$

where $\mathbf{v} = v\hat{\mathbf{x}}$ is the charge's velocity. The new field is the magnetic field of a long, straight current, (10.20), and the force is the magnetic force of the Lorentz force law, (10.2) (Figure 26.2). (With q, u, and v all positive, the force pushes q away from the lines: the moving charge is a current to the right, and I_L flows to the left, and currents in opposite directions repel.)

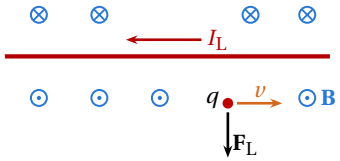


Figure 26.2 Luke's frame: the lines carry the current I_L to the left, its magnetic field (blue) points out of the page below them, and the force $q\mathbf{v}\times\mathbf{B}$ on the moving charge points away from the lines.

Even had people never previously known about magnetic fields, they would have theoretically predicted them after Einstein's theory of relativity and the Lorentz transformations were used to derive the transformation of forces. In short, Gauss's law for **E**, together with relativity, gives Ampère's law and the Lorentz force. The calculation used one particular arrangement of charges; the next section finds the general rule.

26.2 The transformation of E and B

Given the fields $\mathbf{E}_L$ and $\mathbf{B}_L$ that Luke measures at some event, what fields $\mathbf{E}_R$ and $\mathbf{B}_R$ does Rachel measure at the same event? A field can be produced in many ways, and the answer should not depend on which. The fields at a point determine the force on any charge placed there, and no measurement made at that point can tell what produced them. So we choose the simplest sources that produce a uniform field, parallel plates and a long solenoid, find how their fields transform, and expect every field to transform in the same way. This is a plausibility argument, not a proof; Section 26.6.1 supplies the proof.

26.2.1 Moving capacitor plates: the electric field

Let a large parallel-plate capacitor be at rest in Luke's frame, with its plates parallel to the xz plane, a surface charge density σ_0 on the lower plate and $-\sigma_0$ on

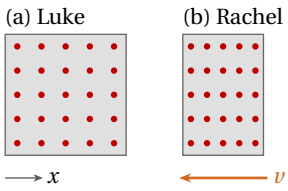


Figure 26.3 One plate of the capacitor, seen face on: (a) at rest in Luke's frame; (b) in Rachel's frame, moving to the left, contracted along x with the same charge, so that its surface charge density is larger by γ (drawn for $\gamma = 1.5$).

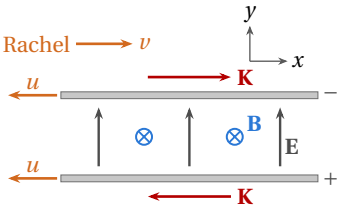


Figure 26.4 A parallel-plate capacitor whose plates move to the left at u in Luke's frame. Its moving charges are surface currents **K**, to the left on the positive plate and to the right on the negative one, so between the plates Luke finds a magnetic field (into the page) as well as the electric field **E**. Rachel moves to the right at v relative to Luke.

the upper one (Figure 26.3). Between the plates $\mathbf{E}_\text{L} = (\sigma_0/\varepsilon_0)\hat{\mathbf{y}}$, and there is no magnetic field.

In Rachel's frame the plates move to the left at v. They are contracted along x by the factor γ, their extent along z is not, and their charge is invariant, so their surface charge density is larger: $\sigma_\text{R} = \gamma\sigma_0$. The laws of physics, Maxwell's equations among them, are the same in Rachel's frame (Postulate 1), and by symmetry the field of each very large plate is still perpendicular to it, with the magnitude $\sigma/2\varepsilon_0$ that Gauss's law gives. So between the plates

$$\mathbf{E}_\text{R} = \frac{\gamma\sigma_0}{\varepsilon_0}\hat{\mathbf{y}} = \gamma\,\mathbf{E}_\text{L}.$$

The component of **E** perpendicular to the relative motion is larger, by γ, in the frame in which the plates move. Plates parallel to the xy plane give the same result for E_z.

Now turn the capacitor so that its plates are perpendicular to the x axis, with $\mathbf{E}_\text{L} = (\sigma_0/\varepsilon_0)\hat{\mathbf{x}}$. In Rachel's frame the contraction now shortens the gap between the plates and leaves their area, and so their surface charge density, unchanged. The field between two large plates does not depend on the gap, so $E_{\text{R},x} = E_{\text{L},x}$: the component of **E** along the relative motion is the same in both frames.

In Rachel's frame the charges on the plates of the first capacitor move, so they are surface currents, and she also measures a magnetic field between them. The next subsection finds it, for the general case in which the plates move in both frames.

26.2.2 Moving plates and a moving solenoid: the magnetic field

Let the plates of the capacitor, again parallel to the xz plane, move to the left at speed u in Luke's frame, with surface charge densities $\pm\sigma_0$ in their own rest frame (Figure 26.4). Three frames are now involved: the plates' own frame; Luke's, in which the plates move to the left at u; and Rachel's, in which they move to the left at

$$u_\text{R} = \frac{u+v}{1+uv/c^2}$$

by velocity addition (25.12). Each of the three speeds, u, v, and u_R, has its own Lorentz factor.

In Luke's frame the surface charge density of the lower plate is $\sigma_\text{L} = \gamma_u\sigma_0$, with $\gamma_u = 1/\sqrt{1-u^2/c^2}$, and the field between the plates is

$$E_{\text{L},y} = \frac{\sigma_\text{L}}{\varepsilon_0}.$$

The moving charges are surface currents: the lower plate, positive and moving to the left, carries the surface current density $\mathbf{K} = -\sigma_\text{L}u\hat{\mathbf{x}}$, and the upper plate, negative, carries $+\sigma_\text{L}u\hat{\mathbf{x}}$. Between two large sheets that carry opposite currents the magnetic field is $\mu_0 K$ (Section 11.2.4), and outside them it is zero. By $\mathbf{B} = (\mu_0/2)\mathbf{K}\times\hat{\mathbf{n}}$, (11.6), the field between these plates points along $-z$:

$$B_{\text{L},z} = -\mu_0\sigma_\text{L}u = -\frac{u}{c^2}E_{\text{L},y},$$

since $\mu_0\varepsilon_0 = 1/c^2$. In Rachel's frame the same steps give $E_{\mathrm{R},y} = \sigma_\mathrm{R}/\varepsilon_0$ and $B_{\mathrm{R},z} = -\mu_0\sigma_\mathrm{R}u_\mathrm{R}$, with $\sigma_\mathrm{R} = \gamma_{u,\mathrm{R}}\sigma_0$, where $\gamma_{u,\mathrm{R}}$ is the plates' Lorentz factor in her frame.

To compare the two, we need the ratio of the plates' Lorentz factors in the two frames, which (26.2) gives with $u_{\mathrm{L},x} = -u$:

$$\gamma_{u,\mathrm{R}} = \gamma_v\gamma_u\left(1+\frac{uv}{c^2}\right), \qquad \text{so} \qquad \sigma_\mathrm{R} = \gamma_v\left(1+\frac{uv}{c^2}\right)\sigma_\mathrm{L}.$$

Then

$$E_{\mathrm{R},y} = \gamma_v\left(1+\frac{uv}{c^2}\right)\frac{\sigma_\mathrm{L}}{\varepsilon_0} = \gamma_v\left(E_{\mathrm{L},y}+\frac{uv}{c^2}E_{\mathrm{L},y}\right) = \gamma_v\left(E_{\mathrm{L},y}-vB_{\mathrm{L},z}\right),$$

and, since $u_\mathrm{R}(1+uv/c^2) = u+v$,

$$B_{\mathrm{R},z} = -\mu_0\gamma_v\sigma_\mathrm{L}(u+v) = \gamma_v\left(B_{\mathrm{L},z}-\frac{v}{c^2}E_{\mathrm{L},y}\right).$$

Part of Luke's magnetic field appears in Rachel's frame as electric field, and part of his electric field appears as magnetic field. Turning the plates so that they are parallel to the xy plane gives, in the same way, $E_{\mathrm{R},z} = \gamma_v(E_{\mathrm{L},z}+vB_{\mathrm{L},y})$ and $B_{\mathrm{R},y} = \gamma_v(B_{\mathrm{L},y}+vE_{\mathrm{L},z}/c^2)$. These relations were found for fields whose **E** and **B** are tied together by the plates' speed, $B_{\mathrm{L},z} = -(u/c^2)E_{\mathrm{L},y}$; but any pair of values of $E_{\mathrm{L},y}$ and $B_{\mathrm{L},z}$ is the sum of the fields of two capacitors that move at different speeds, and the relations are linear, so they hold for every pair.

The component of **B** along the relative motion remains. A long solenoid at rest in Luke's frame, with its axis along x, n turns per unit length, and current I, has the field $B_{\mathrm{L},x} = \mu_0 nI$ inside, (11.7). In Rachel's frame the solenoid is contracted along its axis, so it has γn turns per unit length (Figure 26.5). The current in each turn is smaller. The charge that passes a given point of the wire between two ticks of a clock that moves with the wire is the same in both frames, but Rachel measures the time between those ticks as longer by the factor γ, so in her frame the current is I/γ. Then

$$B_{\mathrm{R},x} = \mu_0\left(\gamma n\right)\frac{I}{\gamma} = B_{\mathrm{L},x}:$$

the component of **B** along the relative motion is the same in both frames.

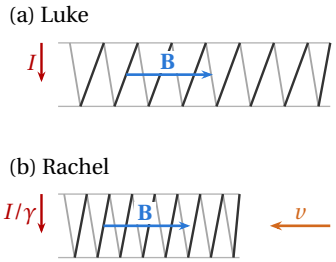


Figure 26.5 A solenoid with its axis along x, (a) at rest in Luke's frame and (b) in Rachel's frame, where it moves to the left at v, its turns are closer together by γ, and its current is smaller by γ. The field inside is the same.

26.2.3 The transformation rules

Collecting the results: if Rachel moves along x at v relative to Luke, the fields she measures at an event are related to the fields that Luke measures at the same event by

$$\begin{array}{ll} E_{\mathrm{R},x} = E_{\mathrm{L},x}, & B_{\mathrm{R},x} = B_{\mathrm{L},x}, \\ E_{\mathrm{R},y} = \gamma\left(E_{\mathrm{L},y}-vB_{\mathrm{L},z}\right), & B_{\mathrm{R},y} = \gamma\left(B_{\mathrm{L},y}+\dfrac{v}{c^2}E_{\mathrm{L},z}\right), \\ E_{\mathrm{R},z} = \gamma\left(E_{\mathrm{L},z}+vB_{\mathrm{L},y}\right), & B_{\mathrm{R},z} = \gamma\left(B_{\mathrm{L},z}-\dfrac{v}{c^2}E_{\mathrm{L},y}\right), \end{array} \tag{26.5}$$

with $\gamma = 1/\sqrt{1 - v^2/c^2}$. The inverse transformation, from Rachel's fields to Luke's, is found by changing the sign of v. In vector form, with $\mathbf{v} = v\hat{\mathbf{x}}$ Rachel's velocity relative to Luke, and with $\parallel$ and $\perp$ meaning the components along and across $\mathbf{v}$,

$$\begin{aligned} \mathbf{E}_{R\parallel} &= \mathbf{E}_{L\parallel}, \quad \mathbf{E}_{R\perp} = \gamma\,(\mathbf{E}_L + \mathbf{v}\times\mathbf{B}_L)_\perp\,, \\ \mathbf{B}_{R\parallel} &= \mathbf{B}_{L\parallel}, \quad \mathbf{B}_{R\perp} = \gamma\left(\mathbf{B}_L - \frac{\mathbf{v}\times\mathbf{E}_L}{c^2}\right)_\perp. \end{aligned} \tag{26.6}$$

The vector form holds for a relative velocity in any direction. The components of the fields along the relative motion do not change; the components across it mix.

The combination $\mathbf{E}_L + \mathbf{v}\times\mathbf{B}_L$ is the force per unit charge, in Luke's frame, on a charge that moves with Rachel. In her frame that charge is at rest, and the force on it, $q\mathbf{E}_R$, is purely electric. So the rule for $\mathbf{E}_{R\perp}$ is the rule (25.41) for the perpendicular components of a force, written per unit charge. Einstein put it this way: "If a unit electric point charge is in motion in an electromagnetic field, the force acting upon it is equal to the electric force which is present at the locality of the charge, and which we ascertain by transformation of the field to a system of co-ordinates at rest relatively to the electrical charge."[2]

Two special cases come up often. If Luke measures no magnetic field, as when all the charges that produce the field are at rest in his frame, then (26.6) gives

$$\mathbf{B}_R = -\frac{\mathbf{v}\times\mathbf{E}_R}{c^2} \qquad (\mathbf{B}_L = 0).$$

Rachel sees those charges move with velocity $-\mathbf{v}$, so the magnetic field of charges that all move with the same velocity $\mathbf{u}$ is $\mathbf{u}\times\mathbf{E}/c^2$, in terms of the electric field measured in the same frame. Section 26.7 uses this for a moving point charge. If instead Luke measures no electric field,

$$\mathbf{E}_R = \mathbf{v}\times\mathbf{B}_R \qquad (\mathbf{E}_L = 0).$$

Exercise P26.4 applies the rules to a moving sheet of charge, and Exercise P26.3 checks that two transformations in succession give the same result as one.

[2] A. Einstein, Ann. Phys. 17, 891–921 (1905), Section 6, in the translation of W. Perrett and G. B. Jeffery (1923); the full reference is in the footnote in Section 15.2.1. Section 6 of the paper derives (26.5) from the requirement that Maxwell's equations have the same form in every inertial frame.

Example 26.1

A **velocity selector** is an arrangement of electric and magnetic fields through which a charge with one particular velocity passes undeflected. Take x to the right, y up, and z out of the page, and let a positive charge q enter, moving to the right at speed v, a region where $\mathbf{E} = E_0\hat{\mathbf{y}}$. (a) What magnitude and direction must the magnetic field have for the magnetic force on the charge to cancel the electric force exactly? (b) In the charge's own frame, what are the electric and magnetic fields? What is the net force on the charge in the frame?

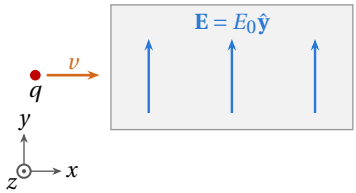
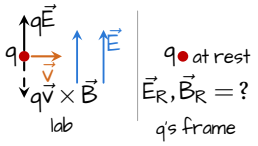


Figure 26.6 Example 26.1: a positive charge q enters a region of uniform electric field $\mathbf{E} = E_0\hat{\mathbf{y}}$ at speed v. The magnetic field is to be found.

Solution:



The first step is a sketch of the charge entering the field region, with the forces in the lab: $q\mathbf{E}$ up, so the magnetic force must point down. In the charge's frame, the second sketch, the charge is at rest.

(a) Call the lab frame Luke's. The electric force is $qE_0\hat{\mathbf{y}}$, so the magnetic force must be $-qE_0\hat{\mathbf{y}}$. A magnetic field along z does it: with $\mathbf{B} = B\hat{\mathbf{z}}$,

$$q\mathbf{v}\times\mathbf{B} = qvB\,\hat{\mathbf{x}}\times\hat{\mathbf{z}} = -qvB\,\hat{\mathbf{y}},$$

so we need $B = E_0/v$, pointing out of the page:

$$\mathbf{B}_\text{L} = \frac{E_0}{v}\hat{\mathbf{z}}.$$

A charge moving faster or slower than v is deflected, which is what makes the arrangement select velocities. (A component of $\mathbf{B}$ along x would exert no force and is not needed.)

(b) The charge's frame is Rachel's, moving along x at v. By (26.5),

$$E_{\text{R},x} = 0, \qquad E_{\text{R},y} = \gamma\left(E_0 - v\frac{E_0}{v}\right) = 0, \qquad E_{\text{R},z} = 0,$$

$$B_{\text{R},z} = \gamma\left(\frac{E_0}{v} - \frac{v}{c^2}E_0\right) = \gamma\frac{E_0}{v}\left(1 - \frac{v^2}{c^2}\right) = \frac{E_0}{\gamma v},$$

and $B_{\text{R},x} = B_{\text{R},y} = 0$. In the charge's frame there is no electric field at all, only a magnetic field $\mathbf{B}_\text{R} = (E_0/\gamma v)\hat{\mathbf{z}}$, smaller than the lab's by γ. The charge is at rest in this frame, so the magnetic field exerts no force on it, and the net force is zero. The two frames agree, as they must: a force that is zero in one frame is zero in every frame, by (25.40).

26.2.4 The magnet and the conductor

Section 15.2.1 described two ways of explaining the EMF induced when a magnet and a loop of wire move relative to each other. Suppose that Luke holds the

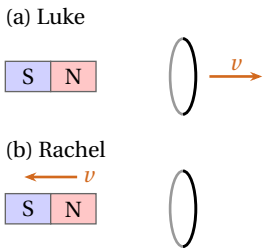


Figure 26.7 The magnet and the conductor. (a) In Luke's frame the magnet is at rest and the loop moves at v. (b) In Rachel's frame the loop is at rest and the magnet moves at v the other way.

magnet, and that Rachel carries the loop away from it at speed v (Figure 26.7). In Luke's frame there is no electric field, only the magnet's magnetic field $\mathbf{B}_\mathrm{L}$; the charges in the moving loop feel the magnetic force $q\mathbf{v}\times\mathbf{B}_\mathrm{L}$, and the EMF around the loop is the motional EMF of Section 15.1, given by the flux rule. In Rachel's frame the magnet moves, the magnetic field at each point of her loop changes, and by (26.6) there is an electric field,

$$\mathbf{E}_\mathrm{R} = \gamma\,\mathbf{v}\times\mathbf{B}_\mathrm{L}.$$

The charges in her loop are at rest, and they feel only the electric force $q\mathbf{E}_\mathrm{R}$; the EMF around the loop is the line integral of $\mathbf{E}_\mathrm{R}$, which Faraday's law gives as the rate of change of the flux. For a magnet and a loop in a laboratory, γ differs from 1 by much less than one part in a billion, and $\mathbf{E}_\mathrm{R} = \mathbf{v}\times\mathbf{B}_\mathrm{L}$: Luke's magnetic force on the moving charges is Rachel's electric force on the same charges. The two explanations of Section 15.2.1, a magnetic force on moving charges and an electric field produced by a changing magnetic field, are one physical effect seen from two frames. At higher speeds the factor γ makes the force in Rachel's frame larger than Luke's by just the factor that the force transformation (25.41) requires, so the two descriptions remain consistent at any speed (Exercise P26.5). In Einstein's words, "it is clear that the asymmetry mentioned in the introduction as arising when we consider the currents produced by the relative motion of a magnet and a conductor, now disappears."

26.3 The field tensor

The rules (26.5) show at once that neither **E** nor **B** can be the space part of a four-vector. A transformation along x changes only the time and x components of a four-vector, (25.16), and leaves its y and z components alone; but it changes E_y and E_z, by mixing them with components of **B**. The six components of the two fields must be transformed together. They are, in fact, the six independent components of a four-tensor of rank 2 (Section 25.7.3), the **field tensor**

$$F^{\mu\nu} = \begin{pmatrix} 0 & E_x/c & E_y/c & E_z/c \\ -E_x/c & 0 & B_z & -B_y \\ -E_y/c & -B_z & 0 & B_x \\ -E_z/c & B_y & -B_x & 0 \end{pmatrix}, \tag{26.7}$$

with its rows and columns numbered from 0, as for Λ. The tensor is antisymmetric, $F^{\nu\mu} = -F^{\mu\nu}$, so its diagonal entries are zero and only six of its entries are independent: the components of $\mathbf{E}/c$ in the time row, $F^{0i} = E_i/c$, and those of **B** in the space block, $F^{12} = B_z$, $F^{31} = B_y$, $F^{23} = B_x$. (The components of **E** are divided by c so that every entry has the units of a magnetic field.)

The claim is that $F^{\mu\nu}$ transforms as a four-tensor, (25.19):

$$F^{\mu\nu}_\mathrm{R} = \Lambda^\mu{}_\lambda \Lambda^\nu{}_\sigma\, F^{\lambda\sigma}_\mathrm{L}, \tag{26.8}$$

with sums over λ and σ. That is sixteen equations, each with a double sum. It is not quite a matrix multiplication, because the second factor of Λ acts on the column index of F_L; written with matrices it is $F_\mathrm{R} = \Lambda F_\mathrm{L} \Lambda^\mathsf{T}$, with the transpose of Λ on the right. To check it, take the component $F^{02} = E_y/c$. Since ${\Lambda^2}_\sigma$ is 1 for $\sigma = 2$ and zero otherwise, and ${\Lambda^0}_\lambda$ is γ for $\lambda = 0$, $-\gamma\beta$ for $\lambda = 1$, and zero otherwise,

$$F_\mathrm{R}^{02} = {\Lambda^0}_\lambda F_\mathrm{L}^{\lambda 2} = \gamma F_\mathrm{L}^{02} - \gamma\beta F_\mathrm{L}^{12}, \qquad \text{or} \qquad \frac{E_{\mathrm{R},y}}{c} = \gamma\left(\frac{E_{\mathrm{L},y}}{c} - \frac{v}{c} B_{\mathrm{L},z}\right),$$

which is the rule for $E_{\mathrm{R},y}$ in (26.5). For the component F^{01} both sums run over $\lambda, \sigma = 0, 1$, and since $F^{00} = F^{11} = 0$ and $F^{10} = -F^{01}$,

$$F_\mathrm{R}^{01} = \gamma^2 F_\mathrm{L}^{01} + \gamma^2\beta^2 F_\mathrm{L}^{10} = \gamma^2\left(1 - \beta^2\right) F_\mathrm{L}^{01} = F_\mathrm{L}^{01},$$

so $E_{\mathrm{R},x} = E_{\mathrm{L},x}$. The other components work out in the same way (Exercise P26.7), and (26.8) reproduces all six rules of (26.5).

Tips for Mathematica and Python. A computer algebra system can check all sixteen components at once. Enter Λ and F_L as 4×4 matrices, with symbols for the field components and $\gamma = 1/\sqrt{1-\beta^2}$, and compute $\Lambda F_\mathrm{L} \Lambda^\mathsf{T}$: in Mathematica, `FullSimplify[L.F.Transpose[L]]`, and in Python, with `sympy`, `(L*F*L.T)` followed by `.applyfunc(sympy.simplify)`. Then compare the result, entry by entry, with F_R built from the fields of (26.5). Writing the double sum of (26.8) out in code, with explicit loops over λ and σ, gives the same result and is a good test of your understanding of the index notation.

A second tensor, the **dual tensor**, is obtained from $F^{\mu\nu}$ by replacing $\mathbf{E}/c$ with $\mathbf{B}$ and $\mathbf{B}$ with $-\mathbf{E}/c$:

$$G^{\mu\nu} = \begin{pmatrix} 0 & B_x & B_y & B_z \\ -B_x & 0 & -E_z/c & E_y/c \\ -B_y & E_z/c & 0 & -E_x/c \\ -B_z & -E_y/c & E_x/c & 0 \end{pmatrix}. \tag{26.9}$$

The rules (26.5) are unchanged by this replacement (it turns the rule for $E_{\mathrm{R},y}$ into the rule for $B_{\mathrm{R},y}$, and so on), so $G^{\mu\nu}$ is a four-tensor too.

Products of tensors in which every index appears twice, once up and once down, are Lorentz scalars, as scalar products of four-vectors are (Section 25.7.1). Two such scalars can be made from the field tensors,

$$F_{\mu\nu}F^{\mu\nu} = 2\left(B^2 - \frac{E^2}{c^2}\right), \qquad G_{\mu\nu}F^{\mu\nu} = -\frac{4\,\mathbf{E}\cdot\mathbf{B}}{c} \tag{26.10}$$

(Exercise P26.6), so $E^2 - c^2B^2$ and $\mathbf{E}\cdot\mathbf{B}$ have the same values in every inertial frame. If the electric and magnetic fields at an event are perpendicular in one frame, for instance, they are perpendicular in every frame, and if $E = cB$ in one frame, then $E = cB$ in every frame. Exercise P26.8 uses the two scalars to decide what other observers can measure, and Section 26.10 uses them to find the frame in which crossed electric and magnetic fields look simplest.

26.4 The four-current and the four-gradient

The sources of the fields, the charge density ρ and the current density $\mathbf{J}$, also belong together. They form a four-vector, the **four-current**

$$J^\mu = \big(c\rho,\, J_x,\, J_y,\, J_z\big). \tag{26.11}$$

To see why, consider a distribution of charge whose charges all move with the same velocity $\mathbf{u}$. In its own rest frame it has a charge density ρ_0, called the **proper charge density**, and no current. In a frame in which it moves at $\mathbf{u}$, every volume is contracted along $\mathbf{u}$ by the factor γ_u and holds the same charge, so $\rho = \gamma_u\rho_0$ and $\mathbf{J} = \rho\mathbf{u} = \gamma_u\rho_0\mathbf{u}$. Together,

$$J^\mu = \rho_0\gamma_u\,(c, \mathbf{u}) = \rho_0\,\eta^\mu,$$

a Lorentz scalar times the four-velocity (25.22), which is a four-vector. When charges of several kinds move at different velocities, as the ions and the electrons of a wire do, the four-current is the sum of the four-currents of the different kinds, and a sum of four-vectors is a four-vector. The time component of J^μ is c times the charge density, just as the time component of the four-momentum is the energy divided by c (Table 26.1): charge density and current density are two parts of one quantity, as energy and momentum are. A current in one frame is therefore partly a charge density in another, which is what Section 26.1 found for its two lines of charge (Exercise P26.11).

	Time part	Space part
x^μ	ct	$\mathbf{r}$
η^μ	$\gamma_u c$	$\gamma_u\mathbf{u}$
p^μ	E/c	$\mathbf{p}$
J^μ	$c\rho$	$\mathbf{J}$
A^μ	V/c	$\mathbf{A}$
k^μ	ω/c	$\mathbf{k}$

Table 26.1 The time and space parts of the four-vectors of Chapter 25 and this chapter.

A point charge q that moves along the world line $w^\mu(\tau)$, where τ is its proper time, has $\rho = q\delta^3(\mathbf{r} - \mathbf{w}(t))$ and $\mathbf{J} = q\mathbf{u}\delta^3(\mathbf{r} - \mathbf{w}(t))$, with $\mathbf{w}$ the space part of w^μ (Section 1.6.2). In four-vector form,

$$J^\mu(x) = qc\int \eta^\mu(\tau)\,\delta^4\big(x - w(\tau)\big)\,d\tau, \tag{26.12}$$

where $\delta^4(x) \equiv \delta(x^0)\,\delta(x^1)\,\delta(x^2)\,\delta(x^3) = \delta(ct)\,\delta^3(\mathbf{r})$. To check it, change the variable of integration from τ to the time t' at which the charge is at $\mathbf{w}(t')$, with $d\tau = dt'/\gamma_u$, and use $\delta(ct - ct') = \delta(t - t')/c$ (substituting $u = ct'$ in the integral over t' gives the factor $1/c$). The integral over t' picks out $t' = t$:

$$J^\mu = qc\gamma_u\,(c, \mathbf{u})\,\frac{1}{c}\,\frac{1}{\gamma_u}\,\delta^3\big(\mathbf{r} - \mathbf{w}(t)\big) = \big(cq,\, q\mathbf{u}\big)\,\delta^3\big(\mathbf{r} - \mathbf{w}(t)\big).$$

In (26.12) every factor but η^μ is a Lorentz scalar (δ^4 is one because its integral over spacetime is 1 in every frame and d^4x is invariant, (25.20)), so this form shows at a glance that the four-current of a point charge is a four-vector. Section 27.1 uses it to find the potentials of a moving point charge.

The four-vector forms of Maxwell's equations need one more ingredient, a four-dimensional gradient. Let f be a scalar field: a function of the event that has the same value at that event in every frame, such as the temperature measured by a thermometer there. Its change between two nearby events,

$$df = \frac{\partial f}{\partial x^\mu}\,dx^\mu = \frac{1}{c}\frac{\partial f}{\partial t}\,(c\,dt) + \frac{\partial f}{\partial x}\,dx + \frac{\partial f}{\partial y}\,dy + \frac{\partial f}{\partial z}\,dz,$$

is a scalar, and dx^μ is a four-vector. For the sum to be a scalar whatever dx^μ is, the four derivatives $\partial f/\partial x^\mu$ must be the lower-index components of a four-vector, so that the sum is a scalar product (25.18). The **four-gradient** is therefore written with its index down,

$$\partial_\mu \equiv \frac{\partial}{\partial x^\mu} = \left(\frac{1}{c}\frac{\partial}{\partial t}, \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right), \qquad \partial^\mu = \left(-\frac{1}{c}\frac{\partial}{\partial t}, \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right), \tag{26.13}$$

where the upper-index form follows from (25.17). The four-gradient obeys the rules of ordinary derivatives, the product rule and the chain rule among them: if f depends on the event only through a scalar s, then $\partial_\mu f = (df/ds)\,\partial_\mu s$. Its scalar product with itself is the d'Alembertian (24.6) of Chapter 24,

$$\partial^\mu \partial_\mu = -\frac{1}{c^2}\frac{\partial^2}{\partial t^2} + \nabla^2 = \Box^2, \tag{26.14}$$

which is therefore the same operator for every inertial observer, as Section 24.2 promised. The scalar product of the four-gradient with a four-vector field a^μ, $\partial_\mu a^\mu$, is called the **four-divergence** of a^μ, and it is a scalar too.

The simplest equation in tensor form is the statement that the four-divergence of the four-current is zero,

$$\partial_\mu J^\mu = 0. \tag{26.15}$$

Unpacking it shows how a tensor equation is read.

1. *In words:* the four-divergence of the four-current is zero.
2. *Units:* ∂_μ has the units of 1/m, and every component of J^μ has the units of $\mathrm{A/m^2}$ ($c\rho$ is in $(\mathrm{m/s})(\mathrm{C/m^3})$), so every term has the units of $\mathrm{A/m^3}$.
3. *Written out:* by the summation convention,
$$\partial_\mu J^\mu = \partial_0 J^0 + \partial_1 J^1 + \partial_2 J^2 + \partial_3 J^3 = \frac{1}{c}\frac{\partial (c\rho)}{\partial t} + \frac{\partial J_x}{\partial x} + \frac{\partial J_y}{\partial y} + \frac{\partial J_z}{\partial z} = \frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J}.$$
4. *Which equation:* $\partial_\mu J^\mu = 0$ is the continuity equation (20.5), the statement that charge is conserved locally.

Since $\partial_\mu J^\mu$ is a Lorentz scalar, if it is zero in one frame it is zero in every frame: charge is conserved for every inertial observer. Section 20.1 argued that relativity requires charge to be conserved locally; (26.15) is the form of that law that every observer writes in the same way.

26.5 Maxwell's equations in tensor form

With the field tensors, the four-current, and the four-gradient, Maxwell's four equations become two, **Maxwell's equations in tensor form**:

$$\boxed{\partial_\nu F^{\mu\nu} = \mu_0 J^\mu, \qquad \partial_\nu G^{\mu\nu} = 0}. \tag{26.16}$$

Each is four equations, one for each value of the free index μ, with a sum over ν, and together they contain all four of Maxwell's equations, as Tutorial 26.1 shows by writing them out. Each side of each equation is a four-vector with the index μ: the four-gradient is a four-vector with its index down, and summing it against one index of a tensor leaves a four-vector. An equation between four-vectors that holds in one inertial frame holds in every inertial frame, so in this form Maxwell's equations are plainly the same for every inertial observer, as Postulate 1 demands of them. Hermann Minkowski, who introduced the four-dimensional view of spacetime (Section 25.2), wrote the equations of electromagnetism in this four-dimensional form in 1908, in a paper on the electrodynamics of moving bodies.[3]

The force on a charge has a tensor form too. The Minkowski force (25.38) on a charge q that moves with four-velocity η^ν is

$$K^\mu = q\,\eta_\nu F^{\mu\nu}, \tag{26.17}$$

a four-vector, as it should be. Its time component, with $\eta_\nu = \gamma_u(-c, u_x, u_y, u_z)$ by (25.17) and $F^{00} = 0$, is

$$K^0 = q\left(\eta_1 F^{01} + \eta_2 F^{02} + \eta_3 F^{03}\right) = \frac{q\gamma_u}{c}\left(u_x E_x + u_y E_y + u_z E_z\right) = \frac{q\gamma_u}{c}\,\mathbf{u}\cdot\mathbf{E}.$$

By (25.38), K^0 is γ_u/c times the rate of change of the charge's energy, $\gamma_u mc^2$, so

$$\frac{d}{dt}\left(\gamma_u mc^2\right) = q\,\mathbf{u}\cdot\mathbf{E},$$

the rate at which the fields do work on the charge. Only the electric field does work; the magnetic field does none (Section 10.1.3). Tutorial 26.1 asks what the other three components of (26.17) say.

26.6 The four-potential

The potentials of Chapter 24 also form a four-vector, the **four-potential**

$$A^\mu = \left(\frac{V}{c},\ A_x,\ A_y,\ A_z\right), \tag{26.18}$$

at least when they are the retarded potentials of the Lorenz gauge, as Section 26.6.1 proves. The field tensor is its four-dimensional curl,

$$F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu. \tag{26.19}$$

To check one component, take $\mu = 0$ and $\nu = 1$, with $\partial^0 = -(1/c)\,\partial/\partial t$ and $\partial^1 = \partial/\partial x$ from (26.13):

$$F^{01} = \partial^0 A^1 - \partial^1 A^0 = -\frac{1}{c}\frac{\partial A_x}{\partial t} - \frac{1}{c}\frac{\partial V}{\partial x} = \frac{1}{c}\left(-\frac{\partial V}{\partial x} - \frac{\partial A_x}{\partial t}\right) = \frac{E_x}{c},$$

[3]The paper is "Die Grundgleichungen für die elektromagnetischen Vorgänge in bewegten Körpern," H. Minkowski, Nachr. Ges. Wiss. Göttingen, Math.-Phys. Kl. 1908, 53–111 (1908). https://eudml.org/doc/58707.

by (15.11). The rest of the time row gives the other components of $\mathbf{E} = -\nabla V - \partial\mathbf{A}/\partial t$, and the space components of (26.19) give $\mathbf{B} = \nabla \times \mathbf{A}$ (Exercise P26.15). So (26.19) is the side of the triangle diagram of Figure 24.10 that goes from the potentials to the fields, written as one tensor equation. A gauge transformation (15.12) becomes $A^\mu \to A^\mu + \partial^\mu\psi$, and it leaves $F^{\mu\nu}$ unchanged, since $\partial^\mu\partial^\nu\psi = \partial^\nu\partial^\mu\psi$.

In this notation the Lorenz condition (24.3) and the two wave equations (24.4) and (24.5) become

$$\partial_\mu A^\mu = 0, \qquad \Box^2 A^\mu = -\mu_0 J^\mu, \tag{26.20}$$

as Tutorial 26.1 asks you to check. Both are equations between scalars or between four-vectors, so a potential that satisfies the Lorenz condition in one inertial frame satisfies it in every inertial frame. That is the advantage of the Lorenz gauge promised in Section 24.2. The Coulomb condition $\nabla\cdot\mathbf{A} = 0$ involves only the space part of A^μ, and potentials that satisfy it in one frame in general do not satisfy it in another (Exercise P26.14).

26.6.1 Proof that the retarded potentials form a four-vector

Potentials are not unique: a gauge transformation changes them without changing the fields. So the claim that $(V/c, \mathbf{A})$ is a four-vector needs to be stated with care. What is true, and what Chapter 27 needs, is this: the retarded potentials (24.11) and (24.12), calculated in two inertial frames for the same sources, are related by the Lorentz transformation, $A^\mu_{\mathrm{R}} = \Lambda^\mu{}_\nu A^\nu_{\mathrm{L}}$ at every event. The proof needs five facts:

1. J^μ is a four-vector (Section 26.4).
2. The interval $(x - x')_\mu (x - x')^\mu$ between two events x and x' is invariant (Section 25.5.1).
3. The four-volume element d^4x is invariant, (25.20).
4. Two events on each other's light cone have the same order in time in every frame (Section 25.5.2).
5. If a function $g(t)$ is zero at the points t_i, and its derivative $g'(t_i)$ is not zero there, then

$$\delta\big(g(t)\big) = \sum_i \frac{\delta(t - t_i)}{|g'(t_i)|}. \tag{26.21}$$

The last is new. Away from the points t_i, $g(t)$ is not zero, and the delta function is zero. Near each t_i, $g(t) \approx g'(t_i)(t - t_i)$, and $\delta\big(g'(t_i)(t - t_i)\big) = \delta(t - t_i)/|g'(t_i)|$ (substituting $u = g'(t_i)(t - t_i)$ in the integral over t gives the factor $1/|g'(t_i)|$, with the absolute value because the limits of integration swap when $g'(t_i)$ is negative). Exercise P26.13 gives some practice with it. Step 2 of the proof meets a function with two such points (Figure 26.8).

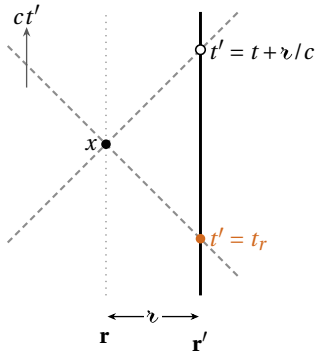


Figure 26.8 Step 2 of the proof: the field event x, its light cone (dashed), and the world line of a source point a distance r away. The interval between them is zero where the world line crosses the light cone, at $t' = t_r$ and at $t' = t + r/c$; the step function keeps only the first.

Proof that the retarded potentials form a four-vector

Step 1. Write the two retarded potentials as one equation. Since $1/\varepsilon_0 c = \mu_0 c$, the potential V/c of (24.11) is $(\mu_0/4\pi)\int c\rho(\mathbf{r}', t_r)/\mathscr{r}\, d\tau'$, so

$$A^\mu(\mathbf{r}, t) = \frac{\mu_0}{4\pi}\int \frac{J^\mu(\mathbf{r}', t_r)}{\mathscr{r}}\, d\tau', \qquad t_r = t - \frac{\mathscr{r}}{c},$$

where $d\tau'$ is the volume element, as in Chapter 24, not a proper time.

Step 2. Replace the retarded time with a delta function. For fixed $\mathbf{r}$ and $\mathbf{r}'$, the interval between the event $x = (ct, \mathbf{r})$ and the event $x' = (ct', \mathbf{r}')$ at the source point,

$$g(t') = (x - x')_\mu (x - x')^\mu = -c^2(t - t')^2 + \mathscr{r}^2,$$

is zero at $t' = t - \mathscr{r}/c$ and at $t' = t + \mathscr{r}/c$, the times at which the source point crosses the past and the future light cones of x, and at both of them $|dg/dt'| = 2c^2|t - t'| = 2c\mathscr{r}$. The step function $\Theta(t - t')$ of Section 1.6.1 keeps only the earlier of the two, the retarded time, so by (26.21)

$$\Theta(t - t')\,\delta\big(g(t')\big) = \frac{\delta(t' - t_r)}{2c\mathscr{r}}.$$

In Step 1, write $J^\mu(\mathbf{r}', t_r)$ as an integral over all times, $\int J^\mu(\mathbf{r}', t')\,\delta(t' - t_r)\,dt'$; then use this result, and write $c\,dt'\,d\tau' = d^4x'$:

$$A^\mu(x) = \frac{\mu_0}{2\pi}\int J^\mu(x')\,\Theta(t - t')\,\delta\big((x - x')_\nu (x - x')^\nu\big)\, d^4x'. \tag{26.22}$$

Step 3. Every factor in the integral but J^μ is a Lorentz scalar: d^4x' by fact 3; the delta function, whose argument is an interval, by fact 2; and the step function by fact 4, because the delta function admits only events x' on the light cone of x, and for those the order in time is the same in every frame. (For events with spacelike separation it is not; the delta function excludes them.) In Rachel's frame, then, with the variables of integration changed from x'_{R} to x'_{L}, where $x'_{\mathrm{R}} = \Lambda x'_{\mathrm{L}}$ and the Jacobian is $\det\Lambda = 1$,

$$A^\mu_{\mathrm{R}}(x_{\mathrm{R}}) = \frac{\mu_0}{2\pi}\int \Lambda^\mu{}_\nu J^\nu_{\mathrm{L}}(x'_{\mathrm{L}})\,\Theta\delta\; d^4x'_{\mathrm{L}} = \Lambda^\mu{}_\nu A^\nu_{\mathrm{L}}(x_{\mathrm{L}}).$$

The retarded potentials transform as a four-vector.

Three consequences follow. First, the retarded potentials satisfy the Lorenz condition (Exercise P24.8), and since $\partial_\mu A^\mu$ is a scalar, they satisfy it in every frame. Second, (26.19) builds $F^{\mu\nu}$ from two four-vectors, ∂^μ and A^ν, by the steps that make a four-tensor, so $F^{\mu\nu}$ is a four-tensor for the field of any distribution of charges and currents, not only for capacitor plates and solenoids. That proves the transformation rules (26.5) in general and closes the gap left by the plausibility argument of Section 26.2. Third, (26.22) is the starting point of Chapter 27: with the four-current (26.12) of a point charge, it gives the potentials of a charge in any motion in a few lines.

The potentials of a charge in uniform motion need no integral at all, as the next example shows.

Example 26.2

A point charge q is at rest at the origin of Rachel's frame, and Rachel moves along x at speed v relative to Luke, with their origins together at $t_L = t_R = 0$. Find the potentials V_L and $\mathbf{A}_L$ in Luke's frame at the event (t_L, x_L, y_L, z_L).

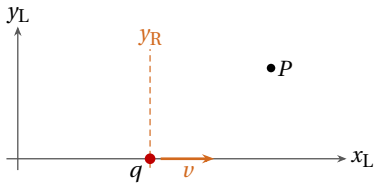
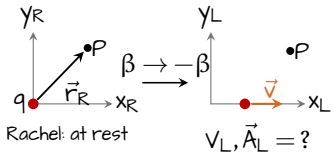


Figure 26.9 Example 26.2: a charge q at rest in Rachel's frame, which moves along x at v relative to Luke, and an event at the point P.

Solution:



The first step is a sketch of the charge in its own frame, where its potentials are the static ones, and in Luke's frame, where it moves; the four-potential carries us from one to the other.

In Rachel's frame the charge is at rest, so its retarded potentials are the potentials of electrostatics, $V_R = q/4\pi\varepsilon_0 r_R$, with $r_R = \sqrt{x_R^2 + y_R^2 + z_R^2}$, and $\mathbf{A}_R = 0$. Its four-potential is $A_R^\mu = (V_R/c, 0, 0, 0)$. The inverse transformation, with $\beta \to -\beta$, gives Luke's:

$$A_L^0 = \gamma\left(A_R^0 + \beta A_R^1\right) = \gamma\,\frac{V_R}{c}, \qquad A_L^1 = \gamma\left(A_R^1 + \beta A_R^0\right) = \gamma\beta\,\frac{V_R}{c},$$

and $A_L^2 = A_L^3 = 0$. So $V_L = \gamma V_R$ and $\mathbf{A}_L = (\gamma v V_R/c^2)\hat{\mathbf{x}} = (v/c^2)V_L\hat{\mathbf{x}}$. Finally, Rachel's coordinates in terms of Luke's are $x_R = \gamma(x_L - vt_L)$, $y_R = y_L$, and $z_R = z_L$, (25.8), so

$$V_L = \frac{q}{4\pi\varepsilon_0}\,\frac{\gamma}{\sqrt{\gamma^2(x_L - vt_L)^2 + y_L^2 + z_L^2}}, \qquad \mathbf{A}_L = \frac{v}{c^2}\,V_L\,\hat{\mathbf{x}}.$$

For $v \to 0$, $\gamma \to 1$ and these are the potentials of a charge at rest. They satisfy the Lorenz condition: V_L depends on x_L and t_L only through $x_L - vt_L$, so $\partial V_L/\partial t_L = -v\,\partial V_L/\partial x_L$, and

$$\frac{1}{c^2}\frac{\partial V_L}{\partial t_L} + \nabla\cdot\mathbf{A}_L = -\frac{v}{c^2}\frac{\partial V_L}{\partial x_L} + \frac{v}{c^2}\frac{\partial V_L}{\partial x_L} = 0.$$

At time t_L the charge is at $x_L = vt_L$. At a point a distance R from it, beside it ($x_L = vt_L$), $V_L = \gamma q/4\pi\varepsilon_0 R$, larger than the potential of a charge at rest by the factor γ; at a point a distance R ahead of it or behind it, $V_L = q/4\pi\varepsilon_0 R$, the same. The potential is centered on the charge's present position, even though the news of where the charge is travels at the speed of light. Section 26.7 finds the fields, and Section 27.1 finds the same potentials from the retarded potentials directly.

Tutorial 26.1 Tensor equations

Printable version

Background

Sections 26.3–26.6 wrote several of the most important equations of electromagnetism in tensor form. Each tensor equation stands for several ordinary equations, one for each value of its free index, and each of its terms hides a sum over every repeated index. In this tutorial you unpack five of them, as Section 26.4 did for $\partial_\mu J^\mu = 0$. The tools are the summation convention (Section 25.7.2); the four-gradient ∂_μ and ∂^μ, (26.13); the field tensor $F^{\mu\nu}$, (26.7), and the dual tensor $G^{\mu\nu}$, (26.9); the four-current $J^\mu = (c\rho, \mathbf{J})$, the four-potential $A^\mu = (V/c, \mathbf{A})$, and the four-velocity $\eta^\mu = \gamma_u(c, \mathbf{u})$. Remember that lowering an index changes the sign of each component that has a 0 in that index.

Part 1 — Five tensor equations

For each of the five equations in the steps below:

1. Identify the quantities in the equation, and say in words what it means. (For $\partial_\mu J^\mu = 0$ one would say, "The four-divergence of the four-current is zero.")
2. Determine the units of the terms on both sides of the equation, and make sure that they match.
3. Use the definitions of the quantities and the summation convention to write the equation out in full.
4. Determine which equation or equations of earlier chapters it corresponds to, and give their equation numbers.

(a) $\partial_\mu A^\mu = 0$. Here $\partial_\mu A^\mu$ is a Lorentz scalar, which is zero in every frame for the potentials used in this book's chapters on radiation.

(b) $\partial_\nu F^{\mu\nu} = \mu_0 J^\mu$. Both sides are four-vectors, so this is four equations, one for each value of μ. Write the four as two: first the equation for $\mu = 0$, then the three for $\mu = 1, 2, 3$ as one vector equation.

(c) $\partial_\nu G^{\mu\nu} = 0$. Again consider $\mu = 0$ first, then $\mu = 1, 2, 3$.

(d) $K^\mu = q\eta_\nu F^{\mu\nu}$. Section 26.5 worked out the $\mu = 0$ component. Which equation do the components $\mu = 1, 2, 3$ correspond to? (Use $K^{1,2,3} = \gamma_u \mathbf{F}$, from (25.38).)

(e) $\partial^\nu \partial_\nu A^\mu = -\mu_0 J^\mu$. By (26.14), $\partial^\nu \partial_\nu$ is the d'Alembertian $\Box^2$, a scalar operator, so acting on a four-vector it gives a four-vector. Consider $\mu = 0$ first, then $\mu = 1, 2, 3$.

Part 2 — Wrap-up

(a) Why would physicists choose to write the equations of electromagnetism in tensor notation like this? What are some advantages and disadvantages?

(b) The two equations of (26.16) contain all four of Maxwell's equations. Is the tensor form simpler than the four equations? In what sense, and what does it hide?

26.7 The fields of a uniformly moving charge

The fields of a moving charge are hard to find from the retarded potentials of Chapter 24: even locating the shadow of a charge that moves at constant velocity took an example of its own (Example 24.1). For a charge in uniform motion, relativity gives the fields with very little work. In the charge's own frame its field is Coulomb's, and the rules of Section 26.2.3 carry that field to any other frame. No retarded time is needed.

Let a point charge q be at rest at the origin of Rachel's frame, as in Example 26.2, so that in Luke's frame it moves along x at speed v, passing his origin at $t_L = 0$. In Rachel's frame the fields are

$$\mathbf{E}_R = \frac{q}{4\pi\varepsilon_0}\frac{x_R\hat{\mathbf{x}} + y_R\hat{\mathbf{y}} + z_R\hat{\mathbf{z}}}{\left(x_R^2 + y_R^2 + z_R^2\right)^{3/2}}, \qquad \mathbf{B}_R = 0.$$

Luke moves at $-v$ relative to Rachel, so the rules (26.5) with $v \to -v$ give his fields at the same event:

$$E_{L,x} = E_{R,x}, \qquad E_{L,y} = \gamma E_{R,y}, \qquad E_{L,z} = \gamma E_{R,z},$$

and, since $\mathbf{B}_R = 0$, $\mathbf{B}_L = \mathbf{v} \times \mathbf{E}_L/c^2$, the special case of Section 26.2.3, with $\mathbf{v} = v\hat{\mathbf{x}}$ the charge's velocity in Luke's frame. To write $\mathbf{E}_L$ in Luke's coordinates, use (25.8): $x_R = \gamma(x_L - vt_L)$, $y_R = y_L$, and $z_R = z_L$. At time t_L the charge is at $x_L = vt_L$, so the vector from the charge's present position to the field point is

$$\mathbf{R} = (x_L - vt_L)\hat{\mathbf{x}} + y_L\hat{\mathbf{y}} + z_L\hat{\mathbf{z}},$$

and Rachel's coordinates are $x_R = \gamma R_x$, $y_R = R_y$, and $z_R = R_z$. (The bold $\mathbf{R}$ is a vector; the upright subscript R still labels Rachel's frame.) Then

$$\mathbf{E}_L = \frac{q}{4\pi\varepsilon_0}\frac{\gamma R_x\hat{\mathbf{x}} + \gamma R_y\hat{\mathbf{y}} + \gamma R_z\hat{\mathbf{z}}}{\left(\gamma^2 R_x^2 + R_y^2 + R_z^2\right)^{3/2}} = \frac{q}{4\pi\varepsilon_0}\frac{\gamma\mathbf{R}}{\left(\gamma^2 R_x^2 + R_y^2 + R_z^2\right)^{3/2}}.$$

The factor γ that the transformation gives the components of the field across the motion is the same factor by which the contraction shortens distances along it, so the field points along $\mathbf{R}$. Let θ be the angle between $\mathbf{v}$ and $\mathbf{R}$, so that $R_x = R\cos\theta$ and $R_y^2 + R_z^2 = R^2\sin^2\theta$. With $\beta = v/c$, $\gamma^2\cos^2\theta + \sin^2\theta = \gamma^2(1 - \beta^2\sin^2\theta)$, and the fields in Luke's frame, with the label L dropped, are

$$\boxed{\mathbf{E} = \frac{q}{4\pi\varepsilon_0}\frac{1-\beta^2}{\left(1-\beta^2\sin^2\theta\right)^{3/2}}\frac{\hat{\mathbf{R}}}{R^2}, \qquad \mathbf{B} = \frac{\mathbf{v}\times\mathbf{E}}{c^2}}. \tag{26.23}$$

These are the fields of a point charge that moves with constant velocity $\mathbf{v}$. The vector $\mathbf{R}$ runs from the charge's present position to the field point, and θ is the angle between $\mathbf{v}$ and $\mathbf{R}$.

Figure 26.10 shows the electric field lines at four speeds. The field is still radial, and it still falls off as $1/R^2$, but it is no longer the same in all directions.

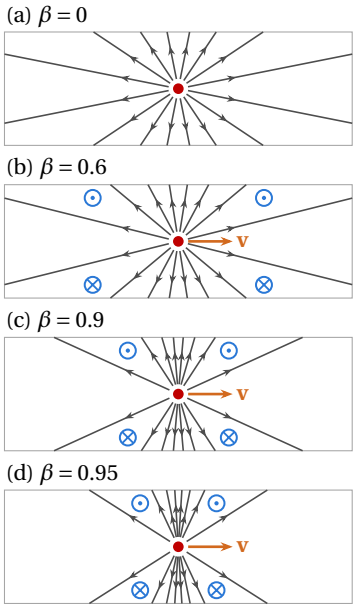


Figure 26.10 Electric field lines (black) of a positive charge moving to the right at (a) $\beta = 0$, (b) $\beta = 0.6$, (c) $\beta = 0.9$, and (d) $\beta = 0.95$, drawn at equal angles in the charge's own frame, and its magnetic field (blue), out of the page above the line of motion and into it below. Animation.

Beside the charge ($\theta = 90°$) it is stronger than Coulomb's field by the factor $(1 - \beta^2)/(1 - \beta^2)^{3/2} = \gamma$, and ahead of the charge and behind it ($\theta = 0$ and $180°$) it is weaker, by the factor $1 - \beta^2 = 1/\gamma^2$. At $\beta = 0.995$, where $\gamma \approx 10$, the field beside the charge is about ten times Coulomb's, and the field ahead and behind is about a hundredth of it. The field is flattened into a pancake around the plane through the charge perpendicular to its motion, and the faster the charge moves, the flatter the pancake. Oliver Heaviside, whose portrait is in Section 2.4, found this field from Maxwell's equations in 1889.[4] Gauss's law still holds: the flux of **E** through a closed surface around the charge is q/ε_0 (Exercise P26.16). The magnetic field circles the line of motion, as the field of a current in a wire does, and like **E** it is concentrated near the plane perpendicular to the motion.

The most surprising thing about (26.23) is where **R** starts. The news from the charge travels at the speed of light, and an observer at the field point sees the charge at its shadow, the position it had at the retarded time (Section 24.3). Yet the field points straight away from the position that the charge has *now*, which the observer cannot yet see. No news has traveled faster than light. The news that reaches the field point left the charge at the retarded time, and it carried the charge's velocity as well as its position. For a charge that has kept that velocity ever since, the field points away from where the charge would be had it kept moving that way, which is where it is. Had the charge changed its velocity after the retarded time, the field would still point away from the position it would have reached without the change, until the news of the change arrived. Section 27.2.1 shows how that news travels, and it finds that part of the field of a charge in any motion is (26.23), with **R** measured from this projected position.

For $v \ll c$, $\gamma \approx 1$, and (26.23) becomes Coulomb's field, measured from the charge's present position, together with the magnetic field

$$\mathbf{B} \approx \frac{\mu_0}{4\pi} \frac{q\mathbf{v} \times \hat{\mathbf{R}}}{R^2},$$

since $1/\varepsilon_0 c^2 = \mu_0$. This is the formula that Section 10.5 called an approximation for a single moving charge, good only at low speeds. For a charge moving at constant velocity, (26.23) is the exact result, with **R** measured from the present position, and the corrections to the low-speed formula are of order β^2 (Exercise P26.22). Tutorial 26.2 uses these fields to find the electromagnetic power that flows through a plane in front of a moving charge.

[4] "On the electromagnetic effects due to the motion of electrification through a dielectric," O. Heaviside, Phil. Mag. 27, 324–339 (1889). https://doi.org/10.1080/14786448908628362.

Tutorial 26.2 Power through a plane

Printable version

Background

This tutorial uses the fields of a point charge moving at constant velocity, (26.23). In them **R** is the vector that points from the *present* location of the charge to the point where you want to know the fields, **v** is the charge's velocity, and θ is the angle between **v** and **R**. Parts 1 and 2 also use the Poynting vector $\mathbf{S} = \mathbf{E} \times \mathbf{B}/\mu_0$ (Section 20.2).

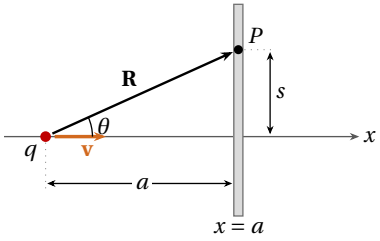


Figure 26.11 Tutorial 26.2: a point charge q, now at the origin, moving at constant velocity **v** toward the plane $x = a$, seen edge on, and an arbitrary point in the plane, a distance s from the x axis.

Part 1 — General

Figure 26.11 shows a point charge q traveling at constant velocity **v** in the x direction, currently at the origin and moving toward a plane at $x = a$.

(a) What will the electric and magnetic fields surrounding the point charge look like in general, and in what directions will they point at the arbitrary point marked in the figure? Draw the fields on a copy of the figure.

(b) In what direction will the Poynting vector point at various locations on the plane? Think about the arbitrary point first, and then about the rest of the plane. Sketch it.

(c) What significance does your answer to (b) have? In other words, what does the Poynting vector represent physically, and what does it mean for the Poynting vector field to look that way?

(d) Repeat (b) and (c) for a plane located behind the charge, for example at $x = -a$.

Part 2 — Finding the power passing through the plane

(a) The goal of the rest of this tutorial is to find the electromagnetic power passing through the infinite plane at $x = a$. Write down your knowns and unknowns, and describe a "game plan" for how you could use what you have been given to find it. What are the steps along the way from what you know to what you want to know? Avoid looking at the following steps when you make your plan; it is fine if your plan is not the way the tutorial does things.

(b) One piece of your game plan likely involves the electric field on the plane. Find it now at the arbitrary point, in terms of v, s, a, and $\hat{\mathbf{R}}$.

(c) Another piece may involve the magnetic field on the plane. We can bypass that step, however, if we are clever. Consider that $\mathbf{E} \times \mathbf{B} = (1/c^2)\,\mathbf{E} \times (\mathbf{v} \times \mathbf{E})$. What about this suggests that there might be a simpler way to find **S** than finding **B**? Why is it important to consider things like this when choosing the steps toward solving a problem?

(d) Derive an expression for the Poynting vector in terms of the electric field alone (no **B**), the velocity **v**, and the angle θ. (Hint: the "BAC-CAB" vector identity should help.)

(e) The electromagnetic power passing through the plane is $\int_{\text{plane}} \mathbf{S} \cdot d\mathbf{a}$. Why? Which component of **S** matters here? Find that component, in terms of the magnitudes of **E** and **v** and the angle θ.



Christian Doppler (1803–1853, Austrian) Doppler was born in Salzburg into a family of stonemasons. He studied at the Polytechnic Institute in Vienna and then higher mathematics, mechanics, and astronomy at the University of Vienna, and from 1835 he taught in Prague, at a technical school and then at the Polytechnic. On 25 May 1842 he presented to the Royal Bohemian Society of Sciences a paper on the colored light of double stars, in which he argued that the frequency of a wave depends on the motion of its source and of its observer. The colors of the stars do not come from their motion, which is far too slow, but the principle was right: in 1845 Christophorus Buys Ballot confirmed it for sound, with musicians who played on a moving train while others noted the pitch they heard. In 1850 Doppler became director of the Institute of Physics at the University of Vienna. He died in Venice three years later, at 49. (Wikipedia)

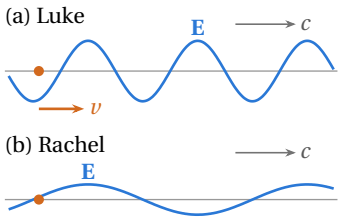


Figure 26.12 A plane wave traveling in the $+x$ direction (a) in Luke's frame, with Rachel moving away from its source at $\beta = 0.6$, and (b) in Rachel's frame, where its wavelength is twice as long and its amplitude half as large.

(f) Complete the problem: do the integral $\int_{\text{plane}} \mathbf{S}\cdot d\mathbf{a}$, and find how much power passes through the plane.

Part 3 — Wrap-up

(a) Think of the standard ways that physicists determine whether an answer is reasonable: limiting cases, dimensional analysis, comparison with other problems, and so forth. Apply at least one of them to your answer.

(b) What are the implications of your answer to Part 2(f)?

26.8 The Doppler effect for light

When a source of sound moves toward you, you hear a higher pitch than it emits, and when it moves away, a lower one; the same happens when you move toward a source at rest or away from it. This is the **Doppler effect**. For sound, the size of the shift depends on whether the source or the listener moves through the air. Light needs no medium, and by Postulate 1 only the relative motion of the source and the observer can matter. The field transformation gives the shift directly.

Let Luke measure a plane wave of angular frequency ω that travels in the $+x$ direction, with its electric field along y (Figure 26.12a):

$$\mathbf{E}_{\text{L}} = E_0 \cos(kx_{\text{L}} - \omega t_{\text{L}})\,\hat{\mathbf{y}}, \qquad \mathbf{B}_{\text{L}} = \frac{E_0}{c}\cos(kx_{\text{L}} - \omega t_{\text{L}})\,\hat{\mathbf{z}},$$

with $\omega = ck$ and $B = E/c$, as for any plane wave in empty space (Section 16.5). Rachel moves in the $+x$ direction at v, away from the source. By (26.5),

$$E_{\text{R},y} = \gamma\left(E_{\text{L},y} - vB_{\text{L},z}\right) = \gamma(1-\beta)\,E_0\cos(kx_{\text{L}} - \omega t_{\text{L}}),$$
$$B_{\text{R},z} = \gamma\left(B_{\text{L},z} - \frac{v}{c^2}E_{\text{L},y}\right) = \gamma(1-\beta)\,\frac{E_0}{c}\cos(kx_{\text{L}} - \omega t_{\text{L}}),$$

and the other components are zero. In Rachel's coordinates, by the inverse transformation (25.10), $x_{\text{L}} = \gamma(x_{\text{R}} + vt_{\text{R}})$ and $t_{\text{L}} = \gamma(t_{\text{R}} + vx_{\text{R}}/c^2)$, so the phase of the wave is

$$kx_{\text{L}} - \omega t_{\text{L}} = \gamma k(x_{\text{R}} + vt_{\text{R}}) - \gamma\omega\left(t_{\text{R}} + \frac{vx_{\text{R}}}{c^2}\right) = \gamma(1-\beta)(kx_{\text{R}} - \omega t_{\text{R}}),$$

using $\omega = ck$. Rachel therefore measures

$$\mathbf{E}_{\text{R}} = \gamma(1-\beta)\,E_0\cos\left[\gamma(1-\beta)(kx_{\text{R}} - \omega t_{\text{R}})\right]\hat{\mathbf{y}},$$

a plane wave whose amplitude, wave number, and angular frequency are all smaller than Luke's by the factor

$$\gamma(1-\beta) = \frac{1-\beta}{\sqrt{(1-\beta)(1+\beta)}} = \sqrt{\frac{1-\beta}{1+\beta}}$$

(Figure 26.12b). Its speed ω_R/k_R is still c, as Postulate 2 requires, and its angular frequency is

$$\boxed{\omega_R = \sqrt{\frac{1-\beta}{1+\beta}}\,\omega_L}. \tag{26.24}$$

For an observer moving away from the source, $\beta > 0$ and the frequency is lower; for one moving toward it, $\beta < 0$ and the frequency is higher. Only the relative velocity of the source and the observer appears. For $|\beta| \ll 1$ the factor is $1-\beta$, the first-order shift familiar from sound; the formula for light differs from the formulas for sound at order β^2. The factors 1/3 and 3 by which Luke saw Rachel's clock run slow and fast in Example 25.3 are (26.24) with $\beta = 0.8$ and $\beta = -0.8$. Einstein derived the relativistic formula, for light traveling in any direction, in Section 7 of his 1905 paper (the reference is in the footnote in Section 15.2.1).

The calculation also shows something more general. At every event the phase of the wave has the same value for both observers: $kx_L - \omega t_L = k_R x_R - \omega_R t_R$. That must be so for any wave. The field transformation multiplies the fields at an event by constants without shifting their phase, so the events at which the field is at a crest, or is zero, are the same for every observer. For a wave traveling in any direction the phase is $\phi = \mathbf{k}\cdot\mathbf{r} - \omega t$, and since it is a scalar, its four-gradient (26.13), $\partial_\mu\phi = (-\omega/c,\, k_x,\, k_y,\, k_z)$, is a four-vector with its index down. The upper-index form is the **four-wavevector**

$$k^\mu = \left(\frac{\omega}{c},\, k_x,\, k_y,\, k_z\right), \tag{26.25}$$

and the phase is $\phi = k_\mu x^\mu$. Since $\omega = ck$, $k_\mu k^\mu = 0$: the four-wavevector of light is lightlike. Table 26.1 lists it with the other four-vectors of the chapter.

The time component of $k_R^\mu = \Lambda^\mu{}_\nu k_L^\nu$ gives the Doppler shift of a wave traveling in any direction: $\omega_R/c = \gamma(\omega_L/c - \beta k_{L,x})$, or

$$\omega_R = \gamma\,\omega_L\left(1 - \beta\cos\theta_L\right), \tag{26.26}$$

where θ_L is the angle between the direction of the wave and Rachel's velocity, measured in Luke's frame. For $\theta_L = 0$ this is (26.24). For $\theta_L = 90°$, a wave that crosses Rachel's path at right angles in Luke's frame, $\omega_R = \gamma\,\omega_L$: Rachel measures a higher frequency, although she moves neither toward the source nor away from it. This **transverse Doppler effect** is time dilation. Rachel moves along the crests of the wave, which pass her at Luke's rate, and her clock runs slow by the factor γ, so she counts γ times as many crests per second of her own time. Herbert Ives and G. R. Stilwell detected this part of the Doppler shift in 1938, in the light emitted by a fast beam of hydrogen.[5] The space components of k^μ give the direction of the wave in Rachel's frame, which turns out to be the aberration (25.14) of Section 25.6 (Exercise P26.24).

[5] "An experimental study of the rate of a moving atomic clock," H. E. Ives and G. R. Stilwell, J. Opt. Soc. Am. 28, 215–226 (1938). https://doi.org/10.1364/JOSA.28.000215.

26.9 The stress–energy tensor

The fields carry energy and momentum (Sections 20.2 and 20.3). In empty space the energy per unit volume is $u = \frac{1}{2}\varepsilon_0 E^2 + B^2/2\mu_0$; energy flows at the rate given by the Poynting vector $\mathbf{S} = \mathbf{E}\times\mathbf{B}/\mu_0$, the energy per unit time per unit area; the momentum per unit volume is $\mathbf{S}/c^2$; and the momentum flux density is the negative of the Maxwell stress tensor, $-T_{ij}$ (Section 20.3.4). For a particle, energy and momentum together form a four-vector, p^μ. For the fields, the densities and the flows of energy and momentum together form a four-tensor, the **stress–energy tensor**

$$T^{\mu\nu} = \begin{pmatrix} u & S_x/c & S_y/c & S_z/c \\ S_x/c & -T_{xx} & -T_{xy} & -T_{xz} \\ S_y/c & -T_{yx} & -T_{yy} & -T_{yz} \\ S_z/c & -T_{zx} & -T_{zy} & -T_{zz} \end{pmatrix}. \tag{26.27}$$

The space part is the momentum flux density, the negative of the Maxwell stress tensor of Section 20.3.3. The first row of $T^{\mu\nu}$ holds the energy density and the energy flow divided by c; each of the other rows holds c times the density of one component of momentum, followed by the flow of that component. The tensor is symmetric, $T^{\nu\mu} = T^{\mu\nu}$, and the symmetry has a physical meaning: the momentum density of the fields is their energy flow divided by c^2, just as the momentum of a particle is its energy times its velocity divided by c^2, (25.32). Built from the field tensor, $T^{\mu\nu}$ is a four-tensor.[6]

☞ *Caution*: In relativity, the word *stress* usually means the space part T^{ij} itself, the momentum flux density, whose diagonal elements are positive where the stress pushes inward on a surface, as a pressure does. The Maxwell stress tensor of Section 20.3.3 has positive diagonal elements where the stress pulls outward, as the stress of solid mechanics and engineering does, so the two differ by a sign.

The conservation of energy and momentum, for the fields and the charges they act on, takes one tensor form,

$$\partial_\nu T^{\mu\nu} = -F^{\mu\nu}J_\nu. \tag{26.28}$$

For $\mu = 0$, the left side is

$$\partial_\nu T^{0\nu} = \frac{1}{c}\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\frac{S_x}{c} + \frac{\partial}{\partial y}\frac{S_y}{c} + \frac{\partial}{\partial z}\frac{S_z}{c} = \frac{1}{c}\left(\frac{\partial u}{\partial t} + \nabla\cdot\mathbf{S}\right),$$

and the right side, with $J_\nu = (-c\rho, J_x, J_y, J_z)$ and $F^{00} = 0$, is $-(F^{01}J_1 + F^{02}J_2 + F^{03}J_3) = -\mathbf{J}\cdot\mathbf{E}/c$. So

$$\frac{\partial u}{\partial t} + \nabla\cdot\mathbf{S} = -\mathbf{J}\cdot\mathbf{E},$$

which is Poynting's theorem (Section 20.2.2): the energy of the fields in a small volume changes because energy flows out through its surface and because the fields do work on the charges in it, at the rate $\mathbf{J}\cdot\mathbf{E}$ per unit volume. In the same way, the components $\mu = 1, 2, 3$ are the conservation of momentum, with the force per unit volume on the charges, $\rho\mathbf{E} + \mathbf{J}\times\mathbf{B}$, on the right side, with a minus sign (Exercise P26.26). The four-vector $F^{\mu\nu}J_\nu$ has as its time component the power

[6] In terms of $F^{\mu\nu}$, $T^{\mu\nu} = (1/\mu_0)\left(F^{\mu\alpha}F^{\nu}{}_{\alpha} - \frac{1}{4}g^{\mu\nu}F_{\alpha\beta}F^{\alpha\beta}\right)$, where $F^{\nu}{}_{\alpha}$ has its second index lowered and $g^{\mu\nu} = \mathrm{diag}(-1,1,1,1)$ is the metric of the footnote in Section 25.7.1. Every term is a product of four-tensors with indices summed in pairs, so $T^{\mu\nu}$ transforms as a four-tensor.

per unit volume delivered to the charges, divided by c, and as its space part the force per unit volume on them, so (26.28) says that the fields lose the energy and momentum that the charges gain. Where there are no charges, the right side is zero, and the energy and momentum of the fields alone are conserved.

26.10 Numerical method: motion in crossed fields

A charge released from rest in uniform, perpendicular electric and magnetic fields, called **crossed fields**, moves on a cycloid and drifts with the average velocity $\mathbf{E}\times\mathbf{B}/B^2$ (Section 10.1.4). That result is Newtonian. Its drift speed E/B can have any size, and the speed of the charge at the top of each arch of the cycloid is twice the drift speed, which exceeds c when $E > cB/2$. This section follows the motion with the relativistic Boris method of Section 25.13 and checks the result with the field transformation.

The transformation shows what to expect. For crossed fields $\mathbf{E}\cdot\mathbf{B} = 0$, and by (26.10) $\mathbf{E}\cdot\mathbf{B} = 0$ in every frame. Suppose that $E < cB$. Then the **drift velocity**

$$\mathbf{v}_d = \frac{\mathbf{E}\times\mathbf{B}}{B^2}, \tag{26.29}$$

of magnitude $E/B < c$, is the velocity of an inertial frame. In that frame the electric field is $\gamma_d(\mathbf{E}+\mathbf{v}_d\times\mathbf{B})_\perp$, by (26.6), with $\gamma_d = 1/\sqrt{1-v_d^2/c^2}$, and since $(\mathbf{E}\times\mathbf{B})\times\mathbf{B} = -B^2\mathbf{E}$ when $\mathbf{E}$ is perpendicular to $\mathbf{B}$, $\mathbf{v}_d\times\mathbf{B} = -\mathbf{E}$: the electric field vanishes. In the frame that moves with the drift velocity there is only a magnetic field, and the invariant $E^2 - c^2B^2$ gives its size: $-c^2B_d^2 = E^2 - c^2B^2$, so

$$B_d = \sqrt{B^2 - \frac{E^2}{c^2}} = \frac{B}{\gamma_d}.$$

In that frame the charge moves on a circle at constant speed (Section 25.12), and in the lab the circle, carried along at $\mathbf{v}_d$, becomes a path of repeated arches, on which the charge never reaches the speed of light. If instead $E > cB$, the drift speed E/B would exceed c, and no frame moves with it; Exercise P26.28 finds what happens then.

The relativistic Boris method needs no change for crossed fields: each step adds half of the electric impulse to $\mathbf{w}$, rotates $\mathbf{w}$ about $\mathbf{B}$, and adds the other half. Example 26.3 follows an electron along one arch of its path and compares the result with what the drift frame predicts.

Tips for Mathematica and Python. The method advances $\mathbf{w}$ at the times halfway between those at which it advances $\mathbf{r}$, so a charge that starts from rest should start with $\mathbf{w} = -(q/2m)\,\mathbf{E}\,\Delta t$, its value half a step before the start; otherwise the times it computes are late by about half a step. To compare a computed path with the drift frame, transform each computed event (t, x, y, z) with (25.8), not each position at one time of the lab: with $\mathbf{v}_d$ along x, $x_\mathrm{R} = \gamma_d(x - v_d t)$ and $t_\mathrm{R} = \gamma_d(t - v_d x/c^2)$, and y and z are unchanged. Events evenly spaced in the lab's time are unevenly spaced in the drift frame's, but they still lie on the circle.

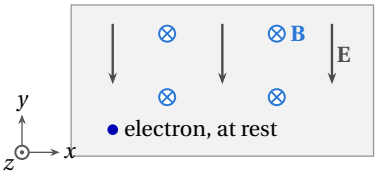
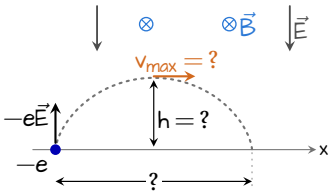


Figure 26.13 Example 26.3: an electron at rest at the origin, in a uniform electric field pointing down the page and a uniform magnetic field pointing into it.

Example 26.3

An electron starts from rest at the origin in the uniform fields $\mathbf{E} = -E_0\hat{\mathbf{y}}$ and $\mathbf{B} = -B_0\hat{\mathbf{z}}$, with $E_0 = 1.50\,\text{MV/m}$ and $B_0 = 10.0\,\text{mT}$, so that $E_0 \approx 0.5cB_0$. Use the relativistic Boris method to follow the electron until it next comes to rest. How high does it rise, what largest speed and kinetic energy does it reach, and where on the x axis and when does it come to rest again? Check the results in the frame in which the electric field vanishes, and compare them with what Newtonian mechanics predicts.

Solution:



The first step is a sketch of the fields and the electron. The electric force on it points up, and once it moves, the magnetic force turns it to the right. The unknowns are the height and length of the arch it follows, and its speed at the top of the arch, where the electric field has done the most work on it.

The program uses the relativistic Boris method of Section 25.13 with $q = -e$ and $m = m_e$, 2000 steps per $2\pi m_e/eB_0 = 3.57\,\text{ns}$ (the cyclotron period of a slow electron), and $\mathbf{w}$ at the start set half a step back, as in the tips above. The electron rises, turns to the right, and comes to rest again on the x axis, at a cusp of its path, after one arch (Figure 26.14a). At the top of the arch it is $y = 0.2275\,\text{m}$ above the axis and moves at $0.8003c$, with $\gamma = 1.668$ and the kinetic energy $(\gamma-1)m_ec^2 = 0.341\,\text{MeV}$. It reaches the cusp at $x = 0.8256\,\text{m}$, at $t = 5.504\,\text{ns}$. With 1000 or 4000 steps per period instead, none of these digits changes.

To check, find the drift frame. Since $\mathbf{E}\times\mathbf{B} = (-E_0\hat{\mathbf{y}})\times(-B_0\hat{\mathbf{z}}) = E_0B_0\hat{\mathbf{x}}$, by (26.29) the drift velocity is $\mathbf{v}_d = (E_0/B_0)\hat{\mathbf{x}}$, with $\beta_d = E_0/cB_0 = 0.5003$ and $\gamma_d = 1.1550$. Let Rachel move with it. By (26.5),

$$E_{R,y} = \gamma_d(-E_0 + v_dB_0) = 0, \qquad B_{R,z} = \gamma_d\left(-B_0 + \frac{v_d}{c^2}E_0\right) = -\frac{B_0}{\gamma_d},$$

a magnetic field of $8.658\,\text{mT}$ into the page and nothing else. In her frame the electron starts with the velocity $-v_d\hat{\mathbf{x}}$, and the magnetic field turns it on a circle at the constant speed v_d, so its Lorentz factor is γ_d. By (25.39), the radius and the period of the circle are

$$R_R = \frac{\gamma_d m_e v_d}{eB_0/\gamma_d} = \frac{\gamma_d^2 m_e v_d}{eB_0} = 0.11377\,\text{m}, \qquad T_R = \frac{2\pi\gamma_d m_e}{eB_0/\gamma_d} = 4.765\,\text{ns}.$$

Distances along y are the same in both frames, so the arch is $2R_R = 0.2275\,\text{m}$ high. At the top of the circle the electron moves at $+v_d$ in Rachel's frame, so its speed in the lab, by velocity addition (25.12), is $2v_d/(1+\beta_d^2) = 0.8003c$. The two cusps are events at the same place in Rachel's frame, T_R apart, so by (25.6) they are $\gamma_dT_R = 5.504\,\text{ns}$ apart in the lab, and $v_d\gamma_dT_R = 0.8256\,\text{m}$ apart along x. All of these agree with the simulation.

Transforming every computed event to Rachel's frame, as in the tips, puts each one on a circle of radius 0.11377 m, to all five digits (Figure 26.14b).

The same program with Newton's law of motion in place of the relativistic one gives a cycloid only 0.171 m high, whose arches repeat every 3.57 ns and every 0.536 m, and on which the electron reaches $1.0007c$ at the top of each arch. The average velocity along x is the same, E_0/B_0, in both. Relativity makes each arch higher and longer and makes the electron take longer to cover it, and it keeps the speed below c.

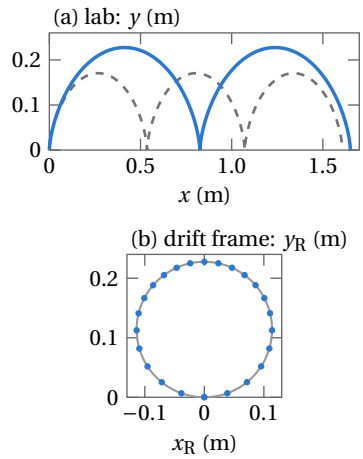


Figure 26.14 (a) The path of the electron of Example 26.3 in the lab (blue), computed with the relativistic Boris method, and the path that Newtonian mechanics predicts (dashed). (b) The computed events of the first arch, transformed to the drift frame (dots), on a circle (gray).

26.11 Summary

The electric and magnetic fields are not two separate vectors but the six components of one four-tensor, $F^{\mu\nu}$, and the rules that carry the fields from one frame to another are the rules for transforming that tensor. Magnetism itself appears when relativity is applied to the electric field of moving charges. The sources ρ and $\mathbf{J}$ form the four-current, the potentials V and $\mathbf{A}$ form the four-potential, and in four-vector notation Maxwell's equations become two equations that read the same for every inertial observer.

With these tools, the fields of a uniformly moving charge, the Doppler effect for light, and the motion of a charge in crossed fields follow from simple transformations, without retarded times. The energy and momentum of the fields form a four-tensor of their own, whose four-divergence expresses the conservation of energy and momentum.

The transformation of the fields

$E_{R,x} = E_{L,x}, \quad E_{R,y} = \gamma\left(E_{L,y} - vB_{L,z}\right),$ $E_{R,z} = \gamma\left(E_{L,z} + vB_{L,y}\right)$	Rachel moves along x at v relative to Luke; the inverse by $v \to -v$
$B_{R,x} = B_{L,x}, \quad B_{R,y} = \gamma\left(B_{L,y} + \frac{v}{c^2}E_{L,z}\right),$ $B_{R,z} = \gamma\left(B_{L,z} - \frac{v}{c^2}E_{L,y}\right)$	the magnetic field
$\mathbf{E}_{R\perp} = \gamma\left(\mathbf{E}_L + \mathbf{v}\times\mathbf{B}_L\right)_\perp,$ $\mathbf{B}_{R\perp} = \gamma\left(\mathbf{B}_L - \frac{\mathbf{v}\times\mathbf{E}_L}{c^2}\right)_\perp$	for $\mathbf{v}$ in any direction; the components along $\mathbf{v}$ do not change
$\mathbf{B} = \frac{\mathbf{u}\times\mathbf{E}}{c^2}$	the magnetic field of charges that all move with velocity $\mathbf{u}$
$E^2 - c^2B^2, \quad \mathbf{E}\cdot\mathbf{B}$	the same in every inertial frame

Four-vectors and four-tensors

$J^\mu = (c\rho, \mathbf{J}) = \rho_0 \eta^\mu$	four-current; for charges that move together, ρ_0 is their proper charge density
$\partial_\mu = \left(\frac{1}{c}\frac{\partial}{\partial t}, \nabla\right), \quad \partial^\mu \partial_\mu = \Box^2$	four-gradient; the d'Alembertian
$A^\mu = (V/c, \mathbf{A})$	four-potential; a four-vector for the retarded potentials of the Lorenz gauge
$k^\mu = (\omega/c, \mathbf{k}), \quad \phi = k_\mu x^\mu$	four-wavevector; the phase ϕ is the same in every frame
$F^{0i} = E_i/c,$ $F^{23} = B_x, \quad F^{31} = B_y, \quad F^{12} = B_z$	the field tensor, antisymmetric; the dual $G^{\mu\nu}$ has $\mathbf{E}/c \to \mathbf{B}$, $\mathbf{B} \to -\mathbf{E}/c$
$F_{\mathrm{R}}^{\mu\nu} = \Lambda^\mu{}_\lambda \Lambda^\nu{}_\sigma F_{\mathrm{L}}^{\lambda\sigma}$	the transformation of a four-tensor; $F_{\mathrm{R}} = \Lambda F_{\mathrm{L}} \Lambda^{\mathsf{T}}$
$T^{00} = u, \quad T^{0i} = T^{i0} = S_i/c$	the stress–energy tensor; its space part T^{ij} is the negative of the Maxwell stress tensor

Electromagnetism in tensor form

$\partial_\mu J^\mu = 0$	the continuity equation: charge is conserved in every frame
$\partial_\nu F^{\mu\nu} = \mu_0 J^\mu, \quad \partial_\nu G^{\mu\nu} = 0$	Maxwell's equations
$K^\mu = q\,\eta_\nu F^{\mu\nu}$	the Lorentz force law, in four-vector form
$F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu$	the fields from the potentials; a gauge transformation adds $\partial^\mu \psi$ to A^μ
$\partial_\mu A^\mu = 0, \quad \Box^2 A^\mu = -\mu_0 J^\mu$	the Lorenz condition and the wave equations for the potentials
$A^\mu(x) = \frac{\mu_0}{2\pi} \int J^\mu(x')\,\Theta(t-t') \times \delta\big((x-x')_\nu (x-x')^\nu\big)\, d^4x'$	the retarded potentials, written so that each factor but J^μ is a Lorentz scalar
$\partial_\nu T^{\mu\nu} = -F^{\mu\nu} J_\nu$	conservation of energy ($\mu = 0$, Poynting's theorem) and of momentum

Applications

$$\mathbf{E} = \frac{q}{4\pi\varepsilon_0}\,\frac{1-\beta^2}{\left(1-\beta^2\sin^2\theta\right)^{3/2}}\,\frac{\hat{\mathbf{R}}}{R^2},$$
$$\mathbf{B} = \frac{\mathbf{v}\times\mathbf{E}}{c^2}$$

a charge moving with constant velocity $\mathbf{v}$; $\mathbf{R}$ runs from its present position

$$\omega_{\mathrm{R}} = \sqrt{\frac{1-\beta}{1+\beta}}\,\omega_{\mathrm{L}},$$
$$\omega_{\mathrm{R}} = \gamma\,\omega_{\mathrm{L}}\left(1-\beta\cos\theta_{\mathrm{L}}\right)$$

the Doppler effect for light: an observer moving away from the source; any direction

$$\mathbf{v}_d = \frac{\mathbf{E}\times\mathbf{B}}{B^2}, \quad B_d = \frac{B}{\gamma_d}$$

crossed fields with $E < cB$: in the frame moving with $\mathbf{v}_d$ there is only a magnetic field

Exercises

Exercises for 26.1 Magnetism from relativity

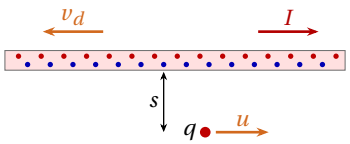


Figure 26.15 The wire of Exercise P26.1 in Luke's frame: its ions (red) at rest, its electrons (blue) drifting to the left, and the charge q moving to the right at u, a distance s from the wire.

P26.1 A long, straight, neutral wire along the x axis carries a current I in the $+x$ direction. In the wire's frame, which is Luke's, its positive ions are at rest, with charge per unit length λ, and its conduction electrons drift in the $-x$ direction at speed v_d, with charge per unit length $-\lambda$, so that $I = \lambda v_d$. A charge q moves parallel to the wire in the $+x$ direction at speed u, a distance s from it (Figure 26.15).

(a) Find the magnetic force on q in Luke's frame.

(b) In the frame of the charge, which is Rachel's, find the charge per unit length of the ions and that of the electrons, using (26.2) for the electrons. Show that the wire carries the net charge per unit length $\lambda_R = -\gamma_u Iu/c^2$.

(c) Find the force on q in Rachel's frame, and transform it to Luke's frame with (25.41). Compare with your answer to (a).

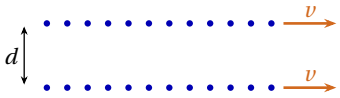


Figure 26.16 The two electron beams of Exercise P26.2, a distance d apart, moving side by side at v.

P26.2 Two long, parallel beams of electrons, a distance d apart, move side by side in the $+x$ direction at the same speed v in the lab (Figure 26.16). In the beams' own rest frame each has the charge per unit length λ_0, which is negative.

(a) In the beams' rest frame, find the force per unit length that one beam exerts on the other.

(b) In the lab, find the charge per unit length and the current of each beam, and the electric and magnetic forces per unit length that each beam exerts on the other. Show that the magnetic attraction cancels the fraction $\beta^2 = v^2/c^2$ of the electric repulsion.

(c) In the beams' rest frame there is only repulsion, at the strength you found in (a). Reconcile the two frames by comparing the forces on the same segment of a beam, of length ℓ_0 in its own frame, with the help of (25.41).

Exercises for 26.2 The transformation of E and B

P26.3 Two checks of the transformation rules (26.5).

(a) Show that applying the rules first with v and then with $-v$ returns the original fields.

(b) Show that two transformations in succession along x, first at v_1 and then at v_2, give the same fields as one transformation at the speed $v = (v_1 + v_2)/(1 + v_1 v_2/c^2)$ of (25.12). It is enough to do this for E_y and B_z. (Hint: show first that $\gamma_1\gamma_2(1 + v_1 v_2/c^2) = \gamma$.)

P26.4 A large, flat sheet of charge lies in the plane $y = 0$. In its own rest frame it has the surface charge density σ_0; in the lab it moves along x at speed u (Figure 26.17).

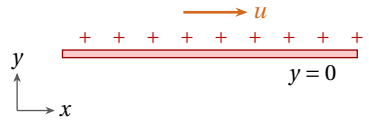


Figure 26.17 The sheet of charge of Exercise P26.4, in the plane $y = 0$, moving along x at u.

(a) In the lab, find the sheet's surface charge density and surface current density, and use Gauss's law and (11.6) to find **E** and **B** above and below the sheet.

(b) Show that on each side $\mathbf{B} = \mathbf{u} \times \mathbf{E}/c^2$, the special case of Section 26.2.3.

(c) Find the same fields by transforming the sheet's rest-frame fields with (26.5).

P26.5 In Luke's frame a uniform magnetic field $\mathbf{B}_L = -B_0\hat{\mathbf{z}}$ fills the region $x < 0$, between the poles of a large magnet, and there is no field for $x > 0$; ignore the fringing field at the edge. A rectangular loop of wire, of width w along x and height h along y, lies in the xy plane and moves in the $+x$ direction at speed v, with its left side in the field and its right side outside it (Figure 26.18). Rachel rides with the loop.

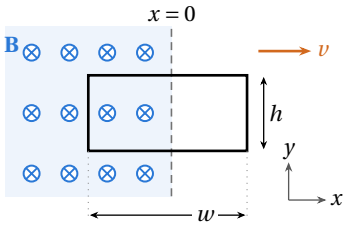


Figure 26.18 Exercise P26.5: a rectangular loop moving at speed v out of a region of uniform magnetic field (into the page) that ends at $x = 0$.

(a) In Luke's frame, find the force per unit charge on the charges in each side of the loop, and the EMF around the loop.

(b) Use (26.5) to find the electric and magnetic fields in Rachel's frame inside the field region. How does the edge of the region move in her frame?

(c) In Rachel's frame the charges of the loop are at rest. Find the EMF around the loop as the line integral of $\mathbf{E}_R$, and again from Faraday's law, from the rate at which the flux through the loop changes.

(d) Compare the EMFs of (a) and (c). Explain why their ratio is what the force transformation (25.41) requires.

Exercises for 26.3 The field tensor

P26.6 Lower both indices of $F^{\mu\nu}$ and of $G^{\mu\nu}$, by changing the sign of every component that has a 0 in an index (Section 25.7.3), and show that $F_{\mu\nu}F^{\mu\nu} = 2(B^2 - E^2/c^2)$ and $G_{\mu\nu}F^{\mu\nu} = -4\mathbf{E}\cdot\mathbf{B}/c$, (26.10).

P26.7 Section 26.3 checked two components of (26.8) by hand.

(a) Find F_R^{03} and F_R^{12} by hand, and show that they agree with (26.5).

(b) With Mathematica or Python, compute all sixteen components of $\Lambda F_{\mathrm{L}} \Lambda^{\mathsf{T}}$ and compare them with F_{R} built from the fields of (26.5).

(c) Do the same for $G^{\mu\nu}$.

P26.8 At a certain event, Luke measures an electric field but no magnetic field: $\mathbf{E}_{\mathrm{L}} \neq 0$ and $\mathbf{B}_{\mathrm{L}} = 0$.

(a) Could Rachel, or any other inertial observer, measure a magnetic field but no electric field at the same event? If so, find that magnetic field; if not, explain how you know.

(b) Answer the same question with the roles of the fields exchanged: Luke measures $\mathbf{B}_{\mathrm{L}} \neq 0$ and $\mathbf{E}_{\mathrm{L}} = 0$. Could another observer measure an electric field but no magnetic field?

(c) At another event Luke measures $\mathbf{E}_{\mathrm{L}} = E_0\hat{\mathbf{y}}$ and $\mathbf{B}_{\mathrm{L}} = (E_0/c)\hat{\mathbf{y}}$. Is there a frame in which the field at this event is purely electric, or one in which it is purely magnetic?

P26.9 A parallel-plate capacitor at rest in Luke's frame has its plates parallel to the xy plane and the uniform field $\mathbf{E}_{\mathrm{L}} = E_0\hat{\mathbf{z}}$ between them.

(a) Rachel moves along x at speed v. Find the fields $\mathbf{E}_{\mathrm{R}}$ and $\mathbf{B}_{\mathrm{R}}$ between the plates.

(b) What current does Rachel see that produces her magnetic field? Check its magnitude with Ampère's law.

(c) Another observer moves along z, perpendicular to the plates, at speed v. Find the fields this observer measures between the plates, and explain the result in terms of the plates' surface charge density.

(d) Check that $E^2 - c^2B^2$ and $\mathbf{E}\cdot\mathbf{B}$ have the same values for all three observers.

Exercises for 26.4 The four-current and the four-gradient

P26.10 A distribution of charge with proper charge density ρ_0 moves at velocity $\mathbf{u}$, forming a current density.

(a) A Lorentz scalar can be formed from the four-current, $J_\mu J^\mu$. What is its value? (Hint: find its value in the rest frame of the charges; it has the same value in every frame.) Check your answer by computing $J_\mu J^\mu$ in a frame in which the charges move at $\mathbf{u}$.

(b) What is $J_\mu J^\mu$ for a neutral wire that carries a current density $\mathbf{J}$? Is the four-current of such a wire timelike, spacelike, or lightlike?

P26.11 In Luke's frame a long, straight wire along the x axis is neutral and carries a uniform current density J in the $+x$ direction.

(a) Use the transformation of the four-current to find the charge density and the current density in Rachel's frame.

(b) Apply your result to the two lines of charge of Section 26.1, and recover (26.3).

(c) A wire carries a current of 10 A, and Rachel moves along it, in the direction of the current, at 1.0 m/s. Find the net charge per unit length that she measures on the wire, and the number of extra electrons per meter that it amounts to. Find the electric field that this charge produces 1.0 cm from the wire, and compare it with the magnetic force per unit charge, vB, that Luke finds on a charge moving with Rachel at that distance.

Exercises for 26.5 Maxwell's equations in tensor form

P26.12 Charge conservation follows from Maxwell's equations.

(a) Show that $\partial_\mu \partial_\nu F^{\mu\nu} = 0$ for any antisymmetric $F^{\mu\nu}$ whose components have continuous second derivatives. (Hint: the order of the two derivatives does not matter, and the names of two summed indices can be exchanged.)

(b) Use (a) and the first of (26.16) to show that $\partial_\mu J^\mu = 0$. Compare with the argument of Section 16.1.2.

Exercises for 26.6 The four-potential

P26.13 Use (26.21) to answer the following.

(a) Write $\delta(x^2 - a^2)$, for $a > 0$, as a sum of delta functions, and evaluate $\int_{-\infty}^{\infty} e^x \delta(x^2 - a^2)\, dx$.

(b) Write $\delta(x^3 - x)$ as a sum of delta functions, and evaluate $\int_{-2}^{2} (x^2 + 1)\delta(x^3 - x)\, dx$ and $\int_{-1/2}^{2} (x^2 + 1)\delta(x^3 - x)\, dx$.

(c) Evaluate $\int_1^\infty e^{-x} \delta(\sin x)\, dx$.

P26.14 A point charge q is at rest at the origin of Luke's frame, where its potentials are $V_\mathrm{L} = q/4\pi\varepsilon_0 r_\mathrm{L}$ and $\mathbf{A}_\mathrm{L} = 0$. These potentials satisfy both the Coulomb condition and the Lorenz condition.

(a) Use (26.18) and the Lorentz transformation to find Rachel's potentials V_R and $\mathbf{A}_\mathrm{R}$ as functions of her coordinates $(x_\mathrm{R}, y_\mathrm{R}, z_\mathrm{R}, t_\mathrm{R})$.

(b) Show that Rachel's potentials satisfy the Lorenz condition.

(c) Show that $\nabla \cdot \mathbf{A}_R$, the divergence with respect to Rachel's coordinates, is zero at each instant only on one plane, so that her potentials do not satisfy the Coulomb condition, although Luke's do.

P26.15 Two more parts of (26.19).

(a) Use (26.13) to find F^{12} and F^{23} in terms of the components of $\mathbf{A}$, and show that they are the components B_z and B_x of $\nabla \times \mathbf{A}$.

(b) Show that the gauge transformation $A^\mu \to A^\mu + \partial^\mu \psi$ is the transformation (15.12) of V and $\mathbf{A}$, and that it leaves every component of $F^{\mu\nu}$ unchanged.

Exercises for 26.7 The fields of a uniformly moving charge

P26.16 Gauss's law for a moving charge.

(a) Show by integration that the flux of the electric field (26.23) through a sphere of radius R centered on the charge's present position is q/ε_0, whatever the speed. (Hint: substitute $u = \cos\theta$, and use $\int du/(a^2 + b^2u^2)^{3/2} = u/(a^2\sqrt{a^2+b^2u^2})$.)

(b) Explain why the answer had to be q/ε_0, without doing the integral.

P26.17 The electrons and positrons of Exercise P25.26, with energies of 50 GeV, travel along a straight beam line, and you stand 1.00 m from it.

(a) What are the largest electric and magnetic fields that you measure as an electron passes? Compare the electric field with that of an electron at rest 1.00 m away.

(b) For how long is the electric field more than half its largest value? (Use the component of $\mathbf{E}$ perpendicular to the beam line, which is the large one near the moment of closest approach.)

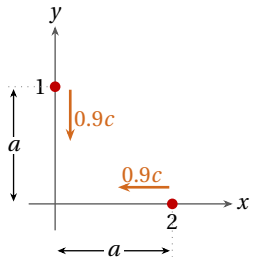


Figure 26.19 The two charges of Exercise P26.18, each a distance a from the origin and moving toward it at $0.9c$.

P26.18 Two particles with the same charge q move at the constant speed $0.9c$ along perpendicular straight lines toward the origin, particle 1 down the y axis and particle 2 to the left along the x axis (Figure 26.19). At a certain instant each is a distance a from the origin. (Other forces, which you may ignore, keep them moving in straight lines.)

(a) Find the electric and magnetic fields that particle 1 produces at the position of particle 2 at that instant, and the force of particle 1 on particle 2.

(b) Find the force of particle 2 on particle 1.

(c) Show that the two forces are not equal and opposite. Explain why this does not violate the conservation of momentum (Section 25.12).

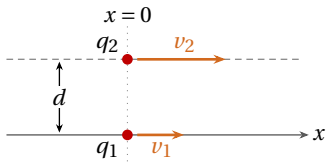


Figure 26.20 The charges of Exercise P26.19 at the moment when q_2 overtakes q_1.

P26.19 Two charges, q_1 and q_2, move in the $+x$ direction in the lab with constant speeds v_1 and $v_2 > v_1$: q_1 along the x axis, and q_2 along a line a distance d above it, in the xy plane (Figure 26.20). At $t = 0$, as both cross the line $x = 0$, q_2 overtakes q_1.

(a) Find the electric and magnetic fields that q_1 produces at the position of q_2 at that instant, in the lab.

(b) Find the same fields in the frame of q_1, in which q_2 moves to the right at $(v_2 - v_1)/(1 - v_1 v_2/c^2)$.

(c) Show that $E^2 - c^2B^2$ and $\mathbf{E}\cdot\mathbf{B}$ have the same values in the two frames, as (26.10) requires.

P26.20 A long, straight line of charge along the x axis has the charge per unit length λ_0 in its own rest frame, which is Rachel's, and it moves in the $+x$ direction at speed v in Luke's frame.

(a) Transform Rachel's fields to Luke's frame with (26.6), and show that in Luke's frame $\mathbf{E} = (\gamma\lambda_0/2\pi\varepsilon_0 s)\hat{\mathbf{s}}$ and $\mathbf{B} = (\mu_0 I/2\pi s)\hat{\boldsymbol{\phi}}$, with $I = \gamma\lambda_0 v$, where s is the distance from the line.

(b) Check both fields with Gauss's law and Ampère's law in Luke's frame.

(c) The field of each charge on the line is (26.23), stronger than Coulomb's beside the charge and weaker ahead and behind. Add up the fields of all the charges, with Luke's charge per unit length $\gamma\lambda_0$, and show that the sum agrees with (a). (Hint: $\int_{-\infty}^{\infty} dx/(x^2 + b^2)^{3/2} = 2/b^2$.)

P26.21 What fraction of the flux of the electric field (26.23) passes through the band of directions within 10° of the plane perpendicular to the motion, $80° < \theta < 100°$? Answer for $\gamma = 1$, 10, and 100. (The integral of Exercise P26.16 helps.) For $\gamma = 10$ and 100, find also the angle δ such that half of the flux lies within δ of that plane, and compare it with $1/\gamma$.

P26.22 The first corrections to the low-speed fields.

(a) Expand the electric field (26.23) in powers of β, and show that to second order

$$\mathbf{E} = \frac{q}{4\pi\varepsilon_0}\left[1 + \beta^2\left(\frac{3}{2}\sin^2\theta - 1\right)\right]\frac{\hat{\mathbf{R}}}{R^2}.$$

(b) In which directions is the correction positive, and in which negative? At what angle does it vanish?

(c) An electron moves at 1.0% of the speed of light. By what fraction does its electric field directly ahead of it differ from Coulomb's field at the same distance, and by what fraction does its field directly beside it differ? How large is the error in the low-speed magnetic field of Section 26.7?



Figure 26.21 Exercise P26.23: light from a source at rest reflected by a mirror that moves away from it at v.

Exercises for 26.8 The Doppler effect for light

P26.23 A mirror moves away from a source of light at speed v, along the direction of the light, and reflects the light straight back (Figure 26.21).

(a) Show that the reflected light returns to the source with the angular frequency $\omega(1-\beta)/(1+\beta)$, where ω is the angular frequency that the source emits. (Hint: find the frequency in the mirror's frame first.)

(b) Radar measures speeds in this way. A radar wave of frequency 24.0 GHz reflects from a car that moves directly away from the radar at 30.0 m/s. Find the change in frequency of the reflected wave, and compare it with the first-order estimate $-2\beta f$.

P26.24 The Doppler effect and aberration from k^μ.

(a) In Luke's frame a plane wave of angular frequency ω_L travels in the xy plane at the angle θ_L to the x axis. Write its four-wavevector, transform it to Rachel's frame, and find ω_R and the angle θ_R at which the wave travels in her frame.

(b) Show that your result for θ_R is the aberration formula (25.14), solved for the angle in Rachel's frame.

(c) A source moves at $\beta = 0.005$ toward a spectrometer at rest in the lab and emits light of wavelength 486.1 nm in its own frame. A mirror behind the source, at rest in the lab, lets the spectrometer see at the same time the light that the source emits away from it. Find the two wavelengths that the spectrometer measures, and show that their average is γ times the wavelength that the source emits. How far from 486.1 nm is the average? (Ives and Stilwell made a measurement of this kind; the reference is in Section 26.8.)

P26.25 Intensity and photons.

(a) The intensity of a plane wave is proportional to the square of its amplitude. Show that Rachel, moving away from the source as in Section 26.8, measures an intensity $(1-\beta)/(1+\beta)$ times Luke's.

(b) Explain the factor in terms of photons. By what factor does the energy of each photon change, and by what factor does the number

of photons that reach a unit area per unit time change? (Hint: a wave train with a given number of crests contains the same number of photons for every observer.)

Exercises for 26.9 The stress–energy tensor

P26.26 The momentum components of (26.28).

(a) Write out the component $\mu = 1$ of (26.28) in terms of S_x, the components T^{1j}, and ρ, **E**, **J**, and **B**.

(b) Interpret each term: which is the rate of change of the momentum density of the fields, which describes the momentum that flows through the surface of a small volume, and which is the force on the charges inside it?

(c) Give an example of a region in which **E** and **B** are both nonzero and $\partial_\nu T^{\mu\nu} = 0$, and one in which $\partial_\nu T^{\mu\nu} \neq 0$. In each case, say what happens to the energy and the momentum of the fields.

P26.27 The plane wave $\mathbf{E} = E_0 \cos(kx - \omega t)\hat{\mathbf{y}}$, $\mathbf{B} = (E_0/c)\cos(kx - \omega t)\hat{\mathbf{z}}$ travels through empty space.

(a) Find every component of its stress–energy tensor. Show that $T^{00} = T^{01} = T^{11}$, that all the other components are zero, and that the trace $T^{\mu}{}_{\mu}$ is zero.

(b) What does T^{11} say about the momentum that the wave carries?

(c) Transform $T^{\mu\nu}$ with (25.19) to Rachel's frame, which moves along x at v, and show that at each event $T^{00}_{\mathrm{R}} = T^{00}_{\mathrm{L}}(1-\beta)/(1+\beta)$. Compare with Exercise P26.25.

Exercises for 26.10 Numerical method: motion in crossed fields

P26.28 Crossed fields with $E > cB$. An electron starts from rest in the uniform fields $\mathbf{E} = -E_0\hat{\mathbf{y}}$ and $\mathbf{B} = -B_0\hat{\mathbf{z}}$ of Example 26.3, but now with $E_0 = 4.50\,\mathrm{MV/m}$ and $B_0 = 10.0\,\mathrm{mT}$.

(a) Show that no frame moves with the velocity (26.29).

(b) Show that in a frame that moves along x at the speed $u = c^2 B_0/E_0$ the magnetic field vanishes, and find the electric field there.

(c) In that frame the electron starts with the velocity $-u\hat{\mathbf{x}}$, perpendicular to the electric field. Show that the component of its momentum along the field grows in proportion to the time, that the component across the field does not change, and that its energy grows without limit.

(d) Follow the electron in the lab with the relativistic Boris method for 20 ns, and plot its kinetic energy against time. Check your program by transforming the computed events to the frame of (b) and comparing with (c).

P26.29 Use the relativistic Boris method for an electron released from rest in the fields of Example 26.3, with $B_0 = 10.0\,\text{mT}$ and with $E_0 = 0.10cB_0$, $0.90cB_0$, and $0.99cB_0$.

(a) For $E_0 = 0.10cB_0$, compare the path with the Newtonian cycloid of Exercise P10.3, adapted to an electron.

(b) For each value of E_0, measure the average velocity along x (the distance between two cusps divided by the time between them) and the time between cusps. Compare them with E_0/B_0 and with the time $2\pi\gamma_d^3 m_e/eB_0$ that Example 26.3 implies.

(c) Explain why the time between cusps grows without limit as E_0 approaches cB_0, although the drift speed stays below c.

Exercises for 26.11 Summary

P26.30 True or false? Explain your reasoning.

(a) The electric field **E** is the space part of a four-vector.

(b) Whether there is an electric field at a given event depends on the frame of the observer.

(c) A charge at rest in Luke's frame produces no magnetic field there, so it produces no magnetic field in any frame.

(d) The electric field of a charge moving at constant velocity points away from the position the charge had at the retarded time.

(e) At a given distance from a charge moving at constant velocity, the electric field is strongest directly ahead of the charge.

(f) The Doppler shift of light depends on whether the source or the observer is moving.

(g) Light that crosses an observer's path at right angles, in the frame of its source, has the same frequency for the observer as for the source.

(h) Maxwell's equations have the same form in every inertial frame.

(i) The energy density of the electromagnetic field has the same value in every inertial frame.

(j) In uniform crossed fields with $E < cB$, there is a frame in which a charge released from rest moves on a circle at constant speed.

P26.31 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as the field lines of a uniformly moving charge at different speeds, the transformation of a field pattern from one frame to another, the field tensor and its invariants in code, the motion of a charge in crossed fields, the Doppler shift of light from a moving source in different directions, or the stress–energy tensor of a plane wave.

Part VIII

Moving Charges and Radiation

> It is now, I believe, generally admitted that the light which we receive from the clear sky is due in one way or another to small suspended particles which divert the light from its regular course. ... Whenever the particles of the foreign matter are sufficiently fine, the light emitted laterally is blue in colour, and, in a direction perpendicular to that of the incident beam, is completely polarized.
>
> — John William Strutt (Lord Rayleigh),
> "On the Light from the Sky, its Polarization and Colour" (1871)*

*"On the light from the sky, its polarization and colour," J. W. Strutt, Phil. Mag. 41, 107–120, 274–279 (1871); reprinted in Lord Rayleigh, *Scientific Papers*, Vol. 1 (Cambridge University Press, Cambridge, 1899), pp. 87–103. https://doi.org/10.1080/14786447108640452.

Chapter 27

Fields of a Point Charge

Chapter 24 found the potentials of any distribution of charges and currents, and Section 26.7 found the fields of a point charge that moves with constant velocity. This chapter finds the fields of a point charge that moves in any way at all. Since the field of any distribution of charges is the sum of the fields of the point charges that make it up, the result is one of the central formulas of electromagnetism.

The direct route to these fields differentiates the retarded potentials, and it is long: the retarded time depends on where and when the fields are wanted, and every derivative has to account for that. This chapter takes the route of four-vectors instead. With the four-current of a point charge, the invariant form of the retarded potentials, (26.22), gives the four-potential in a few lines, and one differentiation, done once for all four components, gives the fields. They come in two parts. One depends only on the velocity of the charge and falls off as $1/\imath^2$, like Coulomb's field. The other is proportional to the acceleration and falls off only as $1/\imath$, and it carries energy away from the charge to arbitrarily large distances: it is radiation. The rest of the chapter is about that radiation: how much of it a charge emits and in which directions, at low speeds (Larmor's formula) and at any speed (Liénard's), and two cases that matter in practice, the synchrotron radiation of charges on curved paths and the bremsstrahlung of charges that are slowed down. The last section computes the fields numerically.

27.1 The Liénard–Wiechert potentials

Let a point charge q move along the world line $w^\mu(\tau)$, where τ is its proper time. In the lab frame the charge is at the position $\mathbf{w}(t')$ at time t', as in Section 24.6.1, and its velocity is $\mathbf{v} = d\mathbf{w}/dt'$. In this chapter only one thing moves, the charge, so its velocity is called $\mathbf{v}$, and $\gamma = 1/\sqrt{1 - v^2/c^2}$ is its Lorentz factor. The charge's four-velocity is $\eta^\mu = \gamma(c, \mathbf{v})$, and its four-current is (26.12). Put that four-current into the invariant form of the retarded potentials, (26.22), and do the integral over x' first. The four-dimensional delta function picks out $x' = w(\tau)$, and

☞ *Caution*: In Chapters 25 and 26, $\mathbf{u}$ was the velocity of an object and $\mathbf{v}$ the velocity of a frame. In this chapter $\mathbf{v}$ is the velocity of the charge.

$$A^\mu(x) = \frac{\mu_0 q c}{2\pi} \int \eta^\mu(\tau)\, \Theta\big(t - t'(\tau)\big)\, \delta\big(\imath_\nu \imath^\nu\big)\, d\tau, \qquad \imath^\nu \equiv x^\nu - w^\nu(\tau),$$

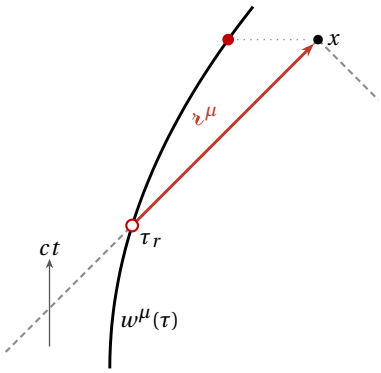


Figure 27.1 The world line of a point charge (black) and the past light cone (dashed) of the field event x. They meet at the shadow event, where the charge was at the retarded time (open circle), and $\mathscr{r}^\mu$ (red) runs from there to x. The solid dot is the charge at the time of the field event.

where $t'(\tau)$ is the lab time at which the charge's clock reads τ. The four-vector $\mathscr{r}^\nu$ runs from an event on the charge's world line to the field event x. Its space part is the vector from the charge's position $\mathbf{w}(t')$ to the field point $\mathbf{r}$, and its time component is $c(t-t')$. The delta function admits only the events of the world line that lie on the light cone of x, where $\mathscr{r}_\nu\mathscr{r}^\nu = 0$, and the step function keeps the one on the past light cone (Figure 27.1). That event is the charge's shadow: the charge was at $\mathbf{w}(t_r)$ at the retarded time t_r, and the news it sent then reaches $\mathbf{r}$ at time t. Since the charge moves more slowly than light, there is exactly one such event (Section 24.6.1). There $\mathscr{r}^0 = c(t-t_r) = \mathscr{r}$, the distance from the shadow to the field point, as in Chapter 24.

To use the composition rule (26.21), we need the derivative of the delta function's argument with respect to τ. Since $d\mathscr{r}^\nu/d\tau = -dw^\nu/d\tau = -\eta^\nu$,

$$\frac{d}{d\tau}\left(\mathscr{r}_\nu\mathscr{r}^\nu\right) = 2\mathscr{r}_\nu\frac{d\mathscr{r}^\nu}{d\tau} = -2\eta_\nu\mathscr{r}^\nu = 2\xi,$$

where

$$\xi \equiv -\eta_\nu\mathscr{r}^\nu = \gamma c\mathscr{r} - \gamma\mathbf{v}\cdot\boldsymbol{\mathscr{r}} = \gamma\left(c\mathscr{r} - \boldsymbol{\mathscr{r}}\cdot\mathbf{v}\right)$$

is a Lorentz scalar. It is positive, since $|\boldsymbol{\mathscr{r}}\cdot\mathbf{v}| \le \mathscr{r}v < \mathscr{r}c$. By (26.21), the step function and the delta function together are $\delta(\tau-\tau_r)/2\xi$, where τ_r is the charge's proper time at the shadow event, and the integral gives

$$A^\mu(x) = \frac{\mu_0 qc}{4\pi}\frac{\eta^\mu}{\xi} = \frac{q}{4\pi\varepsilon_0 c}\frac{\eta^\mu}{\xi}, \tag{27.1}$$

with η^μ and ξ evaluated at the shadow event, and $\mu_0 c = 1/\varepsilon_0 c$. A four-vector divided by a scalar, this is plainly a four-vector. Its time component is V/c and its space part is $\mathbf{A}$, and the factors of γ cancel:

$$\boxed{V(\mathbf{r},t) = \frac{1}{4\pi\varepsilon_0}\frac{qc}{c\mathscr{r} - \boldsymbol{\mathscr{r}}\cdot\mathbf{v}}, \qquad \mathbf{A}(\mathbf{r},t) = \frac{\mathbf{v}}{c^2}V(\mathbf{r},t)}. \tag{27.2}$$

These are the **Liénard–Wiechert potentials**, found by Alfred-Marie Liénard in 1898 and, independently, by Emil Wiechert in 1900. Everything on the right is evaluated at the retarded time: $\boldsymbol{\mathscr{r}} = \mathbf{r} - \mathbf{w}(t_r)$ runs from the shadow to the field point, and $\mathbf{v}$ is the velocity of the charge at t_r.

Written as

$$V = \frac{1}{4\pi\varepsilon_0}\frac{q}{\mathscr{r}\left(1 - \hat{\boldsymbol{\mathscr{r}}}\cdot\mathbf{v}/c\right)},$$

the scalar potential is Coulomb's potential of a charge at the shadow, multiplied by $1/(1-\hat{\boldsymbol{\mathscr{r}}}\cdot\mathbf{v}/c)$. The factor is larger than 1 while the charge moves toward the field point and smaller while it moves away, and it is the factor that stretches and squashes the apparent lengths of Section 24.3.1. That is no coincidence. The retarded potential (24.11) adds up the charge of each volume element at that element's own retarded time, and for a moving charge the far side is counted at an earlier time than the near side. For a charge that approaches, the parts

counted earlier were farther back, so the integral counts the charge as if it were stretched along its motion: however small the charge, the integral of its density at the retarded time is $q/(1-\hat{\boldsymbol{\imath}}\cdot\mathbf{v}/c)$ rather than q (Exercise P27.1). The four-vector route met the same factor as the derivative of the delta function's argument.



Emil Wiechert (1861–1928, German) Wiechert was born in Tilsit, in East Prussia, and studied at the University of Königsberg, where he took his doctorate in 1889 and then taught physics. There, in 1896–97, he helped show that cathode rays are streams of particles far lighter than atoms, at about the same time as J. J. Thomson, though he did not identify the particles as a new kind of particle, the electron. In 1896 he also proposed that the Earth has a dense iron core. He moved to the University of Göttingen in 1897, and in 1898 it made him professor of geophysics, then a new field. There he built up an institute of geophysics and designed a seismograph whose records were far clearer than those of earlier instruments, and his students included Beno Gutenberg, who found the depth of the Earth's core. He found the potentials of a moving point charge in 1900, independently of Liénard. (Wikipedia)

27.1.1 Which velocity?

The velocity in (27.2) is easy to misread, so it is worth saying exactly what it is. The trajectory $\mathbf{w}(t')$ is a function of one time variable, the lab time t' at which the charge is at $\mathbf{w}$, and the velocity at any time is $\mathbf{v}(t') = d\mathbf{w}/dt'$. The $\mathbf{v}$ in (27.2) is this velocity at the instant $t' = t_r$. It is sometimes written $d\mathbf{w}/dt_r$, which means the same thing: the derivative of $\mathbf{w}$ with respect to its own argument, evaluated at the instant t_r. It is the velocity that the charge had when it sent the news that reaches $\mathbf{r}$ at time t, the charge's actual velocity at the retarded time.

It is *not* the velocity at which the observer at $\mathbf{r}$ sees the shadow move. The observer sees the charge at $\mathbf{w}(t_r)$, and as the time of observation t goes on, the retarded time changes too. To see how fast, differentiate the definition of the retarded time, $c(t-t_r) = |\mathbf{r}-\mathbf{w}(t_r)|$, with respect to t at a fixed field point. The distance on the right changes at the rate $-\hat{\boldsymbol{\imath}}\cdot\mathbf{v}$ per unit of t_r, so $c(1-\partial t_r/\partial t) = -\hat{\boldsymbol{\imath}}\cdot\mathbf{v}\,\partial t_r/\partial t$, or

$$\frac{\partial t_r}{\partial t} = \frac{1}{1-\hat{\boldsymbol{\imath}}\cdot\mathbf{v}/c}. \tag{27.3}$$

The shadow, at $\mathbf{w}(t_r)$, therefore moves at $\mathbf{v}\,\partial t_r/\partial t = \mathbf{v}/(1-\hat{\boldsymbol{\imath}}\cdot\mathbf{v}/c)$, which for motion along the line of sight is the apparent speed (24.9) of Section 24.3. When a charge approaches the observer at $v = 0.5c$, for example, its shadow approaches at $0.5c/(1-0.5) = c$, twice as fast as the charge itself ever moves. Section 24.3 pointed out that this apparent velocity is not the velocity of the charge at any time, and the potentials (27.2) contain the charge's actual velocity at the retarded time, not the apparent velocity of its shadow. With the shadow's velocity in its place, (27.2) would give the wrong potentials even for a charge in uniform motion (Exercise P27.2).

Uniform motion is a good check, since Example 26.2 found those potentials another way. Let the charge move with constant velocity $\mathbf{v}$, and let $\mathbf{R}$ be the vector from its present position to the field point. Since $t-t_r = \imath/c$, the charge has moved on by $\mathbf{v}\imath/c$ since the retarded time, so $\mathbf{R} = \boldsymbol{\imath} - \boldsymbol{\beta}\imath$, with $\boldsymbol{\beta} = \mathbf{v}/c$. Then $R^2 = \imath^2 - 2\imath\boldsymbol{\imath}\cdot\boldsymbol{\beta} + \beta^2\imath^2$, and since $\mathbf{R}\times\boldsymbol{\beta} = \boldsymbol{\imath}\times\boldsymbol{\beta}$, also $\beta^2R^2\sin^2\theta = \beta^2\imath^2 - (\boldsymbol{\imath}\cdot\boldsymbol{\beta})^2$, where θ is the angle between $\mathbf{v}$ and $\mathbf{R}$. Subtracting the second from the first gives $R^2(1-\beta^2\sin^2\theta) = (\imath - \boldsymbol{\imath}\cdot\boldsymbol{\beta})^2$, or

$$\imath - \boldsymbol{\imath}\cdot\boldsymbol{\beta} = R\sqrt{1-\beta^2\sin^2\theta}. \tag{27.4}$$

So (27.2) gives $V = q/[4\pi\varepsilon_0 R(1-\beta^2\sin^2\theta)^{1/2}]$ and $\mathbf{A} = (\mathbf{v}/c^2)V$. In Example 26.2, with $R_x = R\cos\theta$ and $R_y^2 + R_z^2 = R^2\sin^2\theta$, Luke's potential is $\gamma q/[4\pi\varepsilon_0 R(\gamma^2\cos^2\theta + \sin^2\theta)^{1/2}]$, and since $\gamma^2\cos^2\theta + \sin^2\theta = \gamma^2(1-\beta^2\sin^2\theta)$, the two agree, as do the vector potentials.

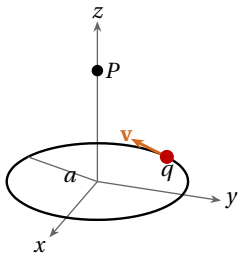
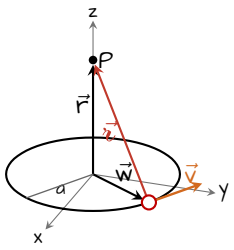


Figure 27.2 Example 27.1: a charge q moving around a circle of radius a in the xy plane, and a point P on the z axis.

Example 27.1

A point charge q moves with constant angular speed ω around a circle of radius a in the xy plane, centered on the origin, so that it is at $\mathbf{w}(t') = a\cos\omega t'\,\hat{\mathbf{x}} + a\sin\omega t'\,\hat{\mathbf{y}}$. Find the Liénard–Wiechert potentials at a point P on the z axis, a distance z from the origin.

Solution:



The first step is a sketch of the circle with the charge at its shadow, and the vectors: $\vec{r}$ to P, $\vec{w}$ to the shadow, and $\vec{\ell}$ from the shadow to P.

The charge's velocity at any time is $\mathbf{v}(t') = a\omega\left(-\sin\omega t'\,\hat{\mathbf{x}} + \cos\omega t'\,\hat{\mathbf{y}}\right)$. The vectors are

$$\begin{aligned}
\mathbf{r} &= z\hat{\mathbf{z}},\\
\mathbf{w}(t_r) &= a\cos\omega t_r\,\hat{\mathbf{x}} + a\sin\omega t_r\,\hat{\mathbf{y}},\\
\boldsymbol{\ell} &= \mathbf{r} - \mathbf{w}(t_r) = -a\cos\omega t_r\,\hat{\mathbf{x}} - a\sin\omega t_r\,\hat{\mathbf{y}} + z\hat{\mathbf{z}},\\
\ell &= \sqrt{a^2+z^2}.
\end{aligned}$$

Every point of the circle is the same distance from P, so ℓ does not depend on t_r, and the retarded time is simply $t_r = t - \sqrt{a^2+z^2}/c$. The velocity is perpendicular to $\boldsymbol{\ell}$:

$$\boldsymbol{\ell}\cdot\mathbf{v}(t_r) = a^2\omega\cos\omega t_r\sin\omega t_r - a^2\omega\sin\omega t_r\cos\omega t_r = 0.$$

So by (27.2),

$$V = \frac{1}{4\pi\varepsilon_0}\,\frac{q}{\sqrt{a^2+z^2}},$$

the potential of a charge at rest at the same distance, constant in time, and

$$\mathbf{A} = \frac{\mathbf{v}(t_r)}{c^2}\,V = \frac{qa\omega}{4\pi\varepsilon_0 c^2\sqrt{a^2+z^2}}\left(-\sin\omega t_r\,\hat{\mathbf{x}} + \cos\omega t_r\,\hat{\mathbf{y}}\right), \qquad t_r = t - \frac{\sqrt{a^2+z^2}}{c}.$$

The vector potential is parallel to the velocity that the charge had at the retarded time, and it turns around the axis at the rate ω, a time $\sqrt{a^2+z^2}/c$ behind the charge. Checks: as $\omega \to 0$ the charge stops, and $\mathbf{A}$ vanishes; and $\mathbf{A}$ is V times a speed divided by c^2, which has the units of $\mathrm{Vs/m} = \mathrm{Tm}$, those of a vector potential. The fields at P cannot be found from these potentials alone, since the fields need derivatives in every direction, and so the potentials off the axis as well; Section 27.2 finds the fields directly.

27.2 The fields of a point charge

The fields follow from the four-potential (27.1) by (26.19), $F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu$. The four-potential depends on the field event x in two ways: directly, through $\mathscr{r}^\mu = x^\mu - w^\mu(\tau_r)$, and through the proper time τ_r of the shadow event, which slides along the world line as x changes. The derivation below takes care of both with the chain rule of Section 26.4. Besides the four-velocity it needs the charge's four-acceleration $\alpha^\mu = d\eta^\mu/d\tau$ at the shadow event, (25.25), and the scalar product $\eta_\mu\eta^\mu = -c^2$, (25.23).

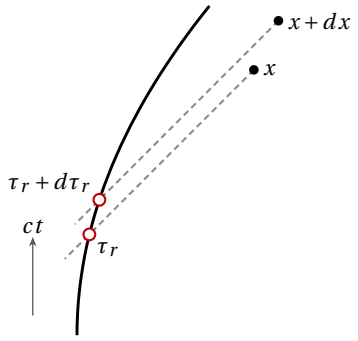


Figure 27.3 Two nearby field events, x and $x+dx$, and their past light cones (dashed). The cones meet the world line (black) at two nearby shadow events (open circles): as the field event moves, the shadow event slides along the world line, and τ_r changes.

The fields from the four-potential

Step 1. Find the four-gradient of τ_r. When the field event moves, the shadow event slides along the world line (Figure 27.3). It is fixed by the condition $\mathscr{r}_\nu\mathscr{r}^\nu = 0$, which holds at every field event, so the four-gradient of $\mathscr{r}_\nu\mathscr{r}^\nu$ is zero:

$$0 = \partial^\mu\left(\mathscr{r}_\nu\mathscr{r}^\nu\right) = 2\mathscr{r}_\nu\,\partial^\mu\mathscr{r}^\nu = 2\mathscr{r}_\nu\left(\partial^\mu x^\nu - \eta^\nu\,\partial^\mu\tau_r\right).$$

By (26.13), $\partial^\mu x^\nu$ is -1 for $\mu=\nu=0$, 1 for $\mu=\nu=1$, 2, or 3, and 0 otherwise, so $\mathscr{r}_\nu\,\partial^\mu x^\nu = \mathscr{r}^\mu$: lowering the index changed the sign of the time component, and $\partial^0 x^0 = -1$ changes it back. With $\mathscr{r}_\nu\eta^\nu = -\xi$,

$$0 = \mathscr{r}^\mu + \xi\,\partial^\mu\tau_r, \qquad \text{so} \qquad \partial^\mu\tau_r = -\frac{\mathscr{r}^\mu}{\xi}.$$

Step 2. Find the four-gradient of $\xi = -\eta_\lambda\mathscr{r}^\lambda$. Here η_λ depends on x through τ_r, with $d\eta_\lambda/d\tau = \alpha_\lambda$, and by the same rule as in Step 1, $\eta_\lambda\,\partial^\mu x^\lambda = \eta^\mu$. So

$$\partial^\mu\xi = -\alpha_\lambda\mathscr{r}^\lambda\,\partial^\mu\tau_r - \eta_\lambda\left(\partial^\mu x^\lambda - \eta^\lambda\,\partial^\mu\tau_r\right) = -\eta^\mu + \left(c^2 + \alpha_\lambda\mathscr{r}^\lambda\right)\frac{\mathscr{r}^\mu}{\xi},$$

since $\eta_\lambda\eta^\lambda = -c^2$.

Step 3. Differentiate the four-potential (27.1), $A^\nu = (q/4\pi\varepsilon_0 c)\,\eta^\nu/\xi$:

$$\partial^\mu A^\nu = \frac{q}{4\pi\varepsilon_0 c}\left(\frac{\alpha^\nu\,\partial^\mu\tau_r}{\xi} - \frac{\eta^\nu\,\partial^\mu\xi}{\xi^2}\right) = \frac{q}{4\pi\varepsilon_0 c}\left[-\frac{\alpha^\nu\mathscr{r}^\mu}{\xi^2} + \frac{\eta^\mu\eta^\nu}{\xi^2} - \left(c^2+\alpha_\lambda\mathscr{r}^\lambda\right)\frac{\eta^\nu\mathscr{r}^\mu}{\xi^3}\right].$$

The middle term is the same with μ and ν exchanged, so it cancels from $F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu$, and the other two give

$$F^{\mu\nu} = \frac{q}{4\pi\varepsilon_0 c}\left(X^\mu\mathscr{r}^\nu - X^\nu\mathscr{r}^\mu\right), \qquad X^\mu = \frac{\xi\,\alpha^\mu + \left(c^2+\alpha_\lambda\mathscr{r}^\lambda\right)\eta^\mu}{\xi^3}. \tag{27.5}$$

Step 4. Read off the fields. By (26.7), $E_i = cF^{0i}$, $B_x = F^{23}$, $B_y = F^{31}$, and $B_z = F^{12}$, so with $\mathscr{r}^0 = \mathscr{r}$, and $\mathbf{X}$ the space part of X^μ,

$$\mathbf{E} = \frac{q}{4\pi\varepsilon_0}\left(X^0\boldsymbol{\mathscr{r}} - \mathbf{X}\mathscr{r}\right), \qquad \mathbf{B} = \frac{q}{4\pi\varepsilon_0 c}\,\mathbf{X}\times\boldsymbol{\mathscr{r}}.$$

Since $\hat{\boldsymbol{\mathscr{r}}}\times\mathbf{E} = -(q/4\pi\varepsilon_0)\mathscr{r}\hat{\boldsymbol{\mathscr{r}}}\times\mathbf{X} = (q/4\pi\varepsilon_0)\mathbf{X}\times\boldsymbol{\mathscr{r}}$, the magnetic field is $\mathbf{B} = \hat{\boldsymbol{\mathscr{r}}}\times\mathbf{E}/c$. For $\mathbf{E}$, put in $\eta^\mu = \gamma(c,\mathbf{v})$ and α^μ from (25.25), with $\mathbf{u}$ replaced by $\mathbf{v}$, and write $\xi = \gamma\kappa$, with $\kappa \equiv c\mathscr{r} - \boldsymbol{\mathscr{r}}\cdot\mathbf{v}$; then $\alpha_\lambda\mathscr{r}^\lambda = \gamma^2\,\mathbf{a}\cdot\boldsymbol{\mathscr{r}} - \gamma^4(\mathbf{v}\cdot\mathbf{a})\kappa/c^2$. In

$$\mathbf{E} = \frac{q}{4\pi\varepsilon_0\xi^3}\left[\xi\left(\alpha^0\boldsymbol{\mathscr{r}} - \boldsymbol{\alpha}\mathscr{r}\right) + \left(c^2+\alpha_\lambda\mathscr{r}^\lambda\right)\left(\eta^0\boldsymbol{\mathscr{r}} - \boldsymbol{\eta}\mathscr{r}\right)\right],$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ are the space parts, the terms with $\mathbf{a}$ come with γ^3 or γ^5, and those with γ^5 cancel. What is left is

$$\mathbf{E} = \frac{q}{4\pi\varepsilon_0\kappa^3}\left[\frac{c^2}{\gamma^2}(c\boldsymbol{\mathscr{r}} - \mathbf{v}\mathscr{r}) + (\mathbf{a}\cdot\boldsymbol{\mathscr{r}})(c\boldsymbol{\mathscr{r}} - \mathbf{v}\mathscr{r}) - \kappa\mathscr{r}\,\mathbf{a}\right].$$

Finally, write $c\boldsymbol{\mathscr{r}} - \mathbf{v}\mathscr{r} = c\mathscr{r}(\hat{\boldsymbol{\mathscr{r}}} - \boldsymbol{\beta})$ and $\kappa = c\mathscr{r}(1 - \hat{\boldsymbol{\mathscr{r}}}\cdot\boldsymbol{\beta})$, with $\boldsymbol{\beta} = \mathbf{v}/c$, and use the BAC-CAB rule (1.49) in the form $\hat{\boldsymbol{\mathscr{r}}}\times(\mathbf{b}\times\mathbf{a}) = \mathbf{b}(\hat{\boldsymbol{\mathscr{r}}}\cdot\mathbf{a}) - \mathbf{a}(\hat{\boldsymbol{\mathscr{r}}}\cdot\mathbf{b})$, with $\mathbf{b} = \hat{\boldsymbol{\mathscr{r}}} - \boldsymbol{\beta}$.

The result is the field of a point charge in arbitrary motion,

$$\boxed{\begin{aligned}\mathbf{E}(\mathbf{r},t) &= \frac{q}{4\pi\varepsilon_0}\frac{1}{(1-\hat{\boldsymbol{\mathscr{r}}}\cdot\boldsymbol{\beta})^3}\left[\frac{(1-\beta^2)(\hat{\boldsymbol{\mathscr{r}}}-\boldsymbol{\beta})}{\mathscr{r}^2} + \frac{\hat{\boldsymbol{\mathscr{r}}}\times[(\hat{\boldsymbol{\mathscr{r}}}-\boldsymbol{\beta})\times\mathbf{a}]}{c^2\mathscr{r}}\right],\\ \mathbf{B}(\mathbf{r},t) &= \frac{1}{c}\hat{\boldsymbol{\mathscr{r}}}\times\mathbf{E}(\mathbf{r},t),\end{aligned}} \tag{27.6}$$

where $\boldsymbol{\beta} = \mathbf{v}/c$, and, as in the potentials, everything on the right is evaluated at the retarded time: $\boldsymbol{\mathscr{r}}$ runs from the shadow to the field point, and $\mathbf{v}$ and $\mathbf{a} = d\mathbf{v}/dt'$ are the charge's actual velocity and acceleration at $t' = t_r$. Exercise P27.5 reaches the same fields by another route, and Exercise P27.9 writes them in a form that uses only what the observer sees.

! **All at the retarded time.** In (27.6), $\boldsymbol{\mathscr{r}}$, $\mathbf{v}$, and $\mathbf{a}$ all belong to the shadow event: the charge's position, velocity, and acceleration at t_r. Where the charge is now, and how it moves now, do not appear.

The electric field has two parts. The first depends on the velocity alone and falls off as $1/\mathscr{r}^2$; for a charge at rest it is Coulomb's field. It is called the **velocity field**, or the generalized Coulomb field. The second is proportional to the acceleration and falls off only as $1/\mathscr{r}$. It is called the **acceleration field**, or the radiation field, for a reason that Section 27.3 explains. The magnetic field is perpendicular to both $\mathbf{E}$ and $\hat{\boldsymbol{\mathscr{r}}}$, and the acceleration field, which is $\hat{\boldsymbol{\mathscr{r}}}$ crossed into a vector, is perpendicular to $\hat{\boldsymbol{\mathscr{r}}}$ as well. Far from the charge, where the acceleration field is much the larger, $\mathbf{E}$, $\mathbf{B}$, and $\hat{\boldsymbol{\mathscr{r}}}$ are therefore perpendicular to one another, as they are in a light wave.

27.2.1 The velocity field

The velocity field is

$$\mathbf{E}_{\text{vel}} = \frac{q}{4\pi\varepsilon_0}\frac{(1-\beta^2)(\hat{\boldsymbol{\mathscr{r}}}-\boldsymbol{\beta})}{\mathscr{r}^2(1-\hat{\boldsymbol{\mathscr{r}}}\cdot\boldsymbol{\beta})^3}. \tag{27.7}$$

It contains the velocity of the charge at the retarded time but not its acceleration, so it is the entire field of a charge that has that velocity at the retarded time and no acceleration then or ever: a charge that has always moved with the velocity $\mathbf{v}$ and always will. That field is (26.23), and it points straight away from the position that such a charge has now. For the real charge, which may have sped up, slowed down, or turned since the retarded time, that position is its **projected position**,

$$\mathbf{w}_{\text{p}} = \mathbf{w}(t_r) + \mathbf{v}(t - t_r),$$

where the charge would be now had it kept the velocity it had at the retarded time (Figure 27.4). So the velocity field of a charge in any motion is (26.23), with **R** measured from the projected position and θ the angle between **R** and the velocity at the retarded time, as Section 26.7 promised (Exercise P27.6 checks it algebraically). The news that reaches the field point left the charge at the retarded time, and it carried the charge's position and velocity at that time; the velocity field is what the field would be had the charge kept that velocity. If the charge has changed its motion since, the news of the change has not yet arrived. For a charge in uniform motion the projected position is the present one, the acceleration field is zero, and (27.6) is (26.23).

Like Coulomb's field, the velocity field falls off as $1/r^2$ far from the charge, and Section 27.3 shows that a field like that carries no energy away to large distances. It is the part of the field that travels along with the charge.

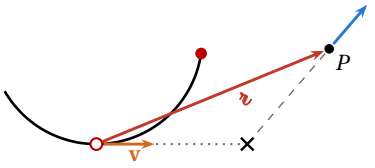


Figure 27.4 A charge moving on a curved path (black). The observer at P sees its shadow (open circle), where it was at the retarded time, moving with the velocity **v** (orange). Had it kept that velocity (dotted line), it would now be at the projected position (×); in fact it is now at the solid dot. The velocity field at P (blue) points straight away from the projected position.

27.2.2 The acceleration field and Thomson's kink

The acceleration field,

$$\mathbf{E}_{\text{rad}} = \frac{q}{4\pi\varepsilon_0 c^2}\frac{\hat{\boldsymbol{r}}\times\left[\left(\hat{\boldsymbol{r}}-\boldsymbol{\beta}\right)\times\mathbf{a}\right]}{r\left(1-\hat{\boldsymbol{r}}\cdot\boldsymbol{\beta}\right)^3}, \tag{27.8}$$

is new. It is proportional to the acceleration of the charge at the retarded time, it is perpendicular to $\hat{\boldsymbol{r}}$, and it falls off only as $1/r$. When the charge moves slowly at the retarded time, $v \ll c$, it is

$$\mathbf{E}_{\text{rad}} = \frac{q}{4\pi\varepsilon_0 c^2 r}\hat{\boldsymbol{r}}\times\left(\hat{\boldsymbol{r}}\times\mathbf{a}\right). \tag{27.9}$$

Its size is easy to see without expanding the cross products. The vector $\hat{\boldsymbol{r}}\times\mathbf{a}$ is perpendicular to $\hat{\boldsymbol{r}}$ and has the magnitude $a\sin\theta$, where θ is the angle between **a** and $\hat{\boldsymbol{r}}$. Crossing the unit vector $\hat{\boldsymbol{r}}$ into a vector perpendicular to it turns that vector through 90° about $\hat{\boldsymbol{r}}$ without changing its length, so $|\hat{\boldsymbol{r}}\times(\hat{\boldsymbol{r}}\times\mathbf{a})| = a\sin\theta$, and

$$E_{\text{rad}} = \frac{q}{4\pi\varepsilon_0}\frac{a\sin\theta}{c^2 r}. \tag{27.10}$$

The field lies in the plane of **a** and $\hat{\boldsymbol{r}}$, and it points against the part of **a** perpendicular to $\hat{\boldsymbol{r}}$, since by the BAC-CAB rule $\hat{\boldsymbol{r}}\times(\hat{\boldsymbol{r}}\times\mathbf{a}) = \hat{\boldsymbol{r}}(\hat{\boldsymbol{r}}\cdot\mathbf{a}) - \mathbf{a}$. A charge that accelerates upward, seen from the side, sends out a radiation field that points down.

J. J. Thomson, whose portrait is in Section 10.1.4, gave a simple picture of this field.[1] Let a charge be at rest at the origin until $t = 0$, then accelerate uniformly for a short time Δt to a small velocity $\Delta\mathbf{v}$, with $\Delta v \ll c$, and then coast. Figure 27.5 shows its field lines at a later time $t \gg \Delta t$. The news of the acceleration has spread into a thin spherical shell of radius ct and thickness $c\,\Delta t$, centered on the origin.

[1] J. J. Thomson, *Electricity and Matter* (Charles Scribner's Sons, New York, 1904), Chapter III. The argument can be made exact at any speed: "Electric field of an accelerating charge," J. R. Tessman and J. T. Finnell, Jr., Am. J. Phys. 35, 523–527 (1967), https://doi.org/10.1119/1.1974162.

Outside the shell the news has not yet arrived, and the field is still Coulomb's field of a charge at the origin. Inside it, the field is that of a charge moving uniformly at $\Delta\mathbf{v}$, which points away from the charge's present position, a distance $\Delta v\, t$ from the origin, and at this low speed it is Coulomb's field of a charge there. The shell holds no charge, so by Gauss's law the field lines continue through it, and since the lines inside and outside carry the same flux in the same directions, each line outside joins the line inside that leaves the charge at the same angle.

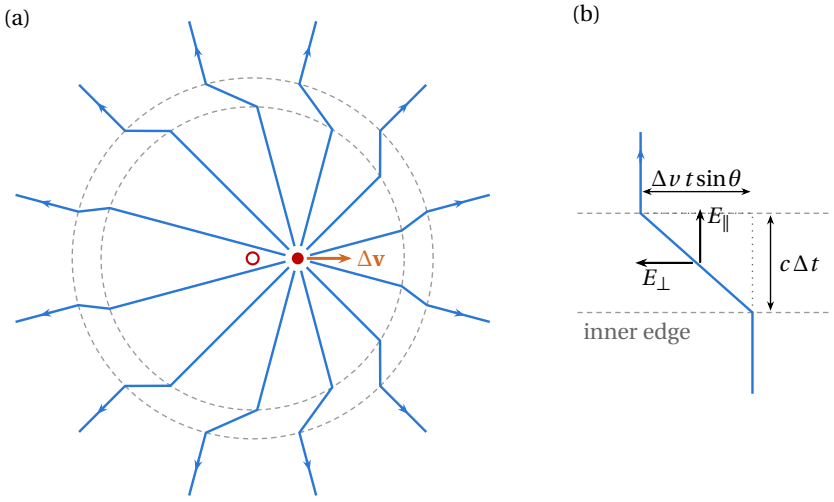


Figure 27.5 Thomson's picture of the acceleration field. (a) A charge at rest at the origin (open circle) was accelerated for a short time Δt to a small velocity $\Delta\mathbf{v}$ (orange) and has since moved uniformly to its present position (solid dot). Outside the sphere of radius ct the news has not arrived, and the field lines point away from the origin; inside the sphere of radius $c(t-\Delta t)$ they point away from the present position; in the shell between, each line kinks. (b) One kink: across the thickness $c\,\Delta t$ of the shell, the line moves sideways by $\Delta v\, t \sin\theta$. The speed and the shell's thickness are exaggerated.

Inside the shell, then, each line runs at a slant. Across the shell's thickness $c\,\Delta t$ it moves sideways by $\Delta v\, t \sin\theta$, where θ is the angle between the line and $\Delta\mathbf{v}$, so the field in the shell has a component $E_\perp$ across the radius, with

$$\frac{E_\perp}{E_\parallel} = \frac{\Delta v\, t \sin\theta}{c\,\Delta t}.$$

The radial component $E_\parallel$ is continuous across the edges of the shell, which carry no charge (Section 4.4), so it is Coulomb's field there, $q/4\pi\varepsilon_0 c^2 t^2$. With $a = \Delta v/\Delta t$ and $\mathscr{r} = ct$,

$$E_\perp = \frac{q}{4\pi\varepsilon_0 c^2 t^2}\,\frac{\Delta v\, t \sin\theta}{c\,\Delta t} = \frac{q}{4\pi\varepsilon_0}\,\frac{a\sin\theta}{c^2\mathscr{r}},$$

exactly (27.10). The direction agrees too: the lines inside start from a point displaced along $\Delta\mathbf{v}$, so in the shell each line slants back, against the part of the acceleration perpendicular to the radius.

The picture also shows why the acceleration field falls off only as $1/\mathscr{r}$. As the shell expands, its thickness stays $c\,\Delta t$, but the sideways step $\Delta v\, t \sin\theta$ grows in

proportion to t, so $E_\perp/E_\parallel$ grows as t, and $E_\perp$ falls off as $t/t^2 = 1/t$, or $1/\mathscr{r}$. The shell, moving outward at the speed of light with a transverse field that falls off as $1/\mathscr{r}$, is a pulse of radiation. The exact field (27.8) extends Thomson's picture to charges of any speed and any acceleration.

27.3 What is radiation?

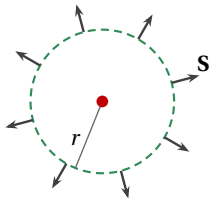


Figure 27.6 A large sphere of radius r (dashed) around a source, and the Poynting vector **S** on it. The power radiated is the limit, as $r \to \infty$, of the power $\oint \mathbf{S}\cdot d\mathbf{a}$ that flows out through the sphere when the news of what the source did reaches it.

Far from a charge, its acceleration field is much larger than its velocity field, since one falls off as $1/\mathscr{r}$ and the other as $1/\mathscr{r}^2$. That difference decides which part of the field can carry energy away for good. Energy that leaves a source and travels outward forever, never to return, is called **radiation**. To measure it, surround the source with a large sphere of radius r (Figure 27.6) and find the power that flows out through it,

$$P(r) = \oint \mathbf{S}\cdot d\mathbf{a}, \qquad \mathbf{S} = \frac{1}{\mu_0}\,\mathbf{E}\times\mathbf{B},$$

where **S** is the Poynting vector (Section 20.2.2), at the moment when the news of what the source did at some time t_0 reaches the sphere. The power radiated at t_0 is the limit of $P(r)$ as $r \to \infty$.

The area of the sphere grows as r^2. A field that falls off as $1/r^2$ gives a Poynting vector that falls off as $1/r^4$, and so a power through the sphere that falls off as $1/r^2$ and vanishes for a large enough sphere: such a field carries energy along with the source, but not away from it. A field that falls off as $1/r$ gives a Poynting vector that falls off as $1/r^2$, and a power through the sphere that does not depend on r: that energy keeps going. (The cross terms between the two kinds of field fall off as $1/r^3$ and vanish too.) So only the acceleration field radiates, as Section 24.5 found for the terms of Jefimenko's equations that contain $\dot\rho$ and $\dot{\mathbf{j}}$, which fall off as $1/\mathscr{r}$ (Exercise P27.10). A charge in uniform motion does not radiate; an accelerated charge does.

For a point charge, center the sphere on the shadow, at $\mathbf{w}(t_r)$, and give it the radius $\mathscr{r} = c(t-t_r)$: at time t, the news of what the charge did at t_r has just reached every point of it. With $\mathbf{B} = \hat{\boldsymbol{\mathscr{r}}}\times\mathbf{E}/c$ and the BAC-CAB rule, the Poynting vector is

$$\mathbf{S} = \frac{1}{\mu_0 c}\,\mathbf{E}\times\left(\hat{\boldsymbol{\mathscr{r}}}\times\mathbf{E}\right) = \frac{1}{\mu_0 c}\left[E^2\hat{\boldsymbol{\mathscr{r}}} - \left(\hat{\boldsymbol{\mathscr{r}}}\cdot\mathbf{E}\right)\mathbf{E}\right].$$

On a large sphere only the acceleration field counts, and it is perpendicular to $\hat{\boldsymbol{\mathscr{r}}}$, so the second term vanishes, and

$$\mathbf{S}_{\text{rad}} = \frac{1}{\mu_0 c}\,E_{\text{rad}}^2\,\hat{\boldsymbol{\mathscr{r}}}. \tag{27.11}$$

The radiation flows straight out from the shadow.

How the radiated power is shared among the directions is described by the power per unit solid angle. The **solid angle** $d\Omega$ of a small patch of a sphere is the patch's area divided by the square of the sphere's radius. In spherical coordinates the area of a patch of a sphere of radius $\mathscr{r}$ is $\mathscr{r}^2\sin\theta\,d\theta\,d\phi$, by (1.15), so

$$d\Omega = \sin\theta\,d\theta\,d\phi, \tag{27.12}$$

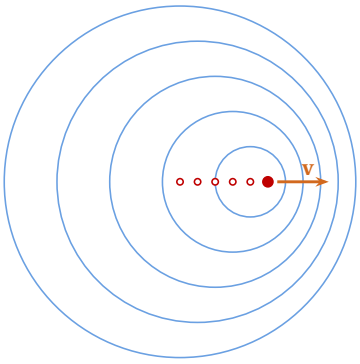


Figure 27.7 The news of a charge moving to the right at $0.5c$. Each circle is as far as the news of one of five equally spaced moments has spread by now, centered on where the charge was at that moment (open circles); the charge is now at the solid dot. Ahead of the charge the circles are crowded together, behind it they are spread apart.

as Section 1.8 lists it. Solid angle is measured in steradians (sr), and the whole sphere is 4π sr, just as a whole circle is 2π radians. A patch of the large sphere with the solid angle $d\Omega$ has the area $r^2\,d\Omega$, and the radiation crosses it at the rate $S_{\text{rad}}r^2\,d\Omega$, which does not depend on r.

One more distinction matters for fast charges: the rate at which energy crosses the sphere is not the rate at which it leaves the charge. The energy that the charge emits in the direction $\hat{\mathbf{r}}$ between the retarded times t_r and $t_r + dt_r$ crosses a distant patch of the sphere between the times t and $t + dt$, and by (27.3), $dt = (1 - \hat{\mathbf{r}}\cdot\mathbf{v}/c)\,dt_r$. When the charge moves toward the patch, it follows the news it has already sent, and the energy emitted during dt_r arrives in the shorter time dt (Figure 27.7). The power per unit solid angle that crosses the sphere is $S_{\text{rad}}r^2$, per unit of the observer's time, so the power per unit solid angle that leaves the charge, per unit of its own time t_r, is

$$\frac{dP}{d\Omega} = \left(1 - \frac{\hat{\mathbf{r}}\cdot\mathbf{v}}{c}\right)\frac{E_{\text{rad}}^2 r^2}{\mu_0 c}. \tag{27.13}$$

This is what is usually wanted: the energy that the charge loses per unit time, sent in each direction. At low speeds the factor $1 - \hat{\mathbf{r}}\cdot\mathbf{v}/c$ is close to 1 and the distinction does not matter, but for fast charges it does (Section 27.5). It is the same compression of the news that raises the frequency of the waves from an approaching source, the Doppler effect.

27.4 Larmor's formula

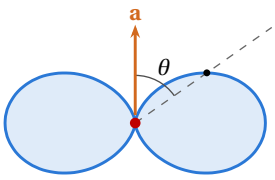


Figure 27.8 The radiation pattern of a slowly moving accelerated charge, (27.14): the distance from the charge to the curve in each direction is proportional to the power per unit solid angle sent that way. The pattern is the same in every plane that contains the acceleration **a** (orange); turned about **a**, it makes a doughnut.

For a charge that moves slowly at the retarded time, $v \ll c$, the factor $1 - \hat{\mathbf{r}}\cdot\mathbf{v}/c$ is close to 1, and the acceleration field is (27.9), with the size (27.10). By (27.11) and (27.13), with $1/(4\pi\varepsilon_0 c^2)^2\mu_0 c = \mu_0/16\pi^2 c$ (since $1/\varepsilon_0 c^2 = \mu_0$),

$$\mathbf{S}_{\text{rad}} = \frac{\mu_0 q^2 a^2}{16\pi^2 c}\frac{\sin^2\theta}{r^2}\hat{\mathbf{r}}, \qquad \frac{dP}{d\Omega} = \frac{\mu_0 q^2 a^2}{16\pi^2 c}\sin^2\theta, \tag{27.14}$$

where θ is the angle between $\hat{\mathbf{r}}$ and the acceleration. The charge radiates most strongly at right angles to its acceleration and not at all along it (Figure 27.8); in three dimensions the pattern is a doughnut around the line of the acceleration. The total power is the integral over all directions, with $d\Omega$ from (27.12):

$$P = \frac{\mu_0 q^2 a^2}{16\pi^2 c}\int_0^{2\pi}\int_0^{\pi}\sin^2\theta\,\sin\theta\,d\theta\,d\phi = \frac{\mu_0 q^2 a^2}{16\pi^2 c}(2\pi)\left(\frac{4}{3}\right),$$

since $\int_0^{\pi}\sin^3\theta\,d\theta = 4/3$. So

$$\boxed{P = \frac{\mu_0 q^2 a^2}{6\pi c}}. \tag{27.15}$$

This is **Larmor's formula**, which Joseph Larmor found in 1897.[2] It holds when

[2] "On the theory of the magnetic influence on spectra; and on the radiation from moving ions," J. Larmor, Phil. Mag. 44, 503–512 (1897), https://doi.org/10.1080/14786449708621095.

the charge moves slowly; Section 27.5 finds the power for any speed. The power depends on the square of the charge, so positive and negative charges radiate alike, and on the square of the acceleration, so a charge radiates whether it speeds up, slows down, or turns. Section 28.2.2 finds the same pattern for an oscillating electric dipole (Exercise P27.16).

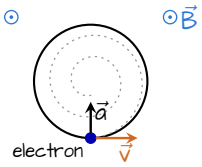


Joseph Larmor (1857–1942, Irish) Larmor was born in Magheragall, County Antrim, and studied at Queen's College, Belfast, and at St John's College, Cambridge, where he was Senior Wrangler in 1880. After five years as a professor at Queen's College, Galway, he returned to Cambridge, and from 1903 to 1932 he was Lucasian Professor of Mathematics, after George Gabriel Stokes. His formula for the radiated power appeared in 1897, in a paper that also explained the effect of a magnetic field on spectral lines by the precession of the electrons' orbits. He was knighted in 1909, and he represented Cambridge University in Parliament from 1911 to 1922. Portrait: Sir Joseph Larmor by Walter Stoneman, bromide print, 1920. © National Portrait Gallery, London, CC BY-NC-ND 3.0 (https://creativecommons.org/licenses/by-nc-nd/3.0/). (Wikipedia)

Example 27.2

A Penning trap holds electrons in a strong uniform magnetic field, here $B = 3.00\,\mathrm{T}$, together with a weak electric field that keeps them from drifting away along the field lines. An electron in the trap moves in a circle perpendicular to the magnetic field, slowly compared with light, and it radiates. Show that the kinetic energy of this circular motion decreases exponentially, and find the time constant, which is called the cooling time.

Solution:



The first step is a sketch of the electron's circle in the field, with its velocity and its acceleration, which points toward the center. The radiation takes energy away, so the circle shrinks.

The magnetic force $|q|vB$ supplies the centripetal force mv^2/R (Section 10.1.4), so the acceleration is

$$a = \frac{v^2}{R} = \frac{|q|vB}{m}.$$

By Larmor's formula (27.15), the electron radiates the power

$$P = \frac{\mu_0 q^2 a^2}{6\pi c} = \frac{\mu_0 q^4 B^2 v^2}{6\pi c m^2}.$$

The kinetic energy is $K = \frac{1}{2}mv^2$, so $v^2 = 2K/m$, and since the magnetic force does no work, the energy decreases only by radiation, at the rate

$$\frac{dK}{dt} = -P = -\frac{\mu_0 q^4 B^2}{3\pi c m^3} K.$$

A quantity that decreases at a rate proportional to itself decays exponentially: $K(t) = K_0 e^{-t/\tau_{cool}}$, with

$$\tau_{cool} = \frac{3\pi c m^3}{\mu_0 q^4 B^2}.$$

For an electron, with $q^4 = e^4$, at $3.00\,\mathrm{T}$, $\tau_{cool} = 0.287\,\mathrm{s}$. The cooling time depends neither on the speed nor on the radius: a faster electron radiates more, in proportion to its energy. As it slows, the radius of its circle, $R = mv/|q|B$, shrinks in proportion to v. Check: the units of $\mu_0 q^4 B^2/cm^3$ are

$$\frac{(\mathrm{kg\,m/s^2A^2})(\mathrm{A^4s^4})(\mathrm{kg^2/s^4A^2})}{(\mathrm{m/s})(\mathrm{kg^3})} = \frac{1}{\mathrm{s}},$$

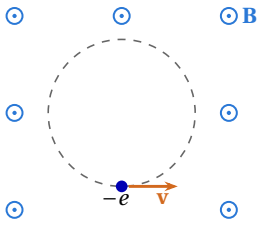


Figure 27.9 Example 27.2: an electron moving in a circle in a uniform magnetic field **B**, which points out of the page.

those of a rate. The electron radiates at its cyclotron frequency (10.6), $eB/2\pi m = 84\,\text{GHz}$, in the microwave range.

The trap is named for Frans Penning, whose vacuum gauge inspired it; Hans Dehmelt built the first one in 1959, and he shared the 1989 Nobel Prize in Physics "for the development of the ion trap technique." Radiation is what cools the electrons in such a trap.

Larmor's formula also explains why the vibrations of a charge die away. A charge held by a force like a spring's oscillates, and as it oscillates it radiates, so its energy decreases. The next example finds how quickly, for the electron of a classical model of an atom.

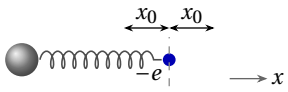
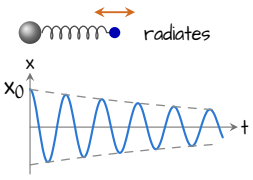


Figure 27.10 Example 27.3: an electron bound by a force like a spring's, oscillating along x with the amplitude x_0.

Example 27.3

In a classical model of an atom, an electron is held in place by a force like a spring's, and it oscillates at the frequency of the light that the atom emits. As it radiates, it loses energy, and its oscillation dies away. (a) Show that the energy of an electron oscillating at the angular frequency ω decays exponentially, at a rate proportional to ω^2. (b) Find the time in which the energy falls by a factor of e, and the number of radians through which the electron oscillates in that time, for the yellow light of sodium, at the wavelength 589 nm.

Solution:



The first step is a sketch of the oscillating electron, and of its position against time: an oscillation whose amplitude shrinks slowly as the energy is radiated.

(a) Let the electron oscillate as $x = x_0 \cos\omega t$, with an amplitude x_0 that changes so slowly that it is nearly constant over one cycle (the answer will justify this). Its acceleration is $a = -\omega^2 x_0 \cos\omega t$, and by Larmor's formula (27.15) it radiates the power

$$P = \frac{\mu_0 e^2 \omega^4 x_0^2}{6\pi c} \cos^2 \omega t.$$

The average of $\cos^2 \omega t$ over a cycle is $\frac{1}{2}$, so the average power is

$$\langle P \rangle = \frac{\mu_0 e^2 \omega^4 x_0^2}{12\pi c}.$$

The energy of the oscillator is $U = \frac{1}{2} m\omega^2 x_0^2$ (at the turning points it is all potential energy), so $x_0^2 = 2U/m\omega^2$, and

$$\frac{dU}{dt} = -\langle P \rangle = -\frac{\mu_0 e^2 \omega^2}{6\pi mc} U = -\Gamma U, \qquad \Gamma = \omega^2 \tau_{\text{rad}}, \qquad \tau_{\text{rad}} \equiv \frac{\mu_0 e^2}{6\pi mc}.$$

So $U(t) = U_0 e^{-\Gamma t}$: the energy decays exponentially, at the rate Γ, proportional to ω^2. The time τ_{rad} depends only on the charge and the mass of the particle; for the electron it is 6.27×10^{-24} s.

(b) At 589 nm, $\omega = 2\pi c/\lambda = 3.20 \times 10^{15}\,\mathrm{s}^{-1}$, so

$$\Gamma = \omega^2 \tau_{rad} = 6.41 \times 10^7\,\mathrm{s}^{-1}, \qquad \frac{1}{\Gamma} = 15.6\,\mathrm{ns}, \qquad \frac{\omega}{\Gamma} = 5.0 \times 10^7.$$

The energy falls by a factor of e in 15.6 ns, during which the electron oscillates through 5.0×10^7 radians, about 8×10^6 cycles, so the amplitude does change slowly, as assumed. The ratio ω/Γ is the oscillator's quality factor Q, and it also sets the width of the spectral line: the light from an oscillator whose energy decays at the rate Γ is spread over a band of angular frequencies about Γ wide, a fraction $1/Q = 2 \times 10^{-8}$ of ω.

For comparison, the measured lifetime of the excited state of sodium that emits this light is 16.24 ns.[3] The classical model, crude as it is, comes within 4% of it. Quantum mechanics changes the classical rate by a factor that depends on the transition, and for this one the factor happens to be close to 1.

A radiating charge also feels a recoil force from its own field. The theory of that force, called the radiation reaction, and its puzzles are beyond this book.

27.5 Liénard's generalization

Larmor's formula holds only for a charge that moves slowly at the retarded time. For a fast charge two things change. Its acceleration field (27.8) contains $\hat{\boldsymbol{r}} - \boldsymbol{\beta}$ in place of $\hat{\boldsymbol{r}}$, and the factor $1/(1-\hat{\boldsymbol{r}}\cdot\boldsymbol{\beta})^3$, which is large in the directions close to $\mathbf{v}$ when β is close to 1. And the power that the charge emits in each direction is not the power that crosses a distant sphere: the two differ by the factor $1 - \hat{\boldsymbol{r}}\cdot\boldsymbol{\beta}$ of (27.13). Putting (27.8) into (27.13), with $1/(4\pi\varepsilon_0 c^2)^2 \mu_0 c = \mu_0/16\pi^2 c$ as in Section 27.4, gives the power per unit solid angle that a charge moving at any speed emits in the direction $\hat{\boldsymbol{r}}$:

$$\frac{dP}{d\Omega} = \frac{\mu_0 q^2}{16\pi^2 c} \frac{|\hat{\boldsymbol{r}} \times [(\hat{\boldsymbol{r}} - \boldsymbol{\beta}) \times \mathbf{a}]|^2}{(1 - \hat{\boldsymbol{r}}\cdot\boldsymbol{\beta})^5}. \tag{27.16}$$

The square of the field brings six powers of $1 - \hat{\boldsymbol{r}}\cdot\boldsymbol{\beta}$ into the denominator, and the factor between the power emitted and the power received takes one of them back. For a fast charge the denominator is small in the directions close to its velocity, and the radiation is thrown forward (Figure 27.11), as Sections 27.6 and 27.7 show.

The total power is the integral of (27.16) over all directions. That integral can be done directly, but it is long. Relativity gives the result in a few lines. Consider

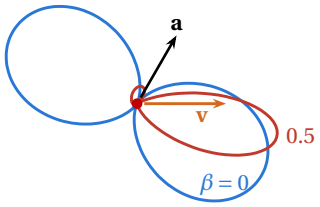


Figure 27.11 The power per unit solid angle, (27.16), radiated by a charge whose acceleration **a** (black) is at 60° to its velocity **v** (orange), in the plane of **v** and **a**: at $\beta = 0$ (blue), Larmor's pattern, and at $\beta = 0.5$ (red). Each curve is scaled to the same maximum. At the higher speed the radiation is thrown forward.

[3] "High precision linewidth measurement of laser-cooled atoms: Resolution of the Na $3p^2P_{3/2}$ lifetime discrepancy," C. W. Oates, K. R. Vogel, and J. L. Hall, Phys. Rev. Lett. 76, 2866–2869 (1996), https://doi.org/10.1103/PhysRevLett.76.2866.

the instantaneous rest frame of the charge at the retarded time (Section 25.9), the inertial frame that moves with the velocity $\mathbf{v}$ that the charge has at that instant. In that frame the charge is momentarily at rest, so Larmor's formula holds exactly, with the proper acceleration $\mathbf{a}_{\text{proper}}$, the acceleration measured in that frame:

$$P_{\text{IRF}} = \frac{\mu_0 q^2 a_{\text{proper}}^2}{6\pi c}.$$

During a short time dt_{IRF} the charge emits the energy $dW_{\text{IRF}} = P_{\text{IRF}}\, dt_{\text{IRF}}$. Radiation carries momentum $1/c$ times its energy in the direction in which it travels (light is massless, $E = pc$, Section 25.10), and the pattern $\sin^2\theta$ sends as much energy in each direction as in the opposite one, so in the instantaneous rest frame the radiation carries away no momentum at all: $d\mathbf{p}_{\text{IRF}} = 0$. Energy and momentum together form a four-vector, $(dW/c, d\mathbf{p})$, so in the lab, where the instantaneous rest frame moves at $\mathbf{v}$, the energy radiated is

$$dW = \gamma\left(dW_{\text{IRF}} + \mathbf{v}\cdot d\mathbf{p}_{\text{IRF}}\right) = \gamma\, dW_{\text{IRF}}.$$



Alfred-Marie Liénard (1869–1958, French) Liénard was born in Amiens. He studied at the École polytechnique (the photograph shows him as a student there, in 1887) and at the École des Mines in Paris, and after three years as a mining engineer he taught at the École des Mines in Saint-Étienne from 1895 to 1908, and then electrical engineering at the École des Mines in Paris. In 1898 he found the potentials of a point charge in arbitrary motion, and from them the power it radiates at any speed. His later work included a test for the stability of linear systems (1914, with Chipart) and an equation for self-sustained oscillations (1928), both named for him. He served in the French Army in World War I and was president of the Société Mathématique de France. (Wikipedia)

In the instantaneous rest frame the emission happens at one place, the charge's position, so in the lab it takes the time $dt = \gamma\, dt_{\text{IRF}}$, dilated by the same factor. The two factors of γ cancel: $P = dW/dt = P_{\text{IRF}}$. The power that a charge radiates is the same in every inertial frame. With $a_{\text{proper}}^2 = \alpha_\mu \alpha^\mu$ from (25.26), with $\mathbf{u}$ replaced by $\mathbf{v}$,

$$\boxed{P = \frac{\mu_0 q^2}{6\pi c}\,\alpha_\mu\alpha^\mu = \frac{\mu_0 q^2 \gamma^6}{6\pi c}\left(a^2 - \frac{|\mathbf{v}\times\mathbf{a}|^2}{c^2}\right)}. \tag{27.17}$$

This is **Liénard's formula**, which Liénard found in 1898, in the same paper as the potentials, by integrating over all directions. Split the acceleration into the parts $\mathbf{a}_\parallel$ along the velocity and $\mathbf{a}_\perp$ across it; then $|\mathbf{v}\times\mathbf{a}| = va_\perp$, and (27.17) becomes

$$P = \frac{\mu_0 q^2}{6\pi c}\left(\gamma^6 a_\parallel^2 + \gamma^4 a_\perp^2\right),$$

in agreement with (25.27), by which the parts of the proper acceleration along and across the velocity are $\gamma^3 a_\parallel$ and $\gamma^2 a_\perp$.

Two cases matter most. When the acceleration is along the velocity, as in a linear accelerator or for a charge that is stopped, $\mathbf{v}\times\mathbf{a} = 0$, and the power is γ^6 times Larmor's. When the acceleration is perpendicular to the velocity, as for a charge moving in a circle at constant speed, $|\mathbf{v}\times\mathbf{a}| = va$, and the power is γ^4 times Larmor's. At low speeds both reduce to Larmor's formula; at high speeds the factors are enormous. An electron with $\gamma = 1000$, an energy of about 0.5 GeV, radiates 10^{12} times the power that Larmor's formula would give for the same acceleration across its path.

The factors of γ may make it look as though relativity has been added to Larmor's formula as a correction. It has not. Liénard found (27.17) in 1898, seven years before Einstein's paper on relativity, from Maxwell's equations alone. Maxwell's equations were consistent with relativity from the start (Chapter 26), and the argument above only uses that fact to save work.

27.6 Synchrotron radiation

A charge that moves in a circle of radius R at constant speed has the acceleration $a = v^2/R$, directed toward the center, perpendicular to its velocity. The radiation of such a charge is called **synchrotron radiation**, after the circular accelerators in which it was first seen.[4] By Liénard's formula (27.17), with $|\mathbf{v}\times\mathbf{a}| = va$ and $v = \beta c$, it radiates the power

$$P = \frac{\mu_0 q^2 \gamma^4 a^2}{6\pi c} = \frac{q^2 c \beta^4 \gamma^4}{6\pi\varepsilon_0 R^2}, \tag{27.18}$$

where the second form uses $a = \beta^2 c^2/R$ and $\mu_0 c^2 = 1/\varepsilon_0$. For a particle of a given energy $E = \gamma mc^2$, the factor γ^4 is proportional to $1/m^4$. An electron and a proton with the same energy, on the same circle, have nearly the same speed and the same acceleration if both are fast, but the electron radiates $(m_\text{p}/m_\text{e})^4 = 1.1\times10^{13}$ times as much power. Electrons in a circular accelerator therefore lose energy to radiation on every turn, and the more so the higher their energy, which is why the highest energies in circular accelerators are reached with protons (Exercise P27.19).

The pattern of the radiation is easiest to see in the charge's instantaneous rest frame, where Larmor's pattern holds: a doughnut around the acceleration, which points toward the center of the circle. Since the acceleration is perpendicular to the velocity, half of the doughnut points forward and half backward (Figure 27.12a). In the lab, aberration (25.14) crowds the half that points forward into a cone around the velocity whose half-angle θ has $\sin\theta = 1/\gamma$: the headlight effect of Section 25.6. The radiation that leaves backward in the instantaneous rest frame is spread over the rest of the sphere, and it carries much less energy, since the transformation lowers the energy of that radiation and raises the energy of the radiation in the forward cone. The exact pattern in the lab follows from (27.16) with $\mathbf{a}$ perpendicular to $\mathbf{v}$. With θ the angle between $\hat{\boldsymbol{\imath}}$ and $\mathbf{v}$, and ϕ the angle around $\mathbf{v}$ measured from the direction of $\mathbf{a}$,

$$\frac{dP}{d\Omega} = \frac{\mu_0 q^2 a^2}{16\pi^2 c}\,\frac{\big(1-\beta\cos\theta\big)^2 - \big(1-\beta^2\big)\sin^2\theta\cos^2\phi}{\big(1-\beta\cos\theta\big)^5} \tag{27.19}$$

(Exercise P27.21). When γ is large, nearly all of the power goes into directions within an angle of about $1/\gamma$ of the velocity (Figure 27.12b).

A charge circling at a high speed thus sends its radiation forward in a narrow beam, tangent to its orbit, which sweeps around with it like the beam of a headlight on a car going around a curve. An observer in the plane of the orbit sees a brief flash each time the beam sweeps past, at the moments when, in Example 24.4, the shadow raced through the part of the orbit where the charge moves toward the observer. The flash is very short. The observer receives radiation only

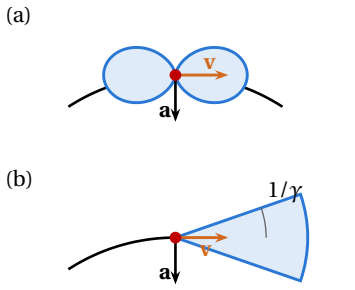


Figure 27.12 A charge moving in a circle (black) at a high speed, with its velocity $\mathbf{v}$ (orange) and its acceleration $\mathbf{a}$ (black). (a) In its instantaneous rest frame, it radiates Larmor's pattern around $\mathbf{a}$, as much backward as forward. (b) In the lab, most of the radiation goes forward, into a narrow cone of half-angle about $1/\gamma$ around $\mathbf{v}$ (here $\gamma = 3$).

[4] It was first seen in 1947, as a spot of visible light from the electrons in General Electric's 70 MeV synchrotron in Schenectady, New York: "Radiation from electrons in a synchrotron," F. R. Elder, A. M. Gurewitsch, R. V. Langmuir, and H. C. Pollock, Phys. Rev. 71, 829–830 (1947), https://doi.org/10.1103/PhysRev.71.829.5.

from the part of the orbit where the velocity points within about $1/\gamma$ of the line of sight, and the charge turns through that angle range, $2/\gamma$, in the time $\Delta t_r \approx 2R/\gamma c$. During that time it moves almost straight toward the observer at almost the speed of light, so by (27.3) the news of that stretch arrives in the much shorter time $(1-\beta)\,\Delta t_r$. Since $1-\beta = 1/[\gamma^2(1+\beta)] \approx 1/2\gamma^2$, the flash lasts about

$$\Delta t \approx \frac{R}{\gamma^3 c}.$$

The charge goes around once in the time $2\pi R/v \approx 2\pi R/c$, but the flash is shorter by a factor of order γ^3, and a pulse that short contains frequencies up to about $1/\Delta t$, which is γ^3 times the frequency of the orbit. A full calculation of the spectrum shows that half of the power is radiated at angular frequencies below the **critical frequency**

$$\omega_c = \frac{3}{2}\frac{\gamma^3 c}{R} \tag{27.20}$$

and half above it, and that far above ω_c the spectrum dies away exponentially.[5] For electrons of a few GeV on circles a few meters in radius, $\hbar\omega_c$ is a few keV, so the radiation reaches into the X-rays. Storage rings built for this purpose, synchrotron light sources, supply intense, tightly collimated X-rays for experiments in physics, chemistry, materials science, and biology. Many of them also send their electrons through undulators, rows of magnets of alternating polarity that make the electrons wiggle from side to side, so that the beams sent forward from the successive wiggles add up.

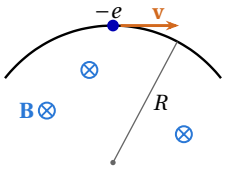
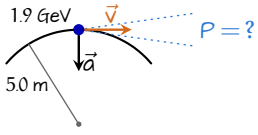


Figure 27.13 Example 27.4: an electron with an energy of 1.9 GeV moving on an arc of radius 5.0 m in a bending magnet, whose field points into the page.

Example 27.4

The Advanced Light Source, a synchrotron light source in Berkeley, California, stores electrons with an energy of 1.9 GeV. In its bending magnets the electrons move on arcs of circles of radius $R = 5.0$ m. (a) Find the power that one electron radiates there, and compare it with Larmor's formula. (b) Find the critical frequency and the corresponding photon energy $\hbar\omega_c$.

Solution:



The first step is a sketch of the electron on its arc, with its velocity along the arc and its acceleration toward the center. They are perpendicular, so the radiation goes forward in a narrow beam.

(a) The electron's rest energy is 0.511 MeV, so its Lorentz factor is

$$\gamma = \frac{E}{mc^2} = \frac{1.9\,\text{GeV}}{0.511\,\text{MeV}} = 3718,$$

[5] The spectrum was worked out in "On the classical radiation of accelerated electrons," J. Schwinger, Phys. Rev. 75, 1912–1925 (1949), https://doi.org/10.1103/PhysRev.75.1912.

and $1-\beta \approx 1/2\gamma^2 = 3.6\times10^{-8}$: the electron's speed differs from c by less than one part in 10^7. Its acceleration is

$$a = \frac{v^2}{R} \approx \frac{c^2}{R} = \frac{(3.00\times10^8\,\mathrm{m/s})^2}{5.0\,\mathrm{m}} = 1.80\times10^{16}\,\mathrm{m/s^2}.$$

The acceleration is perpendicular to the velocity, so by (27.18)

$$P = \frac{\mu_0 e^2\gamma^4 a^2}{6\pi c} = 3.5\times10^{-7}\,\mathrm{W}.$$

Larmor's formula, without the factor $\gamma^4 = 1.9\times10^{14}$, would give only 1.8×10^{-21} W. The bending magnets turn the electrons through a full circle on each trip around the ring, so each electron loses $2\pi RP/v = 230\,\mathrm{keV}$ per turn, about one part in 8000 of its energy, which the ring must give back to it on every turn.

(b) By (27.20),

$$\omega_c = \frac{3}{2}\frac{\gamma^3 c}{R} = 4.6\times10^{18}\,\mathrm{s^{-1}}, \qquad \hbar\omega_c = 3.0\,\mathrm{keV},$$

X-rays with wavelengths near $2\pi c/\omega_c = 0.41\,\mathrm{nm}$, comparable to the spacing of atoms in a crystal. The electron would go around a circle of this radius at the angular frequency $c/R = 6.0\times10^7\,\mathrm{s^{-1}}$, and ω_c is higher by the factor $\frac{3}{2}\gamma^3 = 8\times10^{10}$. Check: the radius agrees with the field of the bending magnets, 1.27 T, since by (25.39) an electron with the momentum $p \approx E/c$ moves in that field on a circle of radius $R = p/eB = 5.0\,\mathrm{m}$.[6]

27.7 Bremsstrahlung

A charge that is slowed down radiates, and the radiation is called **bremsstrahlung**, German for "braking radiation." The electrons that strike the target of an X-ray tube (Figure 27.14) are slowed and deflected in collisions with its atoms, and the continuous part of the X-ray spectrum is their bremsstrahlung. The name is used for the radiation of any collision, whether the charge is slowed, sped up, or turned.

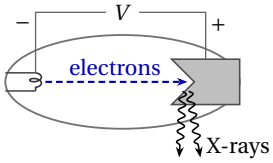


Figure 27.14 An X-ray tube. Electrons from a hot filament are accelerated through the potential difference V and strike a metal target, where they are slowed and deflected, and emit X-rays. Not to scale.

When the acceleration is along the velocity, either way, $\boldsymbol{\beta}\times\mathbf{a} = 0$, so $(\hat{\boldsymbol{r}}-\boldsymbol{\beta})\times\mathbf{a} = \hat{\boldsymbol{r}}\times\mathbf{a}$, and $|\hat{\boldsymbol{r}}\times(\hat{\boldsymbol{r}}\times\mathbf{a})| = a\sin\theta$, where θ is the angle between $\hat{\boldsymbol{r}}$ and the line of motion. By (27.16), the charge emits

$$\frac{dP}{d\Omega} = \frac{\mu_0 q^2 a^2}{16\pi^2 c}\,\frac{\sin^2\theta}{(1-\beta\cos\theta)^5}, \tag{27.21}$$

with θ measured from the velocity $\mathbf{v}$, whether $\mathbf{a}$ points along $\mathbf{v}$ or against it. At low speed this is Larmor's doughnut around the line of motion. As the speed grows,

[6] "Suite of three protein crystallography beamlines with single superconducting bend magnet as the source," A. A. MacDowell et al., J. Synchrotron Rad. 11, 447–455 (2004), https://doi.org/10.1107/S0909049504024835, which gives the ring's energy, 1.9 GeV, the field of its bending magnets, 1.27 T, and their critical photon energy, 3 keV.

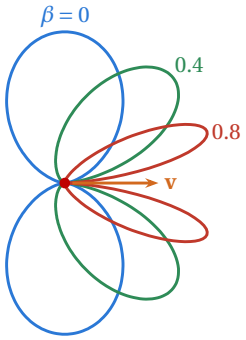


Figure 27.15 The power per unit solid angle radiated by a charge whose acceleration is along its velocity **v** (orange), (27.21), in a plane that contains **v**, at $\beta = 0$, 0.4, and 0.8. Each curve is scaled to the same maximum. As β grows, the lobes tip forward, but no radiation goes straight ahead.



Hendrik Kramers (1894–1952, Dutch) Hendrik Anthony Kramers, known as Hans, studied at Leiden and in 1916 went to Copenhagen to work with Niels Bohr, whose first assistant he became in 1919, the year of his doctorate. His theory of optical dispersion was a precursor of Heisenberg's matrix mechanics. He became professor of theoretical physics at Utrecht in 1926 and succeeded Paul Ehrenfest at Leiden in 1934. After the Second World War he chaired the technical subcommittee of the United Nations Atomic Energy Commission. (Wikipedia)

the denominator tips the two lobes of the pattern forward (Figure 27.15), so that a fast charge sends most of its radiation into a cone around its direction of motion, but none straight ahead, where $\sin\theta = 0$ (Exercise P27.22). The total power is γ^6 times Larmor's, by Liénard's formula (Exercise P27.23).

27.7.1 The spectrum

Larmor's formula gives the power at each instant, but not how the energy is shared among frequencies, which is what a spectrometer measures. The best-known bremsstrahlung spectrum is the continuous spectrum of an X-ray tube, in which electrons accelerated through a potential difference V, typically tens of kilovolts, strike a metal target. Hendrik Kramers explained its shape in 1923, in three steps.[7]

The first step is a single collision. An electron that passes a nucleus is deflected, and its acceleration is a short pulse, lasting about the time τ_c that the electron takes to go past. At low speed the radiation field at each point of a distant sphere is (27.9), proportional to the acceleration at the retarded time, so the radiation is a pulse with the same time dependence. To find how its energy is shared among frequencies, one writes the pulse as a superposition of sinusoidal oscillations of all frequencies. This is its Fourier transform, the version of the Fourier series of Chapter 6 for a function that is not periodic. A result called Parseval's theorem then turns Larmor's formula, which gives the energy radiated as an integral of a^2 over time, into an integral over frequency, and so gives the energy radiated per unit frequency. The outcome is easy to describe. At frequencies well below $1/\tau_c$ the oscillations hardly change during the pulse, and the spectrum is flat, with a height set by the change of the electron's velocity alone, not by the details of the collision. At frequencies well above $1/\tau_c$ the contributions of the different parts of the pulse cancel, and the spectrum falls off. A short collision therefore radiates over a broad band of frequencies, up to about $1/\tau_c$.

The second step adds up the collisions in a thin target, one so thin that the electrons keep their speed. The electrons pass the nuclei at all distances. Close passes are short and sharp, and they radiate up to high frequencies; distant passes are long and gentle, and they radiate only low frequencies. Kramers added up the passes at all distances and found that a thin target radiates equal energy in equal intervals of frequency, up to a highest frequency. That highest frequency does not come from classical physics, in which a close enough pass radiates at any frequency at all. But the spectrum of an X-ray tube does have a sharp end: no X-rays come out with frequencies above eV/h, or wavelengths shorter than

$$\lambda_{\text{min}} = \frac{hc}{eV},$$

0.0413 nm for $V = 30$ kV. This limit was found in 1915.[8] In quantum theory, light

[7] "On the theory of X-ray absorption and of the continuous X-ray spectrum," H. A. Kramers, Phil. Mag. 46, 836–871 (1923), https://doi.org/10.1080/14786442308565244.

[8] "On X-ray wave-lengths," W. Duane and F. L. Hunt, Phys. Rev. 6, 166–172 (1915), https://doi.org/10.1103/PhysRev.6.166.

is emitted in photons of energy $h\nu$, and no photon can carry more energy than the electron has, eV. Kramers took the classical spectrum to hold up to that frequency and to end there. (At frequencies where $h\nu$ is much smaller than the electron's energy, quantum theory and the classical spectrum agree.)

The third step is the thick target of a real tube, in which the electrons slow down as they go deeper, losing most of their energy to the atoms' electrons, not to radiation. A slow electron radiates more strongly than a fast one per unit length of its path, but it also loses its energy faster, and the two effects cancel: with the measured rate at which electrons slow down in matter, each equal amount of energy that an electron loses gives the same flat spectrum, up to the frequency set by the energy that it still has. These flat spectra end at lower and lower frequencies as the electron slows, and together they make a spectrum that falls linearly to zero at the highest frequency. The derivation is beyond the scope of this book, but the result, **Kramers' law**, is simple:

$$I(\nu) = KZ\,(\nu_{\max} - \nu) \qquad (\nu < \nu_{\max}), \qquad \nu_{\max} = \frac{eV}{h}, \tag{27.22}$$

where $I(\nu)\,d\nu$ is the power radiated as X-rays with frequencies between ν and $\nu + d\nu$, Z is the atomic number of the target, and K is a constant proportional to the current of electrons (Figure 27.16). Plotted against wavelength, as X-ray spectra often are, the same spectrum rises from zero at $\lambda_{\min}$ to a maximum and then falls off slowly. Measured spectra also contain the target's characteristic lines, which come from transitions between the inner electron shells of its atoms, and they lose part of their low-frequency end to absorption in the target and in the window of the tube, effects that Kramers' law leaves out.

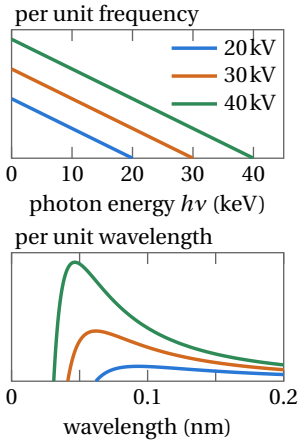


Figure 27.16 Kramers' law (27.22) for a tube at 20, 30, and 40 kV, in arbitrary units: the power per unit frequency against the photon energy (top), and the same spectra per unit wavelength (bottom), each ending sharply at $\lambda_{\min} = hc/eV$.

27.8 Numerical method: fields of an arbitrarily moving charge

The fields (27.6) hold for any motion, but for most motions only a computer can evaluate them. Two tasks come up: finding the fields at given points, and drawing the field lines.

The fields at a point. For a charge whose path $\mathbf{w}(t')$ is known, the field at the point $\mathbf{r}$ and the time t takes two steps. First find the retarded time, the root of $c(t - t') = |\mathbf{r} - \mathbf{w}(t')|$, by the bracketing method of Section 24.6.1. Then evaluate $\boldsymbol{\mathscr{r}} = \mathbf{r} - \mathbf{w}(t_r)$, $\mathbf{v}(t_r)$, and $\mathbf{a}(t_r)$, and put them into (27.6). Repeated on a grid of points, this gives a map of the fields, or of the Poynting vector, at the time t.

Field lines. A field line is a curve that is tangent to $\mathbf{E}$ at each of its points (Section 3.3). Traced step by step from the charge, a field line of a moving charge would need a new retarded time at every step. There is a better way: label the points of a field line by their retarded times.[9] At the time t, the point of the line

[9] This method is from "Electric field lines of relativistically moving point charges," D. Ruhlandt, S. Mühle, and J. Enderlein, Am. J. Phys. 88, 5–10 (2020), https://doi.org/10.1119/10.0000189.

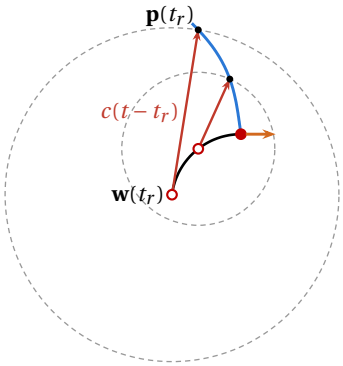


Figure 27.17 The points of a field line (blue) at the time t, for a charge moving on a curved path (black), now at the solid dot. The point $\mathbf{p}(t_r)$ whose retarded time is t_r lies on the circle (dashed) that the news sent from $\mathbf{w}(t_r)$ (open circle) has reached by the time t, a distance $c(t-t_r)$ (red) from $\mathbf{w}(t_r)$.

whose retarded time is t_r lies on the sphere that the news sent at t_r has reached, a distance $c(t-t_r)$ from $\mathbf{w}(t_r)$, in some direction $\hat{\boldsymbol{\imath}}(t_r)$ (Figure 27.17):

$$\mathbf{p}(t_r) = \mathbf{w}(t_r) + c\,(t-t_r)\hat{\boldsymbol{\imath}}(t_r). \tag{27.23}$$

For this curve to be a field line, its tangent $d\mathbf{p}/dt_r$ must be parallel to $\mathbf{E}$ at each of its points. With $\imath = c(t-t_r)$, the tangent is

$$\frac{d\mathbf{p}}{dt_r} = \mathbf{v} - c\hat{\boldsymbol{\imath}} + \imath\frac{d\hat{\boldsymbol{\imath}}}{dt_r},$$

and by (27.6), the field at $\mathbf{p}$ is a positive multiple of $(\hat{\boldsymbol{\imath}}-\boldsymbol{\beta}) + (\gamma^2\imath/c^2)\hat{\boldsymbol{\imath}}\times[(\hat{\boldsymbol{\imath}}-\boldsymbol{\beta})\times\mathbf{a}]$ (multiply the bracket in (27.6) by $\gamma^2\imath^2 = \imath^2/(1-\beta^2)$). The tangent is $-c$ times this vector, and so parallel to $\mathbf{E}$, if

$$\frac{d\hat{\boldsymbol{\imath}}}{dt_r} = \frac{\gamma^2}{c}\left[\left(\hat{\boldsymbol{\imath}}-\boldsymbol{\beta}\right)\times\mathbf{a}\right]\times\hat{\boldsymbol{\imath}}, \tag{27.24}$$

with γ, $\boldsymbol{\beta}$, and $\mathbf{a}$ evaluated at t_r. Since the factor $-c$ is negative, the point $\mathbf{p}$ moves along $\mathbf{E}$, outward from a positive charge, as t_r decreases.

Equation (27.24) is an ordinary differential equation for the unit vector $\hat{\boldsymbol{\imath}}(t_r)$, to be solved with a standard method, such as the fourth-order Runge–Kutta method. No retarded time has to be found, since the retarded time is the variable. The right side is perpendicular to $\hat{\boldsymbol{\imath}}$, so $\hat{\boldsymbol{\imath}}$ remains a unit vector (dividing it by its length after each step removes the roundoff). Where the charge does not accelerate, $\hat{\boldsymbol{\imath}}$ does not change, and the line is straight. And the equation does not contain t: one solution $\hat{\boldsymbol{\imath}}(t_r)$ gives the field line at every time t, through (27.23). As t increases, each point of the line moves straight away from its own $\mathbf{w}(t_r)$ at the speed of light, so the field line is carried outward by the news.

The equation needs a starting value of $\hat{\boldsymbol{\imath}}$ for each line. For a charge that was at rest before some time, start the lines then, at directions spaced evenly in angle, as in the book's other drawings of field lines. For a charge that was moving with constant velocity, a line that leaves the charge at the angle θ' from $\mathbf{v}$ in the charge's rest frame starts with $\hat{\boldsymbol{\imath}}$ at the angle θ given by the aberration formula (25.14), with $\theta_R = \theta'$ and $\theta_L = \theta$ (Exercise P27.26).

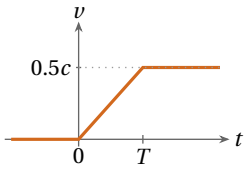
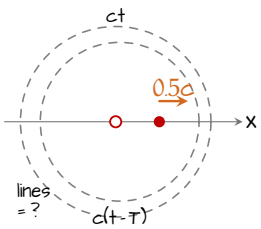


Figure 27.18 Example 27.5: the speed of the charge, which accelerates uniformly from rest to $0.5c$ in the time T and then keeps that speed.

Example 27.5

A point charge q is at rest at the origin until $t = 0$. It then accelerates along $+x$ with constant acceleration for a time T, until its speed is $0.5c$, and from then on it moves with that constant velocity. Draw its electric field lines at the time $t = 6T$.

Solution:



The first step is a sketch of where the charge was and where it is now, and of how far the news of the start and of the end of the acceleration has spread.

Use units in which $c = 1$ and $T = 1$. During the acceleration, $a = 0.5$, $v = 0.5t'$, and $x = 0.25t'^2$, so the charge reaches $x = 0.25$ at $t' = 1$, and at $t = 6$ it is at $x = 0.25 + 0.5(6-1) = 2.75$. The news of the start of the acceleration has spread to the sphere of radius 6 about the origin, and the news of its end to the sphere of radius 5 about $x = 0.25$.

Before $t' = 0$ the charge was at rest, so start twelve lines in the xy plane at $t_r = 0$, at the angles $15°, 45°, \ldots, 345°$ from the x axis. During the acceleration, $0 < t_r < 1$, integrate (27.24) for each line with the fourth-order Runge-Kutta method in 200 steps. Before and after the acceleration $\hat{\mathbf{r}}$ is constant. Then each line consists of the points (27.23) with $t = 6$, for t_r from the past up to 6, where the line reaches the charge.

Figure 27.19 shows the result. Outside the sphere of radius 6 the news of the acceleration has not arrived, and the lines are those of a charge at rest at the origin. Inside the sphere of radius 5 they are those of a charge moving uniformly at $0.5c$: straight lines from its present position, crowded somewhat toward the plane perpendicular to the motion, as in Section 26.7. In the shell between, each line kinks, running nearly along the shell and pointing back, against the acceleration: Thomson's kink of Section 27.2.2, now at a speed where his low-speed picture no longer applies. In the middle of the shell, at the point $(0, 5.5)$, the field is about three times as strong as the Coulomb field of a charge at the origin would be there, and it points mostly backward.

Two checks. The directions $\hat{\mathbf{r}}$ agree with the exact solution of Exercise P27.26 to 10^{-13} radians. And the fields, found point by point with (27.6) and integrated over spheres of radius 5.5 about the origin (which lies inside the shell), 7 about the origin, and 2 about the charge, give the flux q/ε_0 through each, to ten digits, as Gauss's law requires: the kinked lines carry all of the charge's flux.

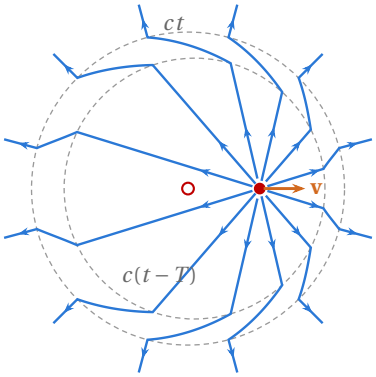


Figure 27.19 The field lines of Example 27.5 at the time $t = 6T$, computed with (27.24). The charge (solid dot) was at rest at the origin (open circle) until $t = 0$ and has moved at $0.5c$ (orange) since $t = T$. Outside the circle of radius ct the lines come from the origin, inside the circle of radius $c(t-T)$ from the charge's present position, and in the shell between they kink.

27.8.1 Tips for Mathematica and Python

- Write the path $\mathbf{w}(t')$, the velocity, and the acceleration as functions. If the acceleration jumps, as in Example 27.5, integrate each smooth stretch separately, with the jumps at the ends of the stretches, so that no step straddles a jump.

- The right side of (27.24) is two cross products: `numpy.cross` in Python, `Cross` in Mathematica. A general solver also works, such as `solve_ivp` in `scipy.integrate` or `NDSolve` in Mathematica.
- To draw the lines in a plane, integrate in three dimensions anyway. Lines that start in a plane that contains the whole motion stay in it.
- For the fields at a point, use the root finder of Section 24.6.1, and check the code on a charge in uniform motion, whose fields (26.23) are known.

27.9 Summary

A point charge in any motion has fields that depend on its position, velocity, and acceleration at the retarded time. The part that depends only on the velocity falls off as $1/r^2$ and travels along with the charge; the part proportional to the acceleration falls off as $1/r$ and carries energy away for good. That energy is radiation, and how much a charge radiates, in which directions, and at which frequencies follows from the acceleration field.

The potentials and the fields

$A^\mu = \dfrac{q}{4\pi\varepsilon_0 c}\dfrac{\eta^\mu}{\xi}, \quad \xi = \gamma\,(cr - \boldsymbol{r}\cdot\mathbf{v})$	the four-potential of a point charge; $\boldsymbol{r}$ runs from the shadow to the field point
$V = \dfrac{1}{4\pi\varepsilon_0}\dfrac{qc}{cr - \boldsymbol{r}\cdot\mathbf{v}}, \quad \mathbf{A} = \dfrac{\mathbf{v}}{c^2}V$	the Liénard–Wiechert potentials; $\mathbf{v}$ is the charge's actual velocity at the retarded time
$\dfrac{\partial t_r}{\partial t} = \dfrac{1}{1-\hat{\boldsymbol{r}}\cdot\mathbf{v}/c}$	the shadow moves at $\mathbf{v}/(1-\hat{\boldsymbol{r}}\cdot\mathbf{v}/c)$, not at $\mathbf{v}$
$\mathbf{E} = \dfrac{q}{4\pi\varepsilon_0\left(1-\hat{\boldsymbol{r}}\cdot\boldsymbol{\beta}\right)^3}\left[\dfrac{\left(1-\beta^2\right)\left(\hat{\boldsymbol{r}}-\boldsymbol{\beta}\right)}{r^2} + \dfrac{\hat{\boldsymbol{r}}\times\left[\left(\hat{\boldsymbol{r}}-\boldsymbol{\beta}\right)\times\mathbf{a}\right]}{c^2 r}\right]$	the velocity field ($1/r^2$) and the acceleration field ($1/r$); everything at the retarded time
$\mathbf{B} = \hat{\boldsymbol{r}}\times\mathbf{E}/c$	perpendicular to $\mathbf{E}$ and to $\hat{\boldsymbol{r}}$
$\mathbf{w}_\mathrm{p} = \mathbf{w}(t_r) + \mathbf{v}(t-t_r)$	the projected position, from which the velocity field points

Radiation

$\mathbf{S}_{\text{rad}} = \dfrac{1}{\mu_0 c} E_{\text{rad}}^2 \hat{\boldsymbol{\mathscr{r}}}$ — only the $1/\mathscr{r}$ field carries energy to infinity

$d\Omega = \sin\theta \, d\theta \, d\phi$ — solid angle; the whole sphere is 4π sr

$\dfrac{dP}{d\Omega} = \left(1 - \dfrac{\hat{\boldsymbol{\mathscr{r}}} \cdot \mathbf{v}}{c}\right) \dfrac{E_{\text{rad}}^2 \mathscr{r}^2}{\mu_0 c}$ — the power emitted per unit solid angle, per unit of the charge's time

$\dfrac{dP}{d\Omega} = \dfrac{\mu_0 q^2 a^2}{16\pi^2 c} \sin^2\theta, \quad P = \dfrac{\mu_0 q^2 a^2}{6\pi c}$ — Larmor's formula, for a slowly moving charge; θ from $\mathbf{a}$

$\Gamma = \omega^2 \tau_{\text{rad}}, \quad \tau_{\text{rad}} = \dfrac{\mu_0 q^2}{6\pi m c}$ — the rate at which the energy of a radiating charged oscillator decays

$P = \dfrac{\mu_0 q^2}{6\pi c} \alpha_\mu \alpha^\mu =$
$\dfrac{\mu_0 q^2 \gamma^6}{6\pi c} \left(a^2 - \dfrac{|\mathbf{v} \times \mathbf{a}|^2}{c^2}\right)$ — Liénard's formula, for any speed; the same in every inertial frame

$\dfrac{dP}{d\Omega} = \dfrac{\mu_0 q^2}{16\pi^2 c} \dfrac{|\hat{\boldsymbol{\mathscr{r}}} \times [(\hat{\boldsymbol{\mathscr{r}}} - \boldsymbol{\beta}) \times \mathbf{a}]|^2}{(1 - \hat{\boldsymbol{\mathscr{r}}} \cdot \boldsymbol{\beta})^5}$ — the pattern at any speed, thrown forward when β is close to 1

Synchrotron radiation and bremsstrahlung

$P = \dfrac{q^2 c \beta^4 \gamma^4}{6\pi\varepsilon_0 R^2}$ — a charge on a circle of radius R; at a given energy, proportional to $1/m^4$

$\theta \approx 1/\gamma, \quad \Delta t \approx \dfrac{R}{\gamma^3 c}, \quad \omega_{\text{c}} = \dfrac{3}{2} \dfrac{\gamma^3 c}{R}$ — the forward beam, the flash an observer receives, and the critical frequency

$\dfrac{dP}{d\Omega} = \dfrac{\mu_0 q^2 a^2}{16\pi^2 c} \dfrac{\sin^2\theta}{(1 - \beta\cos\theta)^5}$ — acceleration along the velocity; θ from $\mathbf{v}$

ω up to about $1/\tau_{\text{c}}$ — the band radiated by one collision that lasts the time τ_{c}, with a nearly flat spectrum

$I(\nu) = KZ(\nu_{\max} - \nu), \quad \nu_{\max} = \dfrac{eV}{h}$ — Kramers' law: the continuous spectrum of an X-ray tube, falling linearly to zero at $\nu_{\max}$

$\lambda_{\min} = \dfrac{hc}{eV}$ — the shortest wavelength from an X-ray tube: a limit set by photons, not by classical physics

Numerical method

$c(t-t_r) = |\mathbf{r} - \mathbf{w}(t_r)|$, then (27.6) — the fields at a point: the retarded time by bracketing

$\mathbf{p}(t_r) = \mathbf{w}(t_r) + c\,(t-t_r)\hat{\boldsymbol{\varepsilon}}(t_r)$ — the points of a field line at the time t, labeled by their retarded times

$$\frac{d\hat{\boldsymbol{\varepsilon}}}{dt_r} = \frac{\gamma^2}{c}\left[\left(\hat{\boldsymbol{\varepsilon}} - \boldsymbol{\beta}\right) \times \mathbf{a}\right] \times \hat{\boldsymbol{\varepsilon}}$$

the equation for the field line's direction from the shadow; the same for every t

Exercises

Exercises for 27.1 The Liénard–Wiechert potentials

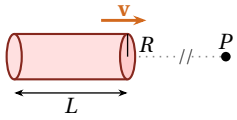


Figure 27.20 The charged cylinder of Exercise P27.1, moving along its axis toward a distant observer at P.

P27.1 The apparent amount of charge. A uniformly charged cylinder of radius R and length L, with the charge density ρ_0, moves at constant speed v along its axis, straight toward a distant observer (Figure 27.20). The retarded potential (24.11) adds up the charge of each volume element at that element's own retarded time.

(a) Use the argument of Section 24.3.1 to show that the charge it counts, $\int \rho(\mathbf{r}', t_r)\, d\tau'$, is $\rho_0 \pi R^2 L/(1 - v/c)$, more than the cylinder's charge, and that for a cylinder moving away from the observer it is $\rho_0 \pi R^2 L/(1 + v/c)$. Take the observer to be far away compared with L and R.

(b) For a small charge q moving with the velocity $\mathbf{v}$, the same argument gives $\int \rho(\mathbf{r}', t_r)\, d\tau' = q/(1 - \hat{\boldsymbol{\imath}} \cdot \mathbf{v}/c)$. Use this to obtain V of (27.2) from (24.11), treating $\imath$ as the same for every part of the small charge.

(c) A second route uses delta functions. Write the density of a point charge as $\rho(\mathbf{r}', t') = q\,\delta^3(\mathbf{r}' - \mathbf{w}(t'))$, and the retarded potential as

$$V(\mathbf{r}, t) = \frac{1}{4\pi\varepsilon_0} \iint \frac{\rho(\mathbf{r}', t')\, \delta\big(t' - t + |\mathbf{r} - \mathbf{r}'|/c\big)}{|\mathbf{r} - \mathbf{r}'|}\, d\tau'\, dt'.$$

Do the integral over $\mathbf{r}'$ first, then use (26.21) for the integral over t', and show that the result is V of (27.2).

P27.2 The wrong velocity. Suppose that the velocity at which the shadow appears to move, $\mathbf{v}/(1 - \hat{\boldsymbol{\imath}} \cdot \mathbf{v}/c)$ by (27.3), were put in place of $\mathbf{v}$ in the Liénard–Wiechert potential V of (27.2).

(a) Show that this would give $V_{\text{app}} = (q/4\pi\varepsilon_0)(1 - \hat{\boldsymbol{\imath}} \cdot \boldsymbol{\beta})/[\imath(1 - 2\hat{\boldsymbol{\imath}} \cdot \boldsymbol{\beta})]$, with $\boldsymbol{\beta} = \mathbf{v}/c$.

(b) For a charge moving with constant velocity, take a field point beside the charge's present position, so that the vector from the present position to the field point, of length R, is perpendicular to $\mathbf{v}$. Show that there $\imath = \gamma R$ and $\hat{\boldsymbol{\imath}} \cdot \boldsymbol{\beta} = \beta^2$, find V and V_{app}, and compare them with Example 26.2. Evaluate V_{app}/V at $\beta = 0.5$.

(c) For $\beta > 1/2$, show that V_{app} of a positive charge would be infinite at some field points and negative at others. What does this say about which velocity belongs in the potentials?

Exercises for 27.2 The fields of a point charge

P27.3 The derivatives of the retarded time. The lab time t_r of the shadow event is related to the charge's proper time τ_r there by $dt_r = \gamma\, d\tau_r$.

(a) Use Step 1 of the derivation in Section 27.2, $\partial^\mu \tau_r = -\mathscr{r}^\mu/\xi$, to show that

$$\nabla t_r = -\frac{\boldsymbol{\mathscr{r}}}{c\mathscr{r} - \boldsymbol{\mathscr{r}}\cdot\mathbf{v}},$$

and that the time component gives (27.3).

(b) For a charge at rest, show that $\nabla t_r = -\hat{\boldsymbol{\mathscr{r}}}/c$, and explain the sign: why does moving the field point away from the charge make the retarded time earlier?

(c) Is $|\nabla t_r|$ larger ahead of a moving charge or behind it? Explain why.

P27.4 The field in the instantaneous rest frame. Let the charge be momentarily at rest at the retarded time, as it is in its instantaneous rest frame.

(a) Show that then $\eta^\mu = (c,0,0,0)$, $\alpha^\mu = (0,\mathbf{a})$, $\xi = c\mathscr{r}$, and $\alpha_\lambda \mathscr{r}^\lambda = \mathbf{a}\cdot\boldsymbol{\mathscr{r}}$.

(b) Use (27.5) and Step 4 of the derivation in Section 27.2 to show that

$$\mathbf{E} = \frac{q}{4\pi\varepsilon_0}\left[\frac{\hat{\boldsymbol{\mathscr{r}}}}{\mathscr{r}^2} + \frac{\hat{\boldsymbol{\mathscr{r}}}\times(\hat{\boldsymbol{\mathscr{r}}}\times\mathbf{a})}{c^2\mathscr{r}}\right],$$

Coulomb's field of a charge at the shadow plus the acceleration field (27.9).

(c) Show that (27.6) with $\mathbf{v} = 0$ gives the same.

P27.5 Another route to the fields. The field of Exercise P27.4, in the charge's instantaneous rest frame at the retarded time, can be carried to the lab with the three-vector rules of Chapters 25 and 26. (This route is A. K. Singal's: "A first principles derivation of the electromagnetic fields of a point charge in arbitrary motion," Am. J. Phys. 79, 1036–1041 (2011), https://doi.org/10.1119/1.3620257.) Let the instantaneous rest frame be Rachel's, moving at $\mathbf{v}$ relative to the lab, which is Luke's. Mark quantities in Rachel's frame with a prime, split every vector into parts parallel and perpendicular to $\mathbf{v}$, and let $\boldsymbol{\beta} = \mathbf{v}/c$.

(a) The shadow event and the field event lie on each other's light cone in both frames, so $\mathscr{r}^\mu$ is $(\mathscr{r},\boldsymbol{\mathscr{r}})$ in Luke's frame and $(\mathscr{r}',\boldsymbol{\mathscr{r}}')$ in Rachel's. Use the Lorentz transformation to show that $\mathscr{r}' = \mathscr{r}/\delta$, with $\delta \equiv 1/\gamma(1 - \hat{\boldsymbol{\mathscr{r}}}\cdot\boldsymbol{\beta})$, and that $\hat{\boldsymbol{\mathscr{r}}}'_\parallel = \delta\gamma(\hat{\boldsymbol{\mathscr{r}}}_\parallel - \boldsymbol{\beta})$ and $\hat{\boldsymbol{\mathscr{r}}}'_\perp = \delta\hat{\boldsymbol{\mathscr{r}}}_\perp$. (This is the aberration of (25.14).)

(b) By (25.27), $\mathbf{a}'_\parallel = \gamma^3\mathbf{a}_\parallel$ and $\mathbf{a}'_\perp = \gamma^2\mathbf{a}_\perp$. Show that $\gamma(\hat{\boldsymbol{\mathscr{r}}}'_\parallel + \boldsymbol{\beta}) = \delta\hat{\boldsymbol{\mathscr{r}}}_\parallel$ and that $(\hat{\boldsymbol{\mathscr{r}}}' + \boldsymbol{\beta})\cdot\mathbf{a}' = \delta\gamma^2\hat{\boldsymbol{\mathscr{r}}}\cdot\mathbf{a}$.

(c) In Rachel's frame $\mathbf{B}' = \hat{\boldsymbol{\imath}}' \times \mathbf{E}'/c$. Carry $\mathbf{E}'$ and $\mathbf{B}'$ to Luke's frame with (26.6), with $\mathbf{v}$ replaced by $-\mathbf{v}$, and show, with the identities of (b), that the result is (27.6). (Do the velocity field first, with $\mathbf{a} = 0$. The whole calculation takes about a page.)

P27.6 Check algebraically that the velocity field (27.7) is (26.23) with $\mathbf{R}$ measured from the projected position.

(a) Let $\mathbf{R} = \mathbf{r} - \mathbf{w}_\mathrm{p}$ be the vector from the projected position to the field point. Show that $\mathbf{R} = \imath(\hat{\boldsymbol{\imath}} - \boldsymbol{\beta})$, so that $\mathbf{E}_\mathrm{vel}$ points along $\mathbf{R}$.

(b) The derivation of (27.4) used only $\mathbf{R} = \boldsymbol{\imath} - \boldsymbol{\beta}\imath$. Explain why it therefore holds here too, with θ the angle between $\mathbf{R}$ and the velocity at the retarded time, and show that $\imath^3(1 - \hat{\boldsymbol{\imath}} \cdot \boldsymbol{\beta})^3 = R^3(1 - \beta^2 \sin^2\theta)^{3/2}$.

(c) Combine (a) and (b) to finish.

P27.7 A point charge q moves along the x axis in any way at all (Figure 27.21): it may speed up, slow down, or reverse. Let v be the x component of its velocity at the retarded time, positive or negative. Do not use the fields of a uniformly moving charge: the charge may accelerate.

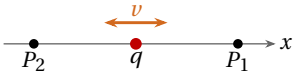


Figure 27.21 The charge of Exercise P27.7, moving along the x axis in any way, and field points P_1 and P_2 on the axis, to its right and to its left.

(a) Show that on the x axis the acceleration field is zero.

(b) Show that at a point on the axis to the right of the shadow, and at a point to the left of it,

$$\mathbf{E} = \frac{q}{4\pi\varepsilon_0 \imath^2}\frac{c+v}{c-v}\hat{\mathbf{x}} \quad \text{(right)}, \qquad \mathbf{E} = -\frac{q}{4\pi\varepsilon_0 \imath^2}\frac{c-v}{c+v}\hat{\mathbf{x}} \quad \text{(left)}.$$

(c) Show that $\mathbf{B} = 0$ everywhere on the axis.

(d) When is the field on the axis stronger than Coulomb's field at the same distance from the shadow, and when is it weaker?

P27.8 Return to the small charged ball of Exercise P24.5, dropped from rest at $t = 0$ from the top of a cliff where you stand (Figure 27.22). Take y upward, and let t_r be the time at which the ball sent the news that reaches you.

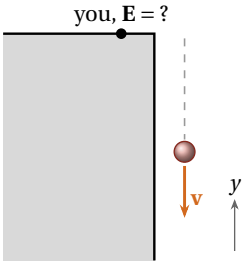


Figure 27.22 The falling ball of Exercise P27.8, and the observer at the top of the cliff, where the field is wanted.

(a) Use Exercise P27.7 to find the electric field at your position in terms of t_r and $\imath$. Does it point up or down, and is it stronger or weaker than Coulomb's field of a charge at the ball's shadow?

(b) Expand your answer to first order in $g\,t_r/c$.

(c) The ball accelerates, yet the acceleration of gravity enters the field only through the velocity $-g\,t_r$. Why?

P27.9 Feynman's formula. Richard Feynman wrote the electric field of a moving point charge as

$$\mathbf{E} = \frac{q}{4\pi\varepsilon_0}\left[\frac{\hat{\boldsymbol{\imath}}}{\imath^2} + \frac{\imath}{c}\frac{d}{dt}\left(\frac{\hat{\boldsymbol{\imath}}}{\imath^2}\right) + \frac{1}{c^2}\frac{d^2\hat{\boldsymbol{\imath}}}{dt^2}\right],$$

where $\hat{\boldsymbol{\imath}}$ points from the shadow to the field point and the derivatives are taken with respect to the observer's time t, at a fixed field point (R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Vol. II, Section 21-1, https://feynmanlectures.caltech.edu/II_21.html). Every quantity in it is something the observer can see: the direction of the shadow, its distance, and how they change. (It agrees with (27.6): "On Feynman's formula for the electromagnetic field of an arbitrarily moving charge," A. R. Janah, T. Padmanabhan, and T. P. Singh, Am. J. Phys. 56, 1036–1038 (1988), https://doi.org/10.1119/1.15334.)

(a) Which of the three terms is largest far from the charge? Explain why it falls off as $1/\imath$.

(b) At a retarded time at which the charge is momentarily at rest, use (27.3) to show that $d\hat{\boldsymbol{\imath}}/dt = 0$ and $d^2\hat{\boldsymbol{\imath}}/dt^2 = -\mathbf{a}_\perp/\imath$, where $\mathbf{a}_\perp$ is the part of the acceleration perpendicular to $\hat{\boldsymbol{\imath}}$, so that the last term is the acceleration field (27.9).

(c) Show that at that time the middle term is zero, so that the formula gives the field of Exercise P27.4.

Exercises for 27.3 What is radiation?

P27.10 Only the $1/\imath$ fields radiate.

(a) Show that far from a moving charge the Poynting vector of the velocity field alone falls off as $1/\imath^4$, and that the cross terms between the velocity field and the acceleration field fall off as $1/\imath^3$. What happens to the power that each carries through a sphere of radius $\imath$ as $\imath \to \infty$?

(b) Sort the terms of Jefimenko's equations, (24.15) and (24.16), by the power of $1/\imath$ with which they fall off far from the sources, and say which products of them can contribute to the radiated power.

(c) A charge that moves with constant velocity carries its fields, and their energy, along with it. Why is that not radiation?

Exercises for 27.4 Larmor's formula

P27.11 Larmor's formula from Thomson's kink. In Figure 27.5, the shell of thickness $c\,\Delta t$ carries the transverse electric field $E_\perp$ found in Section 27.2.2, and with it a magnetic field of magnitude $E_\perp/c$, perpendicular to $E_\perp$ and to the radius.

(a) Show that the energy per unit volume in the shell, beyond that of the Coulomb field, is $\varepsilon_0 E_\perp^2$. (Use the energy densities (5.24) and (12.36).)

(b) Integrate over the shell to find the energy of the pulse, and divide it by the time Δt during which the charge accelerated. Compare with Larmor's formula.

(c) Why does the energy of the pulse not depend on the radius of the shell?

P27.12 The classical atom. In the Bohr model of hydrogen the electron moves in a circle of radius $a_0 = 5.29 \times 10^{-11}\,\text{m}$, the Bohr radius, around the proton. A classical electron would radiate, lose energy, and spiral in.

(a) Find its speed, its acceleration, and the period of its orbit at the radius a_0. Is Larmor's formula valid for it?

(b) Suppose that the orbit stays nearly circular as it shrinks. Use the total energy $-e^2/8\pi\varepsilon_0 r$ of a circular orbit of radius r and Larmor's formula to find dr/dt.

(c) Find how long the electron would take to spiral from a_0 into the proton, and how many times the period of the orbit at a_0 that is. (This collapse is one of the reasons that classical physics cannot describe the atom.)

P27.13 An electron is released from rest at the top of a long, evacuated vertical tube, and it falls freely through 1.00 m. Suppose that no force but gravity acts on it.

(a) Find the power it radiates while it falls, the time the fall takes, and the energy it radiates.

(b) What fraction of the potential energy it loses does it radiate?

(c) Show that for a fall through the height h the fraction is $\tau_{\text{rad}}\sqrt{2g/h}$, with τ_{rad} of Example 27.3, and explain why it is so small.

P27.14 An electron moving at $1.0 \times 10^5\,\text{m/s}$ strikes a metal surface and is brought to rest, with constant deceleration, in a distance of 3.0 nm. What fraction of its kinetic energy does it radiate? Write the answer in terms of τ_{rad} of Example 27.3.

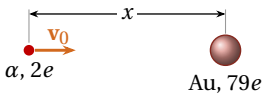


Figure 27.23 The alpha particle of Exercise P27.15, fired straight at a gold nucleus, which is held fixed.

P27.15 An alpha particle, with the charge $q = 2e$ and the mass 6.64×10^{-27} kg, is fired with a kinetic energy of 5.0 MeV straight at a gold nucleus, with the charge $Q = 79e$, which is held fixed (Figure 27.23). It slows down, stops, and comes back out along the same line. It moves slowly compared with light, and in finding its motion you may ignore the energy it radiates.

(a) Show that at the distance x from the nucleus its acceleration is $a = qQ/4\pi\varepsilon_0 mx^2$ and its speed is $v = v_0\sqrt{1 - x_0/x}$, where v_0 is its initial speed and $x_0 = qQ/2\pi\varepsilon_0 mv_0^2$ is the distance of closest approach. Evaluate v_0/c and x_0.

(b) Find the energy it radiates on the whole trip, in and out. (Write $dt = dx/v$.)

(c) Show that the fraction of its kinetic energy that it radiates is

$$\frac{16}{45}\frac{q}{Q}\left(\frac{v_0}{c}\right)^3,$$

and evaluate it.

P27.16 A charge q oscillates along the z axis as $z(t) = d\cos\omega t$, with $d\omega \ll c$.

(a) Find the power it radiates at each instant, and the average over a cycle.

(b) In which directions is the radiation strongest, and in which is there none?

(c) Together with a charge $-q$ fixed at the origin, the oscillating charge forms an electric dipole of moment $\mathbf{p} = qz\hat{\mathbf{z}}$, with the amplitude $p_0 = qd$. Show that the average power is $\mu_0 p_0^2\omega^4/12\pi c$. (Section 28.4 finds the same power for any small oscillating dipole.)

Exercises for 27.5 Liénard's generalization

P27.17 Liénard's formula in two cases.

(a) Use (25.25) to find $\alpha_\mu\alpha^\mu$ for a charge whose acceleration is along its velocity, and for one whose acceleration is perpendicular to its velocity, and show that (27.17) then gives γ^6 and γ^4 times Larmor's power.

(b) A charge moves along a straight line with a constant proper acceleration g, as Rachel's ship does in Example 25.5. Show that it radiates the same power at every moment, although its acceleration in the lab, g/γ^3, decreases as it speeds up.

P27.18 A linear accelerator pushes electrons along a straight line, and an electron gains energy at the rate dE/dx per unit distance.

(a) The force on the electron is along its velocity and has the magnitude $F = dE/dx$. Show from $p = \gamma mv$ that $F = dp/dt = \gamma^3 ma$, and that the electron radiates the power $P = \mu_0 e^2 F^2/6\pi cm^2$, whatever its speed.

(b) Show that the ratio of the radiated power to the rate Fv at which the electron gains energy is

$$\frac{P}{Fv} = \frac{2}{3}\frac{r_e}{mc^2}\frac{dE}{dx}\frac{c}{v},$$

where $r_e = e^2/4\pi\varepsilon_0 mc^2 = 2.818 \times 10^{-15}\,\text{m}$ is the classical electron radius.

(c) A typical linear accelerator gives electrons about 20 MeV per meter. Evaluate the ratio for a fast electron. How large would dE/dx have to be for the electron to radiate as much power as it gains?

Exercises for 27.6 Synchrotron radiation

P27.19 The energy lost per turn.

(a) Show that a charge q moving on a circle of radius R at constant speed loses the energy

$$U_0 = \frac{q^2\beta^3\gamma^4}{3\varepsilon_0 R}$$

on each turn.

(b) For an electron with $\beta \approx 1$ and the energy E, show that

$$U_0 = (88.5\,\text{keV})\,\frac{(E/1\,\text{GeV})^4}{R/1\,\text{m}}.$$

(c) A large electron ring has a bending radius of 3.1 km. How much energy does an electron lose on each turn at 50 GeV, and at 100 GeV?

(d) A proton ring with a bending radius of 2.8 km stores protons with an energy of 6.8 TeV. How much energy does a proton lose on each turn? How much would an electron of the same energy lose on the same circle?

P27.20 A storage ring for a synchrotron light source holds electrons with an energy of 7.0 GeV, and the field in its bending magnets is 0.60 T (Figure 27.12).

(a) Find γ, the radius of the electrons' circles in the bending magnets, and the half-angle $1/\gamma$ of the cone of radiation, in microradians. How wide is the cone at a sample 30 m from the source?

(b) Estimate how long the flash lasts that an observer in the plane of the orbit receives.

(c) Find the critical frequency and the photon energy $\hbar\omega_c$.

P27.21 The pattern of a charge in circular motion.

(a) Derive (27.19) from (27.16): take $\mathbf{v}$ along z and $\mathbf{a}$ along x, and $\hat{\boldsymbol{\varepsilon}} = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$.

(b) Show that for $\beta \to 0$ it becomes Larmor's pattern (27.14).

(c) Show that the numerator is at least $(\cos\theta - \beta)^2$, and find the directions in which the pattern vanishes. Show that they are the directions into which aberration (25.14) carries the two directions in which the charge sends no radiation in its instantaneous rest frame.

(d) Integrate the pattern over all directions numerically, or with computer algebra, for $\beta = 0.5$, and check that the total is γ^4 times Larmor's power.

Exercises for 27.7 Bremsstrahlung

P27.22 The bremsstrahlung lobes.

(a) Show that the pattern (27.21) is largest at the angle $\theta_{\max}$ with

$$\cos\theta_{\max} = \frac{\sqrt{1+15\beta^2}-1}{3\beta}.$$

(b) Evaluate $\theta_{\max}$ for $\beta = 0.5$ and $\beta = 0.9$.

(c) Show that for β close to 1, $\theta_{\max} \approx 1/2\gamma$.

P27.23 The total power of bremsstrahlung.

(a) Show that the total power radiated by a charge whose acceleration is along its velocity is

$$P = \frac{\mu_0 q^2 a^2}{8\pi c}\int_{-1}^{1}\frac{1-u^2}{(1-\beta u)^5}\,du.$$

(b) Evaluate the integral, by hand or with computer algebra, and show that P is γ^6 times Larmor's power, as Liénard's formula says.

P27.24 The end of the X-ray spectrum.

(a) Find the shortest wavelength of the X-rays from a tube operated at 50 kV, and at 120 kV.

(b) A classical electron that is stopped in one collision lasting the time τ_c radiates a spectrum that is flat up to about $1/\tau_c$. For an electron of 50 keV, how short would the collision have to be for the classical spectrum to reach the angular frequency $eV/\hbar$? Compare with the time the electron takes to cross an atom, about 0.1 nm. (Find its speed relativistically.)

(c) What does the classical theory predict for frequencies above $eV/\hbar$, and what is observed?

P27.25 Kramers' law. An X-ray tube with a tungsten target ($Z = 74$) runs at 50 kV, with a current of 20 mA.

(a) Show that Kramers' law (27.22) gives a power per unit wavelength proportional to $(\lambda/\lambda_{\text{min}} - 1)/\lambda^3$, and a number of photons per second per unit wavelength proportional to $(\lambda/\lambda_{\text{min}} - 1)/\lambda^2$, the form in which the law is often quoted. At what wavelength is the number of photons per unit wavelength largest?

(b) Show that the X-ray power is proportional to ZV^2, so that the efficiency of the tube, the X-ray power divided by the power that the electron beam delivers, is proportional to ZV.

(c) Kramers estimated the efficiency as $2.3 \times 10^{-4}\, Z(v_0/c)^2$, with $\frac{1}{2}mv_0^2 = eV$; measurements of his time gave about twice as much. With his estimate, find the efficiency of the tube, its X-ray power, and the power that heats the target.

Exercises for 27.8 Numerical method: fields of an arbitrarily moving charge

P27.26 Field lines of a charge moving along a line. A charge moves along the x axis in any way, and θ is the angle between the x axis and $\hat{\boldsymbol{r}}$ for a field line in the xy plane.

(a) Show that (27.24) gives $d\theta/dt_r = -(\gamma^2 a/c)\sin\theta$, where a is the x component of the acceleration.

(b) With $a = c\, d\beta/dt_r$, show that $d\theta/d\beta = -\sin\theta/(1-\beta^2)$, and integrate to find

$$\tan\frac{\theta}{2} = \sqrt{\frac{1-\beta}{1+\beta}}\tan\frac{\theta'}{2},$$

where θ' is a constant, the value of θ when $\beta = 0$.

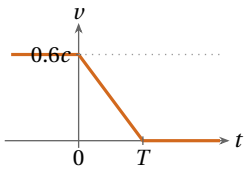


Figure 27.24 The speed of the charge of Exercise P27.27 against time.

(c) Show that this relation is the aberration formula (25.14), with $\theta_R = \theta'$ and $\theta_L = \theta$. What does the result say about the direction of each field line in the charge's instantaneous rest frame?

P27.27 A charge moving along $+x$ at $0.6c$ slows down with constant deceleration and comes to rest at the origin at the time T (Figure 27.24).

(a) With the method of Section 27.8, draw its electric field lines at the time $t = 5T$. Start the lines before the deceleration, at directions spaced evenly in angle in the charge's rest frame (Exercise P27.26).

(b) Where in your picture is the radiation, and which way do the field lines point in it?

(c) Check Gauss's law numerically, with the fields (27.6), through a sphere that cuts through the region of the radiation.

P27.28 A charge moves on a circle of radius R at constant speed, with $\gamma = 2$. An observer in the plane of the orbit, a distance $100R$ from its center, measures the electric field.

(a) Compute the field at the observer during one period, by finding the retarded time and using (27.6), and plot its component perpendicular to the line of sight.

(b) Measure the width of the pulse, and compare it with $R/\gamma^3 c$.

(c) Repeat for $\gamma = 5$.

P27.29 A charge moves on a circle of radius R at the constant speed $0.5c$. Draw its field lines in the plane of the orbit at one instant, out to a distance of about $40R$. Start each line at the charge's present position, with $\hat{\boldsymbol{\imath}}$ at the angle that aberration gives for directions spaced evenly in the charge's instantaneous rest frame, and integrate (27.24) backward in t_r for several periods. Describe the pattern. How far apart are its successive turns, and why?

Exercises for 27.9 Summary

P27.30 True or false? Explain your reasoning.

(a) A charge moving with constant velocity radiates.

(b) The velocity in the Liénard–Wiechert potentials is the velocity at which an observer sees the charge's shadow move.

(c) The velocity field of a moving charge points away from the charge's present position.

(d) Far from an accelerated charge, **E**, **B**, and $\hat{\boldsymbol{r}}$ are perpendicular to one another.

(e) The power that a slowly moving charge radiates crosses a sphere of radius 1 m around it and a sphere of radius 1 km at the same rate, each when the news reaches it.

(f) A slowly moving accelerated charge radiates most strongly in the direction of its acceleration.

(g) The power that a charge radiates is the same in every inertial frame.

(h) An electron and a proton with the same energy, moving on the same circle, radiate the same power.

(i) A fast charge whose acceleration is along its velocity radiates most strongly straight ahead.

(j) A shorter collision radiates a broader band of frequencies.

P27.31 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as the retarded time and the fields of a charge in any motion, the field lines of an accelerated charge, the radiation patterns of fast charges on circles and on straight lines, the pulse and the spectrum that an observer receives from a circling charge, or the spiral of a classical atom that collapses as it radiates.

Chapter 28

Dipole and Multipole Radiation

Chapter 27 found how much a single accelerated charge radiates, and in which directions. This chapter does the same for systems of charges and currents that are small compared with the wavelength of the radiation they send out: the antennas of radios and phones, molecules shaken by sunlight, a spinning magnet. Such a source is described by its multipole moments, as a static distribution of charge was in Section 8.3, and when its moments change in time it radiates.

The chapter starts with two models: an oscillating electric dipole, two charges that change sign back and forth, and an oscillating magnetic dipole, a loop that carries an alternating current. Then it treats any small source at all. Its radiation is a sum of contributions from its electric dipole moment, its magnetic dipole moment, its electric quadrupole moment, and so on, and the accelerated point charge of Chapter 27 returns as the simplest case. The tools are the retarded potentials of Section 24.4 and the meaning of radiation from Section 27.3. The last section computes numerically the radiation of antennas that are not small.

28.1 Radiation from small sources: power and angular distribution

Picture a source of some size d, made of charges and currents, near the origin, and surround it with a large sphere of radius r. When the charges are shaken, electromagnetic waves travel outward, and the questions are how much power gets through the sphere and in which directions it goes. The source must be localized. A source that extends to infinity, such as a sheet of current that fills a whole plane, sends out waves that do not spread out, and it needs a different treatment.

Section 27.3 settled which fields can carry energy all the way out. The area of the sphere grows as r^2, so only the parts of the fields that fall off as $1/r$ give a power through the sphere that does not shrink as the sphere grows. Static fields fall off as $1/r^2$ or faster, so they never radiate: the parts of the fields that radiate come from the changes of the source.

The most important small source is an antenna. Connect a source of alternating voltage to a short wire, and charge sloshes back and forth along it: one end is positive while the other is negative, then the reverse, at the frequency of the source. To find what such a source radiates, we will follow five steps, the same for every source in this chapter:

1. Find the scalar potential V. This is usually the longest step.
2. Find the vector potential $\mathbf{A}$.
3. Find the fields, $\mathbf{E} = -\nabla V - \partial\mathbf{A}/\partial t$ and $\mathbf{B} = \nabla \times \mathbf{A}$.
4. Find the Poynting vector $\mathbf{S} = \mathbf{E} \times \mathbf{B}/\mu_0$, and its average over a cycle, $\langle \mathbf{S} \rangle$.
5. Integrate $\langle \mathbf{S} \rangle$ over the large sphere to get the average radiated power, $\langle P \rangle = \oint \langle \mathbf{S} \rangle \cdot d\mathbf{a}$.

Only the parts of the fields that fall off as $1/r$ contribute to the radiated power, so along the way we will drop the rest, once it is clear which parts they are.

The sources of this chapter oscillate at an angular frequency ω, so their fields and the Poynting vector oscillate too, and what matters for the energy is the average over a cycle. The average of $\cos^2 \omega t$ or $\sin^2 \omega t$ over a cycle is $\frac{1}{2}$. The charges of a small source move slowly, so the distinction that Section 27.3 made between the power emitted and the power received does not arise, and the average power per unit solid angle is

$$\frac{dP}{d\Omega} = r^2 \langle \mathbf{S} \rangle \cdot \hat{\mathbf{r}}, \tag{28.1}$$

evaluated on a large sphere, where it does not depend on r.

A good way to picture $dP/d\Omega$ is as a **radiation pattern**: a surface whose distance from the origin in each direction is proportional to $dP/d\Omega$ in that direction. Figure 27.8 shows a cut through the radiation pattern of a slowly moving accelerated charge. Turned about the acceleration, it makes a doughnut: lots of radiation to the sides, none along the axis (Figure 28.1). The sources of this chapter have patterns of the same kind, and Section 28.5.3 shows how to plot them in three dimensions.

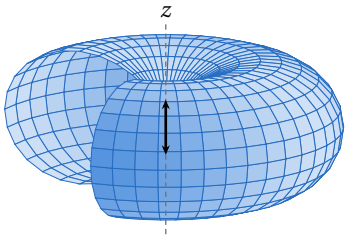


Figure 28.1 The radiation pattern of Figure 27.8 in three dimensions, $dP/d\Omega \propto \sin^2\theta$, with a quarter cut away. A slowly moving accelerated charge and an oscillating dipole (Section 28.2.2) both radiate this way, about the line of the acceleration or of the dipole moment (black arrow). The doughnut pinches to a point at the center, since no power goes along the axis.

Three lengths decide which approximations are allowed: the size d of the source, the distance r from it, and c/ω, the distance light travels while the phase of the source advances by one radian, which is the wavelength $\lambda = 2\pi c/\omega$ divided by 2π. Most of this chapter assumes that

$$d \ll \frac{c}{\omega} \ll r. \tag{28.2}$$

The first inequality says that the source is small compared with the wavelength. The news from all of its parts then reaches a distant point at nearly the same time, and its charges move slowly, since their speeds are of the order of $d\omega$, much less than c. The second inequality says that the field point is many wavelengths away, in what is called the **radiation zone**. A factor of 2π hardly matters when one length

is much smaller than another, so the same conditions are often written $d \ll \lambda \ll r$. Sunlight, with wavelengths from 400 to 700 nm, shaking an air molecule a fraction of a nanometer across and seen from kilometers away, satisfies both inequalities easily. The antenna of an AM radio station does not: its tower is typically about a quarter of a wavelength tall, so the first inequality fails, and Section 28.5 treats such antennas (Exercise P28.1).

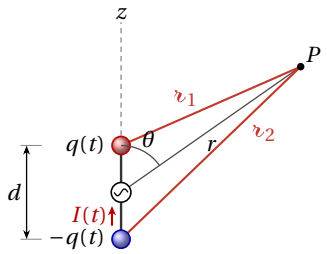


Figure 28.2 A model of an antenna: two small spheres a distance d apart, joined by a wire with a source of alternating voltage (the circle) in the middle. The spheres carry the charges $\pm q(t)$, and the wire carries the current $I(t)$. At the field point P, a distance r from the origin and at the angle θ from the z axis, the distances to the two spheres are r_1 and r_2.

28.2 Electric dipole radiation

The simplest antenna is the one of Figure 28.2: two small metal spheres a distance d apart, joined by a thin straight wire, with a source of alternating voltage in the middle of the wire. The source pushes charge back and forth along the wire, so that the top sphere carries the charge

$$q(t) = q_0 \cos \omega t$$

and the bottom sphere carries $-q(t)$: always equal and opposite, and changing sign twice in every cycle. The charges stay on the spheres, and the wire between them carries the current $I(t) = dq/dt = -q_0 \omega \sin \omega t$, counted as positive upward. The wire is needed: charge cannot appear on one sphere and vanish from the other unless it flows between them. With the origin halfway between the spheres and the z axis through them, the dipole moment (8.11) is

$$\mathbf{p}(t) = q(t)\, d\, \hat{\mathbf{z}} = p_0 \cos \omega t\, \hat{\mathbf{z}}, \qquad p_0 = q_0 d. \tag{28.3}$$

The dipole moment oscillates: up, down, up, down. An antenna that is much shorter than the wavelength and carries the same current all along its length is called a **Hertzian dipole**, after Heinrich Hertz, who first made and detected electromagnetic waves (Section 16.5).

28.2.1 The potentials

It is tempting to take the potential of a static dipole, $V = p \cos\theta / 4\pi\varepsilon_0 r^2$ from (8.14), and put $p(t)$ into it. That would be wrong. It says that the potential everywhere changes at the same moment as the charges, and the news of a change cannot travel faster than light. Instead, each sphere is a point charge at rest whose charge changes in time, and by the retarded potential (24.11) its potential is its charge at the retarded time divided by $4\pi\varepsilon_0$ times its distance. With r_1 and r_2 the distances from the top and bottom spheres to the field point (Figure 28.2),

$$V(r, \theta, t) = \frac{q_0}{4\pi\varepsilon_0} \left[\frac{\cos \omega (t - r_1/c)}{r_1} - \frac{\cos \omega (t - r_2/c)}{r_2} \right]. \tag{28.4}$$

Two cautions. The two distances are different, but each is constant in time, since the spheres do not move; that makes this far easier than the moving charges of

Chapter 27. And each charge enters at its own retarded time: from a point above the dipole the bottom sphere is farther away, so the potential there is made of the top charge as it was a moment ago and the bottom charge as it was a little earlier still.

! **Three facts we will need.** The law of cosines, $C^2 = A^2 + B^2 - 2AB\cos\gamma$; $(1+x)^n \cong 1 + nx$ for $|x| \ll 1$; and $\cos(A+B) \cong \cos A - B\sin A$ for $|B| \ll 1$.

Equation (28.4) is exact, and we now simplify it with the first two inequalities of (28.2). The first approximation is $d \ll r$. The position of the top sphere is $\mathbf{r}' = (d/2)\,\hat{\mathbf{z}}$, so by the law of cosines

$$\mathscr{r}_1^2 = r^2 + \left(\frac{d}{2}\right)^2 - rd\cos\theta \cong r^2\left(1 - \frac{d}{r}\cos\theta\right),$$

where the term $(d/2)^2$ is dropped because it is much smaller than rd. Then the binomial approximation gives

$$\mathscr{r}_1 \cong r\left(1 - \frac{d}{2r}\cos\theta\right), \qquad \frac{1}{\mathscr{r}_1} \cong \frac{1}{r}\left(1 + \frac{d}{2r}\cos\theta\right).$$

For the bottom sphere, at $-(d/2)\,\hat{\mathbf{z}}$, the angle in the triangle is $\pi - \theta$, so the signs change: $\mathscr{r}_2 \cong r[1 + (d/2r)\cos\theta]$ and $1/\mathscr{r}_2 \cong (1/r)[1 - (d/2r)\cos\theta]$.

The second approximation is $d \ll c/\omega$. The argument of the top charge's cosine is

$$\omega\left(t - \frac{\mathscr{r}_1}{c}\right) \cong \omega\left(t - \frac{r}{c}\right) + \frac{\omega d}{2c}\cos\theta,$$

and the second term is small, so

$$\cos\omega\left(t - \frac{\mathscr{r}_1}{c}\right) \cong \cos\omega\left(t - \frac{r}{c}\right) - \frac{\omega d}{2c}\cos\theta\,\sin\omega\left(t - \frac{r}{c}\right).$$

The bottom charge's cosine is the same with the sign of the second term changed. This step needs the source to be small compared with c/ω, not just compared with r: the difference in the retarded times of the two charges, about $(d/c)\cos\theta$, must be a small fraction of a cycle.

Put these into (28.4), multiply out, and keep the terms of first order in d. Every term that does not contain $d\cos\theta/r$ cancels between the two charges, and with $p_0 = q_0 d$,

$$V(r,\theta,t) = \frac{p_0\cos\theta}{4\pi\varepsilon_0 r}\left[\frac{1}{r}\cos\omega\left(t - \frac{r}{c}\right) - \frac{\omega}{c}\sin\omega\left(t - \frac{r}{c}\right)\right]. \tag{28.5}$$

The time $t - r/c$ is the retarded time of the origin, which we will call t_0. The first term is then the static dipole potential (8.14) with $p(t_0)$ in place of p: the naive guess, with the right delay. The second term is new. It comes from the difference in the retarded times of the two charges, and it falls off only as $1/r$. The second term is larger than the first by the factor $\omega r/c$, so in the radiation zone, $r \gg c/\omega$, the first term can be dropped:

$$\boxed{V(r,\theta,t) = -\frac{p_0\omega}{4\pi\varepsilon_0 c}\,\frac{\cos\theta}{r}\,\sin\omega\left(t - \frac{r}{c}\right)}. \tag{28.6}$$

That completes the first step.

The vector potential comes from the current in the wire. By the retarded potential (24.12), with $\mathbf{J}\,d\tau'$ replaced by $I\,dz'\,\hat{\mathbf{z}}$ for a thin wire,

$$\mathbf{A}(r,\theta,t) = \frac{\mu_0}{4\pi}\int_{-d/2}^{d/2} \frac{I(t-\mathscr{r}/c)}{\mathscr{r}}\,dz'\,\hat{\mathbf{z}} = -\frac{\mu_0 q_0 \omega}{4\pi}\,\hat{\mathbf{z}}\int_{-d/2}^{d/2}\frac{\sin\omega(t-\mathscr{r}/c)}{\mathscr{r}}\,dz',$$

where $\mathscr{r}$ is the distance from the point z' of the wire to the field point (Figure 28.3). The wire is short, and by the same two approximations the integrand differs from its value at the midpoint of the wire, $z'=0$, only by fractions of the order of d/r and $\omega d/c$. So the integral is the integrand at the midpoint times the length of the wire, $\sin\omega(t-r/c)\,d/r$, and

$$\boxed{\mathbf{A}(r,\theta,t) = -\frac{\mu_0 p_0 \omega}{4\pi r}\sin\omega\left(t-\frac{r}{c}\right)\hat{\mathbf{z}}}. \tag{28.7}$$

This already falls off as $1/r$, so no third approximation is needed. The vector potential points along $\hat{\mathbf{z}}$ everywhere, like the current that produces it. In which direction, then, does $\mathbf{B}=\nabla\times\mathbf{A}$ point? Around the z axis, along $\hat{\boldsymbol{\phi}}$, as it does around a straight wire.

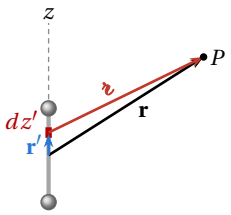


Figure 28.3 The wire of the antenna and the vectors of the script-r method for its vector potential: $\mathbf{r}$ (black) to the field point P, $\mathbf{r}'$ (blue) to a piece dz' of the wire, and $\boldsymbol{\mathscr{r}}$ (red) from that piece to P. The wire is short, so $\mathscr{r}$ differs little from r.

28.2.2 The fields and the power

The third step is the fields. The gradient in spherical coordinates (Section 1.4.6), with the term that vanishes crossed out, is

$$\nabla V = \frac{\partial V}{\partial r}\hat{\mathbf{r}} + \frac{1}{r}\frac{\partial V}{\partial\theta}\hat{\boldsymbol{\theta}} + \cancel{\frac{1}{r\sin\theta}\frac{\partial V}{\partial\phi}\hat{\boldsymbol{\phi}}}.$$

Differentiating (28.6) with respect to r gives two terms: one from the $1/r$, which falls off as $1/r^2$, and one from the sine, which falls off as $1/r$. The θ derivative is divided by r, so it falls off as $1/r^2$ as well. Only the $1/r$ term matters in the radiation zone, so

$$\nabla V \cong -\frac{p_0\omega}{4\pi\varepsilon_0 c}\frac{\cos\theta}{r}\frac{\partial}{\partial r}\sin\omega\left(t-\frac{r}{c}\right)\hat{\mathbf{r}} = \frac{\mu_0 p_0\omega^2}{4\pi}\frac{\cos\theta}{r}\cos\omega\left(t-\frac{r}{c}\right)\hat{\mathbf{r}},$$

since $1/\varepsilon_0 c^2 = \mu_0$. The time derivative of (28.7) is

$$\frac{\partial\mathbf{A}}{\partial t} = -\frac{\mu_0 p_0\omega^2}{4\pi r}\cos\omega\left(t-\frac{r}{c}\right)\hat{\mathbf{z}},$$

so

$$\mathbf{E} = -\nabla V - \frac{\partial\mathbf{A}}{\partial t} = \frac{\mu_0 p_0\omega^2}{4\pi r}\cos\omega\left(t-\frac{r}{c}\right)(\hat{\mathbf{z}}-\cos\theta\,\hat{\mathbf{r}}).$$

With $\hat{\mathbf{z}} = \cos\theta\,\hat{\mathbf{r}} - \sin\theta\,\hat{\boldsymbol{\theta}}$, from (1.8), the radial parts cancel, and

$$\boxed{\mathbf{E}(r,\theta,t) = -\frac{\mu_0 p_0\omega^2}{4\pi}\frac{\sin\theta}{r}\cos\omega\left(t-\frac{r}{c}\right)\hat{\boldsymbol{\theta}}}. \tag{28.8}$$

For the magnetic field, (28.7) has the spherical components $A_r = A_z\cos\theta$ and $A_\theta = -A_z\sin\theta$, and $A_\phi = 0$. The curl in spherical coordinates, with the terms that vanish crossed out, is

$$\nabla\times\mathbf{A} = \frac{1}{r\sin\theta}\left(\cancel{\frac{\partial(\sin\theta\,A_\phi)}{\partial\theta}} - \cancel{\frac{\partial A_\theta}{\partial\phi}}\right)\hat{\mathbf{r}} + \frac{1}{r}\left(\cancel{\frac{1}{\sin\theta}\frac{\partial A_r}{\partial\phi}} - \cancel{\frac{\partial(r\,A_\phi)}{\partial r}}\right)\hat{\boldsymbol{\theta}} + \frac{1}{r}\left(\frac{\partial(r\,A_\theta)}{\partial r} - \frac{\partial A_r}{\partial\theta}\right)\hat{\boldsymbol{\phi}}.$$

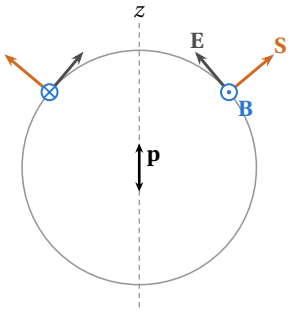


Figure 28.4 The radiation fields of an oscillating electric dipole (center) at one point of a large sphere around it, at one instant: **E** along the meridian, **B** along the circle of latitude, and the Poynting vector **S** straight out. Half a cycle later **E** and **B** have both reversed, and **S** still points out.

The last term, $(1/r)\,\partial A_r/\partial\theta$, falls off as $1/r^2$. The other one is

$$\frac{1}{r}\frac{\partial(r\,A_\theta)}{\partial r} = \frac{1}{r}\frac{\partial}{\partial r}\left[\frac{\mu_0 p_0\omega}{4\pi}\sin\theta\,\sin\omega\left(t-\frac{r}{c}\right)\right] = -\frac{\mu_0 p_0\omega^2}{4\pi c}\frac{\sin\theta}{r}\cos\omega\left(t-\frac{r}{c}\right),$$

so

$$\boxed{\mathbf{B}(r,\theta,t) = -\frac{\mu_0 p_0\omega^2}{4\pi c}\frac{\sin\theta}{r}\cos\omega\left(t-\frac{r}{c}\right)\hat{\boldsymbol{\phi}}}. \tag{28.9}$$

These fields have the features of the plane waves of Section 16.5 (Figure 28.4): **E** and **B** are perpendicular to each other and to the direction of travel, $\hat{\mathbf{r}}$, the magnetic field has the size $B = E/c$, and $\mathbf{E}\times\mathbf{B}$ points outward. But they are spherical waves: the wavefronts are spheres centered on the dipole, and the amplitude falls off as $1/r$, as it must for the same power to cross every sphere. Since the fields came from approximations, it is worth checking that they still satisfy Maxwell's equations. To order $1/r$ they do (Exercise P28.2).

The fourth step is the Poynting vector. With $\hat{\boldsymbol{\theta}}\times\hat{\boldsymbol{\phi}} = \hat{\mathbf{r}}$,

$$\mathbf{S} = \frac{1}{\mu_0}\mathbf{E}\times\mathbf{B} = \frac{\mu_0 p_0^2\omega^4}{16\pi^2 c}\frac{\sin^2\theta}{r^2}\cos^2\omega\left(t-\frac{r}{c}\right)\hat{\mathbf{r}},$$

and its average over a cycle is

$$\boxed{\langle\mathbf{S}\rangle = \frac{\mu_0 p_0^2\omega^4}{32\pi^2 c}\frac{\sin^2\theta}{r^2}\hat{\mathbf{r}}}. \tag{28.10}$$

Its size is the **intensity** of the radiation, the average power per unit area, in watts per square meter. It falls off as $1/r^2$, as it must for spherical waves. By (28.1), $dP/d\Omega = (\mu_0 p_0^2\omega^4/32\pi^2 c)\sin^2\theta$: the same doughnut as the accelerated charge of Figure 27.8, with the dipole moment in the place of the acceleration. The dipole sends nothing along its own axis and the most in the plane perpendicular to it.

The fifth step is the total power. On the sphere $d\mathbf{a} = r^2\sin\theta\,d\theta\,d\phi\,\hat{\mathbf{r}}$, by (1.15). (There is no dr, since this is an integral over a surface.) With $\int_0^\pi \sin^3\theta\,d\theta = 4/3$,

$$\langle P\rangle = \oint\langle\mathbf{S}\rangle\cdot d\mathbf{a} = \frac{\mu_0 p_0^2\omega^4}{32\pi^2 c}\int_0^{2\pi}\int_0^\pi \sin^3\theta\,d\theta\,d\phi = \frac{\mu_0 p_0^2\omega^4}{32\pi^2 c}(2\pi)\left(\frac{4}{3}\right),$$

so

$$\boxed{\langle P\rangle = \frac{\mu_0 p_0^2\omega^4}{12\pi c}}. \tag{28.11}$$

This is the power that Exercise P27.16 found for a single charge q oscillating on a line with the amplitude d, with $p_0 = qd$: a first hint that what radiates is the changing dipole moment, however it is made (Section 28.4). The power grows as the fourth power of the frequency. For the same dipole moment, doubling the frequency multiplies the radiated power by 16.

All of this holds in the radiation zone. For an ideal dipole, with $d \to 0$ at a fixed p_0, the fields can also be found at every distance, without the third approximation (Exercise P28.3), and their field lines show how radiation leaves its source (Figure 28.5).

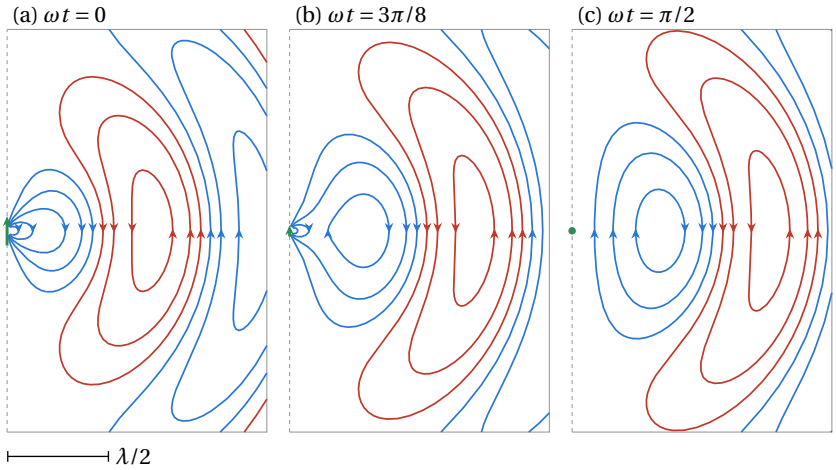


Figure 28.5 The electric field lines of an ideal oscillating dipole, $\mathbf{p} = p_0 \cos\omega t\, \hat{\mathbf{z}}$ (green, drawn in proportion to p), at three moments, in half of a plane through its axis; the other half is a mirror image. (a) When p is largest, the lines near the dipole run from its upper end to its lower end, as for a static dipole. (b) As p shrinks, the outer lines that still end on the dipole are pinched together where they cross the plane $z = 0$, and beyond the pinch a new loop has closed. (c) When p is zero, no lines end on the dipole, and the new loop travels outward at the speed of light, behind the loops of earlier half cycles. Lines that circulate clockwise in the drawing are blue, and those that circulate counterclockwise red. The lines are exact, with no approximation; the bar is half a wavelength. Animation.

Near the dipole, within a distance of about c/ω, the lines run from the positive end of the dipole to the negative end, as in a static dipole. As the dipole moment shrinks toward zero, the lines near the dipole shrink back into it, and the outer parts of the lines close on themselves into loops that are attached to no charge. The loops travel outward at the speed of light, and the next half cycle makes loops that circulate the other way. Far away they are the radiation fields (28.8): perpendicular to $\hat{\mathbf{r}}$, with one loop in every half wavelength (Figure 28.6). Hertz worked out these field lines and drew them at four moments of the cycle in 1889.[1]

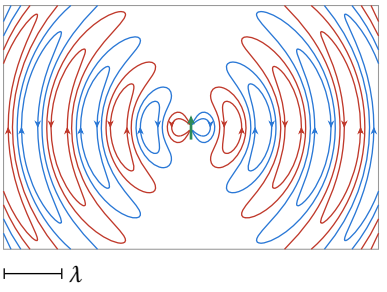


Figure 28.6 The field lines of the dipole of Figure 28.5 at $\omega t = 0$, in a whole plane through it, out to about three wavelengths. Far away the loops are the radiation fields, one loop in every half wavelength; clockwise loops are blue, counterclockwise ones red. The bar is one wavelength. Animation.

[1] The paper is "Die Kräfte electrischer Schwingungen, behandelt nach der Maxwell'schen Theorie," H. Hertz, Ann. Phys. 36, 1–22 (1889), https://doi.org/10.1002/andp.18882720102.

28.2.3 Radiation resistance

An antenna takes power from whatever drives it and sends that power away as radiation. To the driver it looks like a resistor that dissipates the same average power. A current $I(t)$ of amplitude I_0 dissipates the average power $\langle I^2\rangle R = \frac{1}{2}I_0^2 R$ in a resistance R, so the **radiation resistance** R of an antenna is defined by

$$\langle P\rangle = \tfrac{1}{2}I_0^2 R,$$

where I_0 is the amplitude of the current at the point where the antenna is driven. For the antenna of Figure 28.2, $I_0 = q_0\omega$ and $p_0 = q_0 d = I_0 d/\omega$, so by (28.11)

$$\tfrac{1}{2}I_0^2 R = \frac{\mu_0}{12\pi c}\left(\frac{I_0 d}{\omega}\right)^2 \omega^4, \qquad R = \frac{\mu_0 d^2\omega^2}{6\pi c} = \frac{2\pi}{3}\mu_0 c\left(\frac{d}{\lambda}\right)^2,$$

where $\lambda = 2\pi c/\omega$ is the wavelength. The combination $\mu_0 c$ has the units of resistance; it is the impedance of free space of Section 21.4.1,

$$Z_0 = \mu_0 c = 376.73\,\Omega. \tag{28.12}$$

So

$$\boxed{R = \frac{2\pi}{3}Z_0\left(\frac{d}{\lambda}\right)^2 \approx (789\,\Omega)\left(\frac{d}{\lambda}\right)^2}. \tag{28.13}$$

An antenna a tenth of a wavelength long has a radiation resistance of about $8\,\Omega$, and a shorter one has less. The next example shows what a small radiation resistance costs.

Example 28.1

An antenna is made like the one of Figure 28.2: two small spheres joined by a straight copper wire of length $d = 10\,\text{cm}$ and radius $a = 1.0\,\text{mm}$. At high frequencies the current in a wire flows only in a thin layer at its surface, about one skin depth $\delta = \sqrt{2/\mu_0\sigma\omega}$ deep (the skin effect, Section 22.2.2), so the resistance of the wire is about $R_{\text{ohm}} = d/(\sigma 2\pi a\delta)$. What fraction of the power delivered to the antenna is radiated (a) at 100 MHz and (b) at 1.0 MHz? For copper, $\sigma = 5.96\times10^7\,\text{S/m}$ (Section 16.4.1).

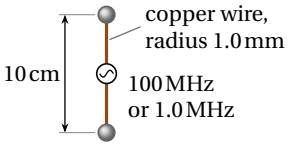
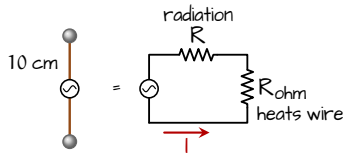


Figure 28.7 Example 28.1: the antenna, with a copper wire 10 cm long and 1.0 mm in radius between the spheres.

Solution:



The first step is a sketch of the antenna as the driver sees it: two resistances in series, the radiation resistance and the resistance of the wire. The same current flows through both, so the power divides between them in proportion to their resistances.

The power that goes into radiation is $\frac{1}{2}I_0^2R$ and the power that heats the wire is $\frac{1}{2}I_0^2R_{\text{ohm}}$, so the fraction of the power that is radiated is

$$\frac{R}{R+R_{\text{ohm}}}.$$

(a) At 100 MHz the wavelength is $\lambda = c/(10^8\,\text{Hz}) = 3.00\,\text{m}$, so by (28.13)

$$R = (789\,\Omega)\left(\frac{0.10\,\text{m}}{3.00\,\text{m}}\right)^2 = 0.878\,\Omega.$$

With $\omega = 2\pi\times10^8\,\text{s}^{-1}$ the skin depth is

$$\delta = \sqrt{\frac{2}{\mu_0\sigma\omega}} = 6.52\times10^{-6}\,\text{m},$$

much less than the radius of the wire, so the thin-layer formula applies, and

$$R_{\text{ohm}} = \frac{0.10\,\text{m}}{(5.96\times10^7\,\text{S/m})(2\pi)(1.0\times10^{-3}\,\text{m})(6.52\times10^{-6}\,\text{m})} = 0.041\,\Omega.$$

The fraction radiated is $0.878/(0.878+0.041) = 0.955$, so 95.5% of the power is radiated.

(b) At 1.0 MHz the wavelength is 300 m, a hundred times longer, so R is 10^4 times smaller, $8.8\times10^{-5}\,\Omega$. The skin depth grows as $1/\sqrt{\omega}$, to 65 μm, still much less than the radius, so R_{ohm} is 10 times smaller, $0.0041\,\Omega$. Now the fraction radiated is $(8.8\times10^{-5})/(0.0042) = 0.021$: only 2.1% of the power is radiated, and the rest heats the wire. Between the two frequencies the fraction radiated rises steeply, and it is one half near 13 MHz (Figure 28.8).

Check: the model needs $d \ll c/\omega$. At 1.0 MHz, $c/\omega = 48\,\text{m}$, far larger than 10 cm. At 100 MHz, $c/\omega = 0.48\,\text{m}$, only about five times d, so the answer to (a) is an estimate. A short antenna is an efficient radiator only if its length is not too small a fraction of the wavelength.

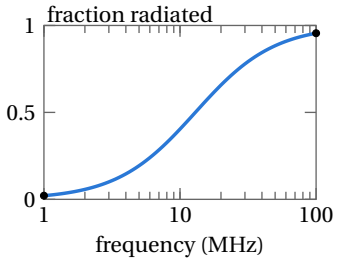


Figure 28.8 The fraction of the power radiated by the antenna of Example 28.1, against the frequency (on a logarithmic scale), with the example's two answers (dots).

The model's current is the same all along the wire, which needs the spheres at the ends to hold the charge. A straight wire with no spheres, fed at its center, has a current that falls to zero at its ends, and its radiation resistance is a quarter of (28.13) (Exercise P28.6). A wire half a wavelength long has a radiation resistance of about 73 Ω (Section 28.5). The cable that carries power to an antenna sees its radiation resistance as the load at its end (Section 23.8).

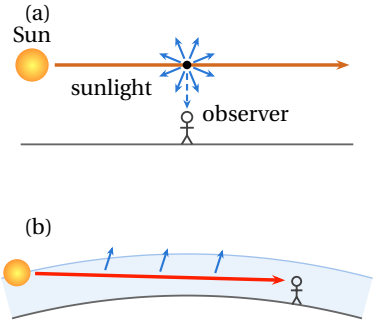


Figure 28.9 (a) An air molecule in a sunbeam scatters light in all directions, so light reaches an observer who is not looking at the Sun. Blue light is scattered much more than red. (b) At sunset the light from the Sun crosses much more air on its way to the observer, and most of its blue has been scattered out.

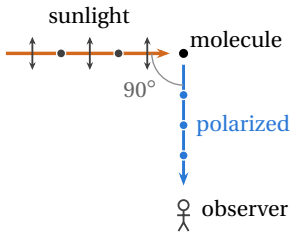


Figure 28.10 Sunlight (orange arrow) drives an air molecule; the observer looks at the molecule at 90° to the sunbeam. The sunlight's field has components along the line of sight and across it (double arrows). A dipole sends nothing along its own axis, so only the component across the line of sight sends light to the observer, and that light is polarized.

28.2.4 Why the sky is blue

Sunlight shakes the molecules of the air. The electric field of the light, oscillating at the angular frequency ω, pushes the electrons of each molecule one way and its nuclei the other, and gives the molecule an induced dipole moment $\mathbf{p} = \alpha\mathbf{E}$ (Section 9.1.5), which oscillates at the same frequency. Each molecule is a tiny antenna. With $p_0 = \alpha E_0$, where E_0 is the amplitude of the light's field, it radiates by (28.11) the average power

$$\langle P \rangle = \frac{\mu_0 \alpha^2 E_0^2 \omega^4}{12\pi c} \tag{28.14}$$

in all directions except along its dipole moment (Figure 28.9): it scatters light out of the sunbeam. The polarizability of a molecule depends on the frequency, but only a little for visible light, whose frequencies are well below those at which the molecules of the air absorb, in the ultraviolet. So the scattered power is proportional to ω^4, or to $1/\lambda^4$. Blue light, with a wavelength around 450 nm, is scattered $(700/450)^4 \approx 5.9$ times as strongly as red light at 700 nm. Look at the sky anywhere away from the Sun, and you see sunlight that the air has scattered toward you, and it is mostly blue. When I teach this material, this is one of my favorite moments, because I can point to the ω^4 in (28.11) and say: and that is why the sky is blue.

The same scattering makes sunsets red. Near sunset the light from the Sun reaches you through a far longer path in the air, and most of its blue has been scattered out of it on the way; what is left is red and orange. Violet light is scattered even more strongly than blue, and yet the sky does not look violet. Sunlight contains less violet than blue, and our eyes are much less sensitive to it, and the mixture of colors that the sky sends us looks to the eye like blue mixed with white.[2] Clouds are white because their water droplets are many times larger than the wavelength of light. The formulas of this chapter do not apply to them, and such droplets scatter all the visible colors about equally.

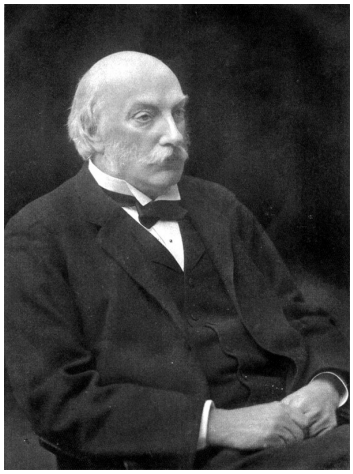
Skylight is also polarized (Figure 28.10). Sunlight is unpolarized: its electric field points in all directions perpendicular to the sunbeam, changing at random. Look at a patch of sky 90° from the Sun. The molecules there are driven by two components of the sunlight's field, one along your line of sight and one perpendicular to it. A dipole sends nothing along its own axis, so the first component sends you nothing, and all the light you receive comes from the second. That light is completely polarized, perpendicular to the plane that contains the Sun, the molecule, and you. In other directions it is partly polarized (Exercise P28.8). Lord Rayleigh worked out the theory of the color and the polarization of the sky in 1871, and in 1899 he showed that it applies to the molecules of the air themselves.[3]

[2] "Human color vision and the unsaturated blue color of the daytime sky," G. S. Smith, Am. J. Phys. 73, 590–597 (2005), https://doi.org/10.1119/1.1858479.

[3] "On the light from the sky, its polarization and colour," J. W. Strutt, Phil. Mag. 41, 107–120 and 274–279 (1871), https://doi.org/10.1080/14786447108640452 and https://doi.org/10.1080/14786447108640479; "On the transmission of light through an atmosphere containing small parti-

The scattering of light by particles much smaller than its wavelength is called **Rayleigh scattering**.

The polarization of the sky can be seen through a polarizing filter, and some animals use it: honeybees find their way by the pattern of polarization of the blue sky, even when the Sun itself is hidden.[4]



John William Strutt, Lord Rayleigh (1842–1919, English) John William Strutt was born at Langford Grove, near Maldon in Essex, and studied mathematics at Trinity College, Cambridge, where he was Senior Wrangler in 1865. His papers of 1871 on the light from the sky explained its color and its polarization. Two years later he succeeded his father as the third Baron Rayleigh. His two volumes of *The Theory of Sound* appeared in 1877–78. He followed James Clerk Maxwell as Cavendish Professor at Cambridge from 1879 to 1884, and then continued his experiments at his home, Terling Place. Careful measurements of the density of nitrogen led him and William Ramsay to the discovery of argon in 1894, and he received the Nobel Prize in Physics in 1904 "for his investigations of the densities of the most important gases and for his discovery of argon in connection with these studies." (Wikipedia)

28.3 Magnetic dipole radiation

The second model is a loop of wire of radius b in the xy plane, centered at the origin, that carries the alternating current $I(t) = I_0 \cos\omega t$ (Figure 28.11). By (12.23) its magnetic dipole moment is

$$\mathbf{m}(t) = I(t)\,\pi b^2\,\hat{\mathbf{z}} = m_0 \cos\omega t\,\hat{\mathbf{z}}, \qquad m_0 = I_0 \pi b^2. \tag{28.15}$$

A loop like this is another kind of antenna, and we follow the same five steps.

The first step is easy this time. The current flows around the loop and no charge piles up anywhere, so $\rho = 0$ and

$$V = 0.$$

The second step, the vector potential, is the hard one. By (24.12), with $\mathbf{J}\,d\tau'$ replaced by $I\,d\boldsymbol{\ell}'$ for a thin wire,

$$\mathbf{A}(\mathbf{r}, t) = \frac{\mu_0}{4\pi} \oint \frac{I(t - \mathscr{r}/c)}{\mathscr{r}}\, d\boldsymbol{\ell}'. \tag{28.16}$$

The integral is simplest if the field point is placed above the x axis, which costs nothing, since the loop looks the same from every direction around the z axis. As in the script-r method (Section 3.5), we write

$$\begin{aligned}
\mathbf{r} &= r\sin\theta\,\hat{\mathbf{x}} + r\cos\theta\,\hat{\mathbf{z}},\\
\mathbf{r}' &= b\cos\phi'\,\hat{\mathbf{x}} + b\sin\phi'\,\hat{\mathbf{y}},\\
\boldsymbol{\mathscr{r}} &= \mathbf{r} - \mathbf{r}' = (r\sin\theta - b\cos\phi')\,\hat{\mathbf{x}} - b\sin\phi'\,\hat{\mathbf{y}} + r\cos\theta\,\hat{\mathbf{z}},\\
\mathscr{r} &= \sqrt{r^2 + b^2 - 2rb\cos\psi}, \qquad \cos\psi = \hat{\mathbf{r}}\cdot\hat{\mathbf{r}}' = \sin\theta\cos\phi',
\end{aligned}$$

where ψ is the angle between $\mathbf{r}$ and $\mathbf{r}'$. The element of the loop is

$$d\boldsymbol{\ell}' = b\,d\phi'\,\hat{\boldsymbol{\phi}}' = b\,d\phi'\left(-\sin\phi'\,\hat{\mathbf{x}} + \cos\phi'\,\hat{\mathbf{y}}\right),$$

by (1.7). Colton's First Commandment (Section 3.5.1) applies here with full force: $\hat{\boldsymbol{\phi}}'$ changes direction around the loop, so it must stay inside the integral as $-\sin\phi'\,\hat{\mathbf{x}} + \cos\phi'\,\hat{\mathbf{y}}$.

cles in suspension, and on the origin of the blue of the sky," Lord Rayleigh, Phil. Mag. 47, 375–384 (1899), https://doi.org/10.1080/14786449908621276.

[4] Karl von Frisch found this: "Die Polarisation des Himmelslichtes als orientierender Faktor bei den Tänzen der Bienen," K. v. Frisch, Experientia 5, 142–148 (1949), https://doi.org/10.1007/BF02174424. The short video *Show Me Some Science! Polarization of the Sky*, by the CNS Little Shop of Physics of Colorado State University, shows the pattern and how bees use it, with words on the screen and music, https://www.youtube.com/watch?v=TP5JOfrPguQ.

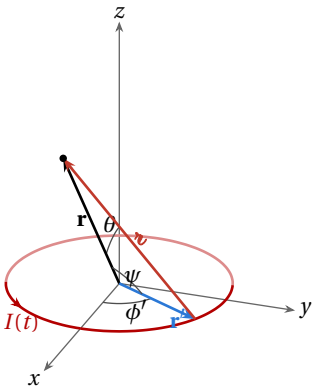


Figure 28.11 A loop of radius b in the xy plane carries the current $I(t)$. The field point is placed above the x axis, at the angle θ from the z axis. A point of the loop at the angle ϕ' has the position $\mathbf{r}'$ (blue), and $\boldsymbol{\ell}$ (red) runs from it to the field point; ψ is the angle between $\mathbf{r}$ and $\mathbf{r}'$.

☞ *Caution*: At the field point $\hat{\boldsymbol{\phi}} = \hat{\mathbf{y}}$, and it is tempting to pull $\hat{\mathbf{y}}$ out of the integral. That loses the factor $\cos\phi'$ that gives the whole answer.

The loop is small, $b \ll r$ and $b \ll c/\omega$, and the same two approximations as for the electric dipole apply, with $b\sin\theta\cos\phi'$ in the place of $(d/2)\cos\theta$:

$$\frac{1}{\ell} \cong \frac{1}{r}\left(1 + \frac{b}{r}\sin\theta\cos\phi'\right),$$
$$\cos\omega\left(t - \frac{\ell}{c}\right) \cong \cos\omega\left(t - \frac{r}{c}\right) - \frac{\omega b}{c}\sin\theta\cos\phi'\,\sin\omega\left(t - \frac{r}{c}\right).$$

To first order in b, their product is

$$\frac{\cos\omega(t-\ell/c)}{\ell} \cong \frac{1}{r}\cos\omega\left(t-\frac{r}{c}\right) + \frac{b\sin\theta\cos\phi'}{r}\left[\frac{1}{r}\cos\omega\left(t-\frac{r}{c}\right) - \frac{\omega}{c}\sin\omega\left(t-\frac{r}{c}\right)\right].$$

Multiplied by $-\sin\phi'\,\hat{\mathbf{x}} + \cos\phi'\,\hat{\mathbf{y}}$ and integrated from 0 to 2π, this gives four integrals:

$$\int_0^{2\pi}\sin\phi'\,d\phi' = 0, \qquad \int_0^{2\pi}\sin\phi'\cos\phi'\,d\phi' = 0,$$
$$\int_0^{2\pi}\cos\phi'\,d\phi' = 0, \qquad \int_0^{2\pi}\cos^2\phi'\,d\phi' = \pi.$$

Only the last survives, and with $I_0\pi b^2 = m_0$,

$$\mathbf{A} = \frac{\mu_0 m_0}{4\pi}\frac{\sin\theta}{r}\left[\frac{1}{r}\cos\omega\left(t-\frac{r}{c}\right) - \frac{\omega}{c}\sin\omega\left(t-\frac{r}{c}\right)\right]\hat{\mathbf{y}}.$$

At the field point, above the x axis (Figure 28.11), $\hat{\mathbf{y}}$ is $\hat{\boldsymbol{\phi}}$, and since the loop looks the same from every direction around the z axis, $\mathbf{A}$ points along $\hat{\boldsymbol{\phi}}$ at every field point:

$$\mathbf{A}(r,\theta,t) = \frac{\mu_0 m_0}{4\pi}\frac{\sin\theta}{r}\left[\frac{1}{r}\cos\omega\left(t-\frac{r}{c}\right) - \frac{\omega}{c}\sin\omega\left(t-\frac{r}{c}\right)\right]\hat{\boldsymbol{\phi}}. \tag{28.17}$$

As for the electric dipole, the first term is the static dipole's vector potential (12.25) with the dipole moment at the retarded time of the origin, and the second term is new. In the radiation zone only the second term counts:

$$\boxed{\mathbf{A}(r,\theta,t) = -\frac{\mu_0 m_0 \omega}{4\pi c}\frac{\sin\theta}{r}\sin\omega\left(t-\frac{r}{c}\right)\hat{\boldsymbol{\phi}}}. \tag{28.18}$$

The third step gives the fields. Since $V = 0$, the electric field is $-\partial\mathbf{A}/\partial t$. For the magnetic field, $\mathbf{A}$ has only a ϕ component, so the curl in spherical coordinates, with the terms that vanish crossed out, is

$$\nabla\times\mathbf{A} = \frac{1}{r\sin\theta}\left(\frac{\partial(\sin\theta\,A_\phi)}{\partial\theta} - \cancel{\frac{\partial A_\theta}{\partial\phi}}\right)\hat{\mathbf{r}} + \frac{1}{r}\left(\cancel{\frac{1}{\sin\theta}\frac{\partial A_r}{\partial\phi}} - \frac{\partial(r\,A_\phi)}{\partial r}\right)\hat{\boldsymbol{\theta}} + \frac{1}{r}\left(\cancel{\frac{\partial(r\,A_\theta)}{\partial r}} - \cancel{\frac{\partial A_r}{\partial\theta}}\right)\hat{\boldsymbol{\phi}}.$$

The $\hat{\mathbf{r}}$ term falls off as $1/r^2$, and the $\hat{\boldsymbol{\theta}}$ term, with rA_ϕ differentiated through the sine, gives the $1/r$ part. So

$$\boxed{\mathbf{E} = \frac{\mu_0 m_0 \omega^2}{4\pi c}\frac{\sin\theta}{r}\cos\omega\left(t-\frac{r}{c}\right)\hat{\boldsymbol{\phi}}, \qquad \mathbf{B} = -\frac{\mu_0 m_0 \omega^2}{4\pi c^2}\frac{\sin\theta}{r}\cos\omega\left(t-\frac{r}{c}\right)\hat{\boldsymbol{\theta}}}. \tag{28.19}$$

Again $\mathbf{E}$ and $\mathbf{B}$ are perpendicular to each other and to $\hat{\mathbf{r}}$, and $B = E/c$, but now the electric field circles the z axis and the magnetic field lies along the meridians: the roles of the two fields are exchanged.

The last two steps go as before. With $\hat{\boldsymbol{\phi}} \times \hat{\boldsymbol{\theta}} = -\hat{\mathbf{r}}$,

$$\langle \mathbf{S} \rangle = \frac{\mu_0 m_0^2 \omega^4}{32\pi^2 c^3}\frac{\sin^2\theta}{r^2}\hat{\mathbf{r}},$$

the same doughnut as before, and

$$\boxed{\langle P \rangle = \frac{\mu_0 m_0^2 \omega^4}{12\pi c^3}}. \tag{28.20}$$

Every formula of this section can be read off from the electric dipole's by replacing p_0 with m_0/c, $\mathbf{E}$ with $c\mathbf{B}$, and $c\mathbf{B}$ with $-\mathbf{E}$. Exercise P28.10 shows why, with the parallel equations of Chapter 14.

How does a magnetic dipole antenna compare with an electric one? By (28.11) and (28.20) the ratio of their powers is $(m_0/p_0 c)^2$. Take a loop of radius b and a two-sphere antenna whose wire has the same length, πb, and the same current amplitude I_0. Then $m_0 = I_0 \pi b^2$ and $p_0 = q_0 d = (I_0/\omega)\pi b$, so

$$\frac{\langle P \rangle_{\text{loop}}}{\langle P \rangle_{\text{dipole}}} = \left(\frac{m_0}{p_0 c}\right)^2 = \left(\frac{\omega b}{c}\right)^2,$$

which is small, since the loop is small compared with c/ω. For sources of comparable size and current, the electric dipole radiates much more. The radiation resistance of a loop is the subject of Exercise P28.9.

28.4 Radiation from an arbitrary small source: the multipole expansion

The antennas of Sections 28.2 and 28.3 each had one moment that changed in time, an electric dipole moment or a magnetic one, and each radiated according to that moment. Most small sources are not so simple: an atom or a molecule, a short antenna of some irregular shape, a spinning magnet. This section treats any source that is small compared with c/ω, whatever its charges and currents do. As in the multipole expansion of the static potential (Section 8.3), the source is described by its moments, and its radiation is a sum of terms in a definite order: the electric dipole term first, then the magnetic dipole and electric quadrupole terms together, and so on (Figure 28.12). In the static expansion each term falls off one power of r faster than the one before. Here every term falls off as $1/r$, and what puts them in order is how small the source is compared with c/ω.

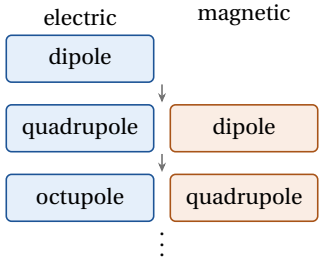


Figure 28.12 The order of the terms of the multipole expansion of the radiation of a small source. Each row is one order: the electric dipole alone, then the electric quadrupole and the magnetic dipole together, then the electric octupole and the magnetic quadrupole, and so on.

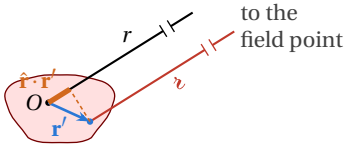


Figure 28.13 A small source (red) and a far field point in the direction $\hat{\mathbf{r}}$. The distance ϱ from a point $\mathbf{r}'$ of the source is shorter than the distance r from the origin by $\hat{\mathbf{r}}\cdot\mathbf{r}'$ (orange), when the field point is far away.

28.4.1 The fields far from a small source

Put the origin inside the source and the field point far away, in the radiation zone. Only the parts of the fields that fall off as $1/r$ matter for the radiation, so in the retarded vector potential (24.12) the distance ϱ in the denominator can be replaced by r. In the retarded time it cannot be replaced so simply. By the law of cosines, as in Section 28.2.1, $\varrho \cong r - \hat{\mathbf{r}}\cdot\mathbf{r}'$ when $r' \ll r$ (Figure 28.13), so

$$t_r = t - \frac{\varrho}{c} \cong t_0 + \frac{\hat{\mathbf{r}}\cdot\mathbf{r}'}{c}, \qquad t_0 = t - \frac{r}{c},$$

where t_0 is again the retarded time of the origin. The term $\hat{\mathbf{r}}\cdot\mathbf{r}'/c$ is the difference between the retarded times of a point of the source and of the origin. It is small, but it cannot be dropped: in Section 28.2.1 this difference is what produced the part of the potential that falls off as $1/r$. So, far away,

$$\mathbf{A}(\mathbf{r},t) \cong \frac{\mu_0}{4\pi r}\int \mathbf{J}\left(\mathbf{r}', t_0 + \frac{\hat{\mathbf{r}}\cdot\mathbf{r}'}{c}\right) d\tau'. \tag{28.21}$$

In the radiation zone the fields can be found from $\mathbf{A}$ alone. Each component of (28.21) depends on the position of the field point in three ways: through the factor $1/r$, through the direction $\hat{\mathbf{r}}$, and through $t_0 = t - r/c$. Differentiating the first two with respect to x, y, or z gives terms that fall off as $1/r^2$. Only the dependence on t_0 gives a term that falls off as $1/r$, and since $\partial t_0/\partial x = -x/rc = -\hat{r}_x/c$, and likewise for y and z, to order $1/r$ the operator ∇ acting on $\mathbf{A}$ has the same effect as $-(\hat{\mathbf{r}}/c)\,\partial/\partial t$. So

$$\nabla\times\mathbf{A} \cong -\frac{1}{c}\hat{\mathbf{r}}\times\frac{\partial\mathbf{A}}{\partial t}, \qquad \nabla\cdot\mathbf{A} \cong -\frac{1}{c}\hat{\mathbf{r}}\cdot\frac{\partial\mathbf{A}}{\partial t}.$$

The retarded potentials satisfy the Lorenz condition (24.3), $\nabla\cdot\mathbf{A} = -(1/c^2)\,\partial V/\partial t$, so $\partial V/\partial t = c\,\hat{\mathbf{r}}\cdot\partial\mathbf{A}/\partial t$. Apart from a part that does not change in time, and so does not radiate, $V = c\,\hat{\mathbf{r}}\cdot\mathbf{A}$ in the radiation zone. (For the antenna of Section 28.2, $c\,\hat{\mathbf{r}}\cdot\mathbf{A}$ with $\mathbf{A}$ from (28.7) is exactly (28.6).) To order $1/r$, then, $-\nabla V = (\hat{\mathbf{r}}/c)\,\partial V/\partial t = \hat{\mathbf{r}}\,(\hat{\mathbf{r}}\cdot\partial\mathbf{A}/\partial t)$, and by the BAC–CAB rule (1.49)

$$\mathbf{E} = -\nabla V - \frac{\partial\mathbf{A}}{\partial t} \cong \hat{\mathbf{r}}\left(\hat{\mathbf{r}}\cdot\frac{\partial\mathbf{A}}{\partial t}\right) - \frac{\partial\mathbf{A}}{\partial t} = \hat{\mathbf{r}}\times\left(\hat{\mathbf{r}}\times\frac{\partial\mathbf{A}}{\partial t}\right).$$

So in the radiation zone

$$\boxed{\mathbf{B} = -\frac{1}{c}\hat{\mathbf{r}}\times\frac{\partial\mathbf{A}}{\partial t}, \qquad \mathbf{E} = c\mathbf{B}\times\hat{\mathbf{r}}} \,. \tag{28.22}$$

Only the part of $\partial\mathbf{A}/\partial t$ perpendicular to $\hat{\mathbf{r}}$ matters: the part along $\hat{\mathbf{r}}$ drops out of the cross product, and it cancels between $-\nabla V$ and $-\partial\mathbf{A}/\partial t$. The fields are perpendicular to each other and to $\hat{\mathbf{r}}$, with $E = cB$, as they were for both dipoles.

If the source is small compared with c/ω, its current density changes little during the time $\hat{\mathbf{r}}\cdot\mathbf{r}'/c$, which is never longer than the time light takes to cross the source. To lowest order, then, that time can be dropped from (28.21), and

$$\mathbf{A} \cong \frac{\mu_0}{4\pi r}\int \mathbf{J}(\mathbf{r}', t_0)\, d\tau' = \frac{\mu_0}{4\pi r}\,\dot{\mathbf{p}}(t_0),$$

since the integral of the current density of any localized source is the rate of change of its electric dipole moment, $\int \mathbf{J}\, d\tau' = \dot{\mathbf{p}}$ (Exercise P28.11). By (28.22),

$$\mathbf{B} = -\frac{\mu_0}{4\pi c r}\,\hat{\mathbf{r}}\times\ddot{\mathbf{p}}, \qquad \mathbf{E} = \frac{\mu_0}{4\pi r}\,\hat{\mathbf{r}}\times\left(\hat{\mathbf{r}}\times\ddot{\mathbf{p}}\right), \tag{28.23}$$

with $\ddot{\mathbf{p}}$ evaluated at t_0. These are the fields of **electric dipole radiation**. Since $\mathbf{B}$ is perpendicular to $\hat{\mathbf{r}}$, the Poynting vector is $\mathbf{S} = \mathbf{E}\times\mathbf{B}/\mu_0 = (c/\mu_0)B^2\,\hat{\mathbf{r}}$, and with $|\hat{\mathbf{r}}\times\ddot{\mathbf{p}}| = \ddot{p}\sin\theta$, where θ is now the angle between $\hat{\mathbf{r}}$ and $\ddot{\mathbf{p}}$,

$$\mathbf{S} = \frac{\mu_0\ddot{p}^2}{16\pi^2 c}\,\frac{\sin^2\theta}{r^2}\,\hat{\mathbf{r}}, \qquad \frac{dP}{d\Omega} = r^2\,\mathbf{S}\cdot\hat{\mathbf{r}} = \frac{\mu_0\ddot{p}^2}{16\pi^2 c}\sin^2\theta.$$

The integral of $\sin^2\theta$ over all directions is $\int_0^{2\pi}\int_0^{\pi}\sin^3\theta\, d\theta\, d\phi = 8\pi/3$, so

$$\boxed{P = \frac{\mu_0\ddot{p}^2}{6\pi c}}. \tag{28.24}$$

This is the power at each instant, for any time dependence, with $\ddot{p}$ evaluated at the retarded time of the origin. A source whose motion contains several frequencies must satisfy the inequalities (28.2) for all of them: the source must be small compared with c/ω for the highest frequency, and the field point far compared with c/ω for the lowest. For the antenna of Section 28.2, $\mathbf{p} = p_0\cos\omega t\,\hat{\mathbf{z}}$, so $\ddot{p}^2 = p_0^2\omega^4\cos^2\omega t_0$, whose average over a cycle is $p_0^2\omega^4/2$, and (28.24) gives back (28.11).

The simplest small source of all is a single charge q that moves slowly near the origin (Figure 28.14). At the position $\mathbf{w}(t)$ it has the dipole moment $\mathbf{p} = q\mathbf{w}$, so $\ddot{\mathbf{p}} = q\mathbf{a}$, and (28.24) becomes $P = \mu_0 q^2 a^2/6\pi c$, which is Larmor's formula (27.15).[5] The accelerated charge of Chapter 27 is the simplest case of electric dipole radiation.

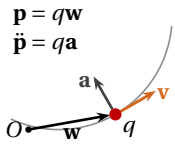


Figure 28.14 A charge q moving slowly near the origin. Its dipole moment is q times its position $\mathbf{w}$, so the second derivative of its dipole moment is q times its acceleration.

The expansion of $\mathbf{A}$ begins with the electric dipole moment. The expansion of V would begin one step earlier, with the monopole term $Q/4\pi\varepsilon_0 r$, where Q is the total charge of the source. But charge is conserved, so Q never changes, and that term is static and radiates nothing: there is no monopole radiation. A sphere of charge that swells and shrinks, keeping its total charge and its spherical symmetry, sends out no radiation at all (Exercise P28.13). Sound is different. A

[5] In Section 27.1 the same retarded potentials gave a moving point charge the factor $1/(1-\hat{\boldsymbol{\varepsilon}}\cdot\mathbf{v}/c)$, because each part of the charge is counted at its own retarded time. The expansion here takes the whole source at the one time t_0 and puts the differences in retarded time into the terms that contain time derivatives, so no such factor appears in it. For a single slow charge, the first-order part of the factor is carried by the next term of the expansion, in Section 28.4.2 (Exercise P28.12).

pulsating sphere in air is the simplest source of sound, since the volume of air that it pushes aside changes; the amount of charge inside a sphere drawn around a source of electromagnetic waves cannot change. Many sources have no changing electric dipole moment either. For them the leading term vanishes, and the next term of the expansion decides what they radiate.

28.4.2 The next order

To go further, keep the next term of the Taylor expansion of the current density in (28.21) about the time t_0:

$$\mathbf{J}\left(\mathbf{r}', t_0 + \frac{\hat{\mathbf{r}}\cdot\mathbf{r}'}{c}\right) \cong \mathbf{J}(\mathbf{r}', t_0) + \frac{\hat{\mathbf{r}}\cdot\mathbf{r}'}{c}\dot{\mathbf{J}}(\mathbf{r}', t_0).$$

Then

$$\mathbf{A} \cong \frac{\mu_0}{4\pi r}\left[\dot{\mathbf{p}} + \frac{1}{c}\frac{d}{dt}\int (\hat{\mathbf{r}}\cdot\mathbf{r}')\,\mathbf{J}\,d\tau'\right],$$

with everything evaluated at t_0. The integral in the second term holds two more moments of the source, its magnetic dipole moment $\mathbf{m}$ and its electric quadrupole moment $\overleftrightarrow{\mathbf{Q}}$:

$$\int (\hat{\mathbf{r}}\cdot\mathbf{r}')\,\mathbf{J}\,d\tau' = \mathbf{m}\times\hat{\mathbf{r}} + \frac{1}{3}\frac{d}{dt}\left(\overleftrightarrow{\mathbf{Q}}\cdot\hat{\mathbf{r}}\right) + (\text{a part along } \hat{\mathbf{r}}),$$

where $\overleftrightarrow{\mathbf{Q}}\cdot\hat{\mathbf{r}}$ is the vector whose components are $\sum_j Q_{ij}\hat{r}_j$: the matrix (8.19) times the column of the components of $\hat{\mathbf{r}}$. The derivation comes next.

! **Two parts of an array.** Any array M_{ij} is a part that changes sign when i and j are exchanged plus a part that does not: $M_{ij} = \frac{1}{2}(M_{ij} - M_{ji}) + \frac{1}{2}(M_{ij} + M_{ji})$.

The next order of the expansion

The i component of the integral is $\sum_j \hat{r}_j \int x'_j J_i\,d\tau'$. Split $\int x'_j J_i\,d\tau'$ into a part that changes sign when i and j are exchanged and a part that does not:

$$\int x'_j J_i\,d\tau' = \frac{1}{2}\int \left(x'_j J_i - x'_i J_j\right) d\tau' + \frac{1}{2}\int \left(x'_j J_i + x'_i J_j\right) d\tau'.$$

The antisymmetric part. Summed with $\hat{r}_j$, it gives

$$\frac{1}{2}\int \left[(\hat{\mathbf{r}}\cdot\mathbf{r}')\,J_i - x'_i\,(\hat{\mathbf{r}}\cdot\mathbf{J})\right] d\tau',$$

which is the i component of $\frac{1}{2}\int \hat{\mathbf{r}}\times\left(\mathbf{J}\times\mathbf{r}'\right)d\tau'$, by the BAC–CAB rule. Since $\mathbf{J}\times\mathbf{r}' = -\mathbf{r}'\times\mathbf{J}$, this is $-\hat{\mathbf{r}}\times\mathbf{m} = \mathbf{m}\times\hat{\mathbf{r}}$, where $\mathbf{m} = \frac{1}{2}\int \mathbf{r}'\times\mathbf{J}\,d\tau'$ is the magnetic dipole moment (12.24).

The symmetric part. By the product rule, $\nabla'\cdot(x'_i x'_j\mathbf{J}) = x'_i x'_j\,\nabla'\cdot\mathbf{J} + x'_j J_i + x'_i J_j$. Integrate both sides over a volume that contains the whole source. By the divergence theorem the left side becomes the flux of $x'_i x'_j\mathbf{J}$ through the volume's surface, which is zero, since no current crosses it. By the continuity equation (20.5), $\nabla'\cdot\mathbf{J} = -\partial\rho/\partial t$. So

$$\int \left(x'_j J_i + x'_i J_j\right) d\tau' = \frac{d}{dt}\int x'_i x'_j\,\rho\,d\tau'.$$

By the definition (8.17), $\int x'_i x'_j \rho\, d\tau' = \frac{2}{3} Q_{ij} + \frac{1}{3}\delta_{ij}\int r'^2 \rho\, d\tau'$. Summed with $\frac{1}{2}\hat{r}_j$, the symmetric part is therefore

$$\frac{1}{3}\sum_j \dot{Q}_{ij}\hat{r}_j + \frac{1}{6}\hat{r}_i \frac{d}{dt}\int r'^2 \rho\, d\tau'.$$

The first term is the i component of $\frac{1}{3}\,d(\overleftrightarrow{\mathbf{Q}}\cdot\hat{\mathbf{r}})/dt$. The second points along $\hat{\mathbf{r}}$, and by (28.22) a part of $\mathbf{A}$ along $\hat{\mathbf{r}}$ produces no radiation fields, so it can be dropped.

So, to this order and apart from a part along $\hat{\mathbf{r}}$,

$$\mathbf{A} \cong \frac{\mu_0}{4\pi r}\left[\dot{\mathbf{p}} + \frac{1}{c}\dot{\mathbf{m}}\times\hat{\mathbf{r}} + \frac{1}{3c}\frac{d^2}{dt^2}\left(\overleftrightarrow{\mathbf{Q}}\cdot\hat{\mathbf{r}}\right)\right],$$

and by (28.22)

$$\boxed{\mathbf{B} = -\frac{\mu_0}{4\pi c r}\hat{\mathbf{r}}\times\left[\ddot{\mathbf{p}} + \frac{1}{c}\ddot{\mathbf{m}}\times\hat{\mathbf{r}} + \frac{1}{3c}\frac{d^3}{dt^3}\left(\overleftrightarrow{\mathbf{Q}}\cdot\hat{\mathbf{r}}\right)\right]}, \tag{28.25}$$

with the moments evaluated at t_0, and $\mathbf{E} = c\mathbf{B}\times\hat{\mathbf{r}}$. The three terms are the fields of electric dipole radiation, **magnetic dipole radiation**, and **electric quadrupole radiation**, often called E1, M1, and E2 radiation for short. The magnetic dipole term gives back the fields of the loop of Section 28.3. For $\mathbf{m} = m_0\cos\omega t\,\hat{\mathbf{z}}$, the vector $\hat{\mathbf{r}}\times(\ddot{\mathbf{m}}\times\hat{\mathbf{r}})$ is the part of $\ddot{\mathbf{m}}$ perpendicular to $\hat{\mathbf{r}}$,

$$-m_0\omega^2\cos\omega t_0\,(\hat{\mathbf{z}} - \cos\theta\,\hat{\mathbf{r}}) = m_0\omega^2\cos\omega t_0\,\sin\theta\,\hat{\boldsymbol{\theta}},$$

and (28.25) becomes the magnetic field of (28.19).

Each term gives a power by the same steps as the electric dipole term. The Poynting vector contains the square of the bracket in (28.25), and so cross terms between different terms, which shape the radiation pattern. But each cross term integrates to zero over the large sphere, and the powers simply add (Exercise P28.16):

$$\boxed{P_{\mathrm{E1}} = \frac{\mu_0\ddot{p}^2}{6\pi c}, \qquad P_{\mathrm{M1}} = \frac{\mu_0\ddot{m}^2}{6\pi c^3}, \qquad P_{\mathrm{E2}} = \frac{\mu_0}{180\pi c^3}\sum_{i,j}\dddot{Q}_{ij}^2}. \tag{28.26}$$

The electric quadrupole power needs one more integral over directions, $\oint|\hat{\mathbf{r}}\times(\overleftrightarrow{\mathbf{Q}}\cdot\hat{\mathbf{r}})|^2\,d\Omega = (4\pi/5)\sum_{i,j}Q_{ij}^2$, which holds because $\overleftrightarrow{\mathbf{Q}}$ is symmetric and its diagonal components add to zero (Exercise P8.9). For moments that oscillate as $\cos\omega t$, with the amplitudes p_0, m_0, and $Q_{0,ij}$, the average of each squared derivative over a cycle is half its largest value, and the average powers are

$$\langle P_{\mathrm{E1}}\rangle = \frac{\mu_0 p_0^2\omega^4}{12\pi c}, \qquad \langle P_{\mathrm{M1}}\rangle = \frac{\mu_0 m_0^2\omega^4}{12\pi c^3}, \qquad \langle P_{\mathrm{E2}}\rangle = \frac{\mu_0\omega^6}{360\pi c^3}\sum_{i,j}Q_{0,ij}^2. \tag{28.27}$$

! **The other convention.** With $Q_{ij}^{\mathrm{alt}} = 2Q_{ij}$,

$$P_{\mathrm{E2}} = \frac{\mu_0}{720\pi c^3}\sum_{i,j}\left(\ddot{Q}_{ij}^{\mathrm{alt}}\right)^2,$$

and the average for a moment that oscillates as $\cos\omega t$ has 1440π in place of 720π and ω^6 times the squared amplitudes.

With the other convention for the quadrupole moment, $Q_{ij}^{\text{alt}} = 2Q_{ij}$ (Section 8.5.3), the coefficient of P_{E2} becomes $1/720\pi c^3$, and that of its average $1/1440\pi c^3$, the form found in many other references. The electric and magnetic dipole terms each have the doughnut pattern $\sin^2\theta$, about the direction of $\ddot{\mathbf{p}}$ or $\ddot{\mathbf{m}}$. The pattern of the electric quadrupole term depends on the shape of $\overleftrightarrow{\mathbf{Q}}$ (Example 28.2).

The electric dipole term is the leading term of the expansion. The magnetic dipole and electric quadrupole terms come together at the next order, from the same term of the Taylor expansion, so they are typically about the same size (Tutorial 28.1 demonstrates this); then come the magnetic quadrupole and the electric octupole, and so on.[6]

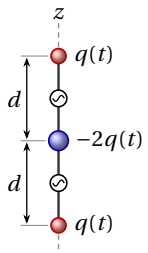
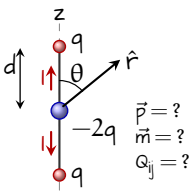


Figure 28.15 Example 28.2: the charges $q(t)$ at $z = \pm d$ and $-2q(t)$ at the origin, joined by wires.

Example 28.2

Two antennas like the one of Figure 28.2 are placed end to end along the z axis and driven in opposite phase, so that the outer spheres, at $z = \pm d$, both carry the charge $q(t) = q_0 \cos\omega t$, and the two inner spheres, merged into one at the origin, carry $-2q(t)$. The source is small, $d \ll c/\omega$. What are its electric dipole, magnetic dipole, and electric quadrupole moments? In which directions does it radiate, and what is the average radiated power?

Solution:



The first step is a sketch of the source: the three charges on the z axis, the currents in the wires between them, and a far direction at the angle θ from the axis. The moments come from the positions of the charges.

Electric dipole moment: with the charges at $z = d$, $-d$, and 0,

$$\mathbf{p} = q\,d\,\hat{\mathbf{z}} + q\,(-d)\,\hat{\mathbf{z}} + (-2q)(0) = 0$$

at every instant. The two antennas have opposite dipole moments, and they cancel: there is no electric dipole radiation.

Magnetic dipole moment: the currents flow along the z axis, where $\mathbf{r}'$ is parallel to $\mathbf{J}$, so $\mathbf{r}' \times \mathbf{J} = 0$ and, by (12.24), $\mathbf{m} = 0$.

Quadrupole moment: by (8.17) for point charges, a charge q at $z = d$ on the axis contributes $q(\frac{3}{2}d^2 - \frac{1}{2}d^2) = qd^2$ to Q_{zz} and $-\frac{1}{2}qd^2$ to Q_{xx} and to Q_{yy}, and nothing off the diagonal. The charge at $z = -d$ contributes the same, and the charge at the origin

[6] The gravitational waves of a small source begin one order later, with its mass quadrupole moment. The gravitational counterpart of the electric dipole moment is the total mass times the position of the center of mass, whose rate of change is the total momentum, which is constant; and the counterpart of the magnetic dipole moment is the angular momentum, which is conserved too.

nothing, so

$$\overleftrightarrow{\mathbf{Q}} = q(t)\,d^2\begin{pmatrix}-1 & 0 & 0\\ 0 & -1 & 0\\ 0 & 0 & 2\end{pmatrix}.$$

For the direction $\hat{\mathbf{r}} = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$,

$$\overleftrightarrow{\mathbf{Q}}\cdot\hat{\mathbf{r}} = qd^2\,(-\hat{\mathbf{r}} + 3\cos\theta\,\hat{\mathbf{z}}), \qquad \hat{\mathbf{r}}\times\left(\overleftrightarrow{\mathbf{Q}}\cdot\hat{\mathbf{r}}\right) = 3qd^2\cos\theta\,\hat{\mathbf{r}}\times\hat{\mathbf{z}} = -3qd^2\sin\theta\cos\theta\,\hat{\boldsymbol{\phi}}.$$

By (28.25) the radiation field is

$$\mathbf{B} = \frac{\mu_0 d^2\,\dddot{q}}{4\pi c^2 r}\sin\theta\cos\theta\,\hat{\boldsymbol{\phi}}, \qquad \dddot{q} = q_0\omega^3\sin\omega t_0,$$

and with $\langle \dddot{q}^{\,2}\rangle = q_0^2\omega^6/2$,

$$\frac{dP}{d\Omega} = \frac{c}{\mu_0}r^2\langle B^2\rangle = \frac{\mu_0 q_0^2 d^4\omega^6}{32\pi^2 c^3}\sin^2\theta\cos^2\theta.$$

The pattern has four lobes, at 45° and 135° from the axis, with no radiation along the axis and none in the xy plane (Figure 28.16). With $\int_0^\pi \sin^3\theta\cos^2\theta\,d\theta = 4/15$,

$$\langle P\rangle = \frac{\mu_0 q_0^2 d^4\omega^6}{32\pi^2 c^3}(2\pi)\left(\frac{4}{15}\right) = \frac{\mu_0 q_0^2 d^4\omega^6}{60\pi c^3}.$$

The same comes from (28.27), with $\sum_{i,j} Q_{0,ij}^2 = q_0^2 d^4(1+1+4)$.

Check: why four lobes? The two antennas oscillate in opposite phase. In the xy plane their waves travel equal distances, so they arrive in opposite phase and cancel, and along the axis neither antenna sends anything. In between, one antenna is closer than the other by $d\cos\theta$, the cancellation is not complete, and the radiation is strongest halfway between.

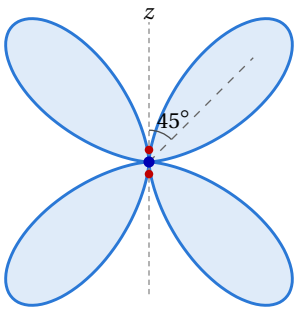


Figure 28.16 The radiation pattern of the source of Example 28.2, $dP/d\Omega \propto \sin^2\theta\cos^2\theta$, in a plane through the z axis, with the three charges on the axis. Turned about the axis, it makes two cones.

The rest of this section is about sources whose charges move in circles. Tutorial 28.1 compares the sizes of the three kinds of radiation and applies them to three arrangements of charges on circles.

Tutorial 28.1 Multipole radiation Printable version

Background

Section 28.4 found the first three terms of the radiation of a small source: the powers (28.26) of electric dipole, magnetic dipole, and electric quadrupole radiation at each instant, and their averages (28.27) for moments that oscillate as $\cos\omega t$. The electric and magnetic dipole terms each have the angular dependence $\sin^2\theta$, where θ is measured from $\ddot{\mathbf{p}}$ or $\ddot{\mathbf{m}}$. In this tutorial you compare the sizes of the three terms, and then you apply them to charges that move in circles.

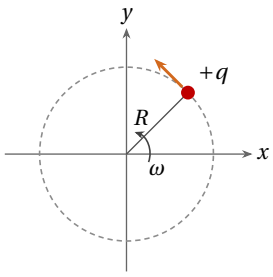


Figure 28.17 Tutorial 28.1, Part 2: a charge q moving counterclockwise around a circle of radius R in the xy plane.

Part 1 — Relative magnitudes

First, compare the sizes of the three kinds of radiation. Picture a set of charges and currents that change in time but return to their original configuration every period $T = 2\pi/\omega$. Suppose that the electric dipole, the magnetic dipole, and the electric quadrupole all have the same characteristic length d, time T, and charge q. A *characteristic* length, time, or charge is a rough size for that quantity: d is about the size of the source, T is the period, and q is about how much charge moves. Answers in terms of these scales are estimates that show how a quantity depends on d, T, and q; for example, the charges move at speeds of about d/T.

(a) In terms of these characteristic scales, what is the electric dipole moment? (Recall that $\mathbf{p} = \sum_k q_k \mathbf{r}_k$.) What is the average power of the electric dipole radiation?

(b) In terms of the characteristic scales, what is the magnetic dipole moment? (Recall that $\mathbf{m} = I\mathbf{a}$, where in terms of the characteristic scales the current is a charge divided by a time and the area is a length squared.) What is the average power of the magnetic dipole radiation?

(c) Now compare. Use your answers to (a) and (b) to find $\langle P_{\mathrm{M1}}\rangle/\langle P_{\mathrm{E1}}\rangle$, and put it in terms of the characteristic speed $v = d/T$. Under what conditions is the magnetic dipole radiation much weaker than the electric dipole radiation?

(d) In terms of the characteristic scales, what is the quadrupole moment? (Just give one representative component, such as Q_{xy}. Recall that for point charges $Q_{ij} = \sum_k q_k(\frac{3}{2} r_{k,i} r_{k,j} - \frac{1}{2} r_k^2 \delta_{ij})$, from Section 8.5.) Roughly what is the average power of the electric quadrupole radiation?

(e) Use your answers to (a) and (d) to find $\langle P_{\mathrm{E2}}\rangle/\langle P_{\mathrm{E1}}\rangle$, also in terms of $v = d/T$. When is it small?

(f) What do your last two answers say about the relative sizes of P_{M1} and P_{E2}, and about how to use the multipole expansion of radiation? For comparison, in the multipole expansion of the electric potential (Section 8.3) the terms come in the order monopole, dipole, quadrupole, and so on, and far away you can generally stop at the first one that is not zero.

Part 2 — A point charge moving in a circle

Now consider three arrangements of charges that move in circles. First, a single charge q moves around a circle of radius R at the constant speed v, with $\omega = v/R$, so that its position is $\mathbf{r}' = R\cos\omega t\,\hat{\mathbf{x}} + R\sin\omega t\,\hat{\mathbf{y}}$ (Figure 28.17).

(a) Why might it be useful to describe the radiation of charges that move in circles with the multipole expansion, rather than, say, with Larmor's formula for each charge?

(b) Find the electric dipole moment $\mathbf{p}$ of the charge. Does it have a harmonic time dependence?

(c) Find the magnetic dipole moment $\mathbf{m}$ of the charge. Take its current to be q divided by the period of its motion. Does $\mathbf{m}$ have a harmonic time dependence?

(d) Does this system emit electric dipole radiation? If so, how much power? Does it emit magnetic dipole radiation? If so, how much? Which kind dominates?

Part 3 — Two opposite charges on a rigid rotor

Now consider two opposite charges, $+q$ and $-q$, fixed to the ends of a rigid rod of length $2R$ that rotates about its center at the angular velocity ω (Figure 28.18). This arrangement is called a rigid rotor.

(a) Find the electric dipole moment. Does it have a harmonic time dependence?

(b) Find the magnetic dipole moment. Does it have a harmonic time dependence?

(c) Does this system emit electric dipole radiation? If so, how much? Does it emit magnetic dipole radiation? If so, how much? Which kind dominates?

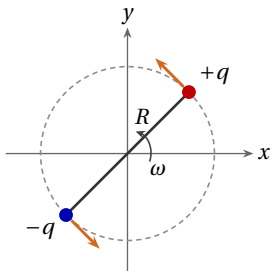


Figure 28.18 Tutorial 28.1, Part 3: charges $+q$ and $-q$ at the ends of a rigid rod of length $2R$ that rotates about its center at the angular velocity ω.

Part 4 — Two equal charges on a rigid rotor

Now put two equal charges, both $+q$, on the ends of the rotor (Figure 28.19).

(a) Find the electric dipole moment. Does it have a harmonic time dependence?

(b) Find the magnetic dipole moment. Does it have a harmonic time dependence?

(c) Does this system emit electric dipole radiation? If so, how much? Does it emit magnetic dipole radiation? If so, how much? Which kind dominates?

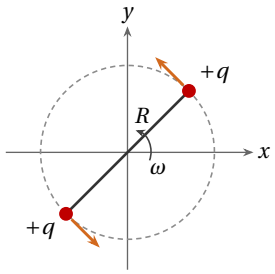


Figure 28.19 Tutorial 28.1, Part 4: two charges $+q$ at the ends of the rotor of Part 3.

Part 5 — Wrap-up

(a) Thinking about what happens physically, explain the difference between your answers to Parts 3 and 4.

(b) Think of the standard ways that physicists determine whether an answer is reasonable: limiting cases, dimensional analysis, comparison with other problems, and so forth. Apply at least one of them to one of your answers.

A spinning magnet radiates too, as long as its magnetic dipole moment is not along its axis of rotation, and the most powerful spinning magnets known are pulsars. A neutron star is the collapsed core of a star that has exploded as a supernova, with a mass somewhat larger than the Sun's packed into a sphere about 10 km in radius. Many neutron stars spin rapidly and have strong magnetic fields, and those whose beams of radio waves sweep past the Earth are seen as pulsars, which send out one pulse of radio waves in every turn.[7]

[7] Pulsars were discovered by Jocelyn Bell (now Bell Burnell) and Antony Hewish in 1967: "Observation of a rapidly pulsating radio source," A. Hewish, S. J. Bell, J. D. H. Pilkington, P. F. Scott, and R. A. Collins, Nature 217, 709–713 (1968), https://doi.org/10.1038/217709a0. Before the discovery was announced, Franco Pacini had proposed that a rapidly rotating neutron star with a strong magnetic field would radiate and could supply the energy of the Crab Nebula, and in 1968 Thomas Gold proposed that the pulsating sources are rotating neutron stars: "Energy emission from a neutron star," F. Pacini, Nature 216, 567–568 (1967), https://doi.org/10.1038/216567a0, and "Rotating neutron stars as the origin of the pulsating radio sources," T. Gold, Nature 218, 731–732 (1968), https://doi.org/10.1038/218731a0.

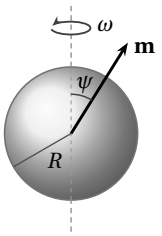
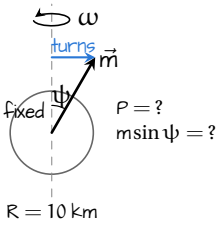


Figure 28.20 Example 28.3: a neutron star of radius R rotating at the angular velocity ω, with its magnetic dipole moment **m** at the angle ψ to the axis of rotation.

Example 28.3

The pulsar at the center of the Crab Nebula is slowing down. In September 2026 its rotation frequency was $\nu = 29.534\,\text{Hz}$, and it was decreasing at the rate $\dot{\nu} = -3.658\times10^{-10}\,\text{s}^{-2}$.[8] Model it as a sphere of radius $R = 10\,\text{km}$ with the moment of inertia $I = 1.0\times10^{38}\,\text{kg}\cdot\text{m}^2$, typical values for a neutron star, and with a magnetic dipole moment **m** fixed in the star at the angle ψ to its axis of rotation. (a) At what rate does the pulsar lose rotational energy? (b) If magnetic dipole radiation carries all of that energy away, how large is the part of **m** perpendicular to the axis, and how strong is the magnetic field at the star's surface?

Solution:



The first step is a sketch of the star, its axis of rotation, and **m** at the angle ψ, split into a part along the axis, which stays fixed, and a part perpendicular to it, which turns with the star.

(a) The rotational energy is $K = \frac{1}{2}I\omega^2$ with $\omega = 2\pi\nu$, so

$$\frac{dK}{dt} = I\omega\dot{\omega} = 4\pi^2 I\nu\dot{\nu}$$
$$= 4\pi^2\left(1.0\times10^{38}\,\text{kg}\cdot\text{m}^2\right)\left(29.534\,\text{s}^{-1}\right)\left(-3.658\times10^{-10}\,\text{s}^{-2}\right) = -4.3\times10^{31}\,\text{W}.$$

The pulsar loses rotational energy at the rate $4.3\times10^{31}\,\text{W}$.

(b) The part of **m** along the axis, $m\cos\psi$, does not change and does not radiate. The part perpendicular to the axis, of size $m\sin\psi$, turns with the star at the angular velocity ω, so its second derivative has the constant size $m\sin\psi\,\omega^2$, and by (28.26)

$$P_{\text{M1}} = \frac{\mu_0\left(m\sin\psi\right)^2\omega^4}{6\pi c^3}.$$

Set this equal to $4.3\times10^{31}\,\text{W}$, with $\omega = 2\pi\nu = 185.6\,\text{s}^{-1}$:

$$m\sin\psi = \sqrt{\frac{6\pi c^3 P_{\text{M1}}}{\mu_0\omega^4}} = 3.8\times10^{27}\,\text{A}\cdot\text{m}^2.$$

The field of a magnetic dipole at the distance R on its equator is $\mu_0 m/4\pi R^3$, by (12.26), so if **m** is nearly perpendicular to the axis, the field at the surface is about

$$B = \frac{\mu_0 m}{4\pi R^3} = \frac{\left(10^{-7}\,\text{T}\cdot\text{m/A}\right)\left(3.8\times10^{27}\,\text{A}\cdot\text{m}^2\right)}{\left(1.0\times10^4\,\text{m}\right)^3} = 3.8\times10^8\,\text{T},$$

[8] "Jodrell Bank Crab Pulsar Monthly Ephemeris," https://www.jb.man.ac.uk/pulsar/crab.html, the entry for 15 September 2026; the measurements are described in "23 years of Crab pulsar rotational history," A. G. Lyne, R. S. Pritchard, and F. Graham Smith, Mon. Not. R. Astron. Soc. 265, 1003–1012 (1993), https://doi.org/10.1093/mnras/265.4.1003.

and twice that at the magnetic poles. For a smaller angle ψ the field must be stronger still.

Check: the formulas need the source to be small compared with $c/\omega = 1.6 \times 10^6$ m, and the star's radius, 10^4 m, is. The field is the size that Table 10.1 gives for the surface of a neutron star.

! **In gauss.** Since 1 gauss = 10^{-4} T (Section 17.1.7), the field of Example 28.3 is 3.8×10^{12} gauss.

The example treats the pulsar as a magnet in empty space. A real pulsar is surrounded by plasma, which carries currents and changes how the star loses energy, but calculations that include the plasma find a power of the same order, $\mu_0 m^2 \omega^4 (1 + \sin^2\psi)/4\pi c^3$, which is not zero even when $\mathbf{m}$ lies along the axis.[9] The power that the Crab pulsar loses, about 10^5 times the power of the Sun, is what keeps the surrounding Crab Nebula shining.[10]

28.5 Numerical methods: antennas of any length

The multipole expansion needs a source that is small compared with c/ω, and many antennas are not. An antenna half a wavelength long is π times c/ω, and the towers of AM radio stations are about a quarter of a wavelength tall (Section 28.1). For such antennas we go back to (28.21), which did not assume a small source, and evaluate it numerically.

28.5.1 The far field of a given current

Equation (28.21) assumed only that the field point is far away: far compared with the size d of the source, so that $1/\mathscr{r} \cong 1/r$, and far enough that the next term of $\mathscr{r}$, of the order of d^2/r, changes the retarded time by only a small fraction of a cycle, which needs $r \gg \omega d^2/c$. It did not assume that the source is small compared with c/ω. Suppose that the current oscillates at one frequency. With the complex time dependence $e^{-i\omega t}$ for fields (Section 18.2.4), write it as $\mathbf{J}(\mathbf{r}', t) = \mathrm{Re}[\tilde{\mathbf{J}}(\mathbf{r}')\, e^{-i\omega t}]$, where the complex amplitude $\tilde{\mathbf{J}}$ can have a different phase at each point. The shift of the time in (28.21) then becomes a factor $e^{-ik\hat{\mathbf{r}}\cdot\mathbf{r}'}$, with $k = \omega/c$, and

$$\mathbf{A}(\mathbf{r}, t) = \mathrm{Re}\left[\frac{\mu_0 e^{-i\omega t_0}}{4\pi r}\int \tilde{\mathbf{J}}(\mathbf{r}')\, e^{-ik\hat{\mathbf{r}}\cdot\mathbf{r}'}\, d\tau'\right]. \tag{28.28}$$

The integral depends only on the direction $\hat{\mathbf{r}}$. The factor $e^{-ik\hat{\mathbf{r}}\cdot\mathbf{r}'}$ is the phase of the news from the point $\mathbf{r}'$, which starts toward the field point a distance $\hat{\mathbf{r}}\cdot\mathbf{r}'$ ahead of the news from the origin, and the integral adds up the news from every part of the source with its phase.

For a thin straight antenna along the z axis, $\tilde{\mathbf{J}}\, d\tau'$ becomes $\tilde{I}(z')\, dz'\, \hat{\mathbf{z}}$, where $\tilde{I}$ is the complex amplitude of the current, and $\hat{\mathbf{r}}\cdot\mathbf{r}' = z'\cos\theta$ (Figure 28.21). Then

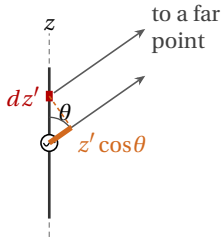


Figure 28.21 A straight antenna along the z axis, fed at its center, and the news from a piece dz' of it and from its center toward a far point at the angle θ. The news from dz' starts the distance $z'\cos\theta$ (orange) closer to the far point.

[9] "Time-dependent force-free pulsar magnetospheres: axisymmetric and oblique rotators," A. Spitkovsky, Astrophys. J. 648, L51–L54 (2006), https://doi.org/10.1086/507518; the formula

(28.28) is $\mathbf{A} = \mathrm{Re}[(\mu_0 e^{-i\omega t_0}/4\pi r)\,F(\theta)]\,\hat{\mathbf{z}}$, with

$$F(\theta) = \int \tilde{I}(z')\, e^{-ikz'\cos\theta}\, dz'.$$

By (28.22), $B = (\sin\theta/c)\,|\partial A_z/\partial t|$, since $|\hat{\mathbf{r}} \times \hat{\mathbf{z}}| = \sin\theta$. The average over a cycle of the square of $\mathrm{Re}[Xe^{-i\omega t}]$ is $\frac{1}{2}|X|^2$ (Section 18.4), so with $Z_0 = \mu_0 c$ the average power per unit solid angle, $(c/\mu_0)\,r^2\langle B^2\rangle$, is

$$\boxed{\frac{dP}{d\Omega} = \frac{Z_0 k^2}{32\pi^2}\sin^2\theta\,|F(\theta)|^2}. \tag{28.29}$$

The total power is $\langle P\rangle = 2\pi\int_0^\pi (dP/d\Omega)\sin\theta\,d\theta$, and the radiation resistance is $R = 2\langle P\rangle/|\tilde{I}_{\text{feed}}|^2$, where $\tilde{I}_{\text{feed}}$ is the current where the antenna is driven. The two integrals, over z' and over θ, seldom have closed forms, but a computer does them easily. As a check, take a uniform current $\tilde{I} = I_0$ along a short length d. Then $F = I_0 d$, the integral of $\sin^2\theta$ over all directions is $8\pi/3$, $\langle P\rangle = Z_0 k^2 I_0^2 d^2/12\pi$, and $R = Z_0 k^2 d^2/6\pi = (2\pi/3)Z_0(d/\lambda)^2$, which is (28.13).

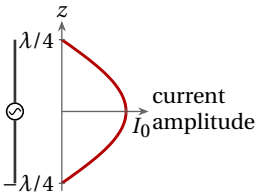
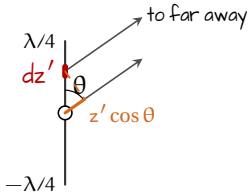


Figure 28.22 Example 28.4: the half-wave antenna, fed at its center, and the amplitude of its current along its length.

Example 28.4

A straight wire half a wavelength long lies along the z axis from $-\lambda/4$ to $\lambda/4$ and is fed at its center. Its current falls from I_0 at the center to zero at the ends as a cosine, $I(z,t) = I_0\cos kz\cos\omega t$, which is close to the current in a real thin antenna. Find its radiation pattern and its radiation resistance.

Solution:



The first step is a sketch of the antenna, its current, and a far direction at the angle θ. The news from the piece dz' at z' starts a distance $z'\cos\theta$ ahead of the news from the center.

Every point of the wire oscillates in phase, so the amplitude is real, $\tilde{I}(z') = I_0\cos kz'$. The imaginary part of $e^{-ikz'\cos\theta}$ is odd in z' and drops out of the integral, which leaves

$$F(\theta) = 2I_0\int_0^{\lambda/4}\cos kz'\cos(kz'\cos\theta)\,dz' = \frac{2I_0}{k}\,\frac{\cos\left(\frac{\pi}{2}\cos\theta\right)}{\sin^2\theta},$$

which a computer algebra system gives at once (or write the product of cosines as a sum). By (28.29),

$$\frac{dP}{d\Omega} = \frac{Z_0 I_0^2}{8\pi^2}\,\frac{\cos^2\left(\frac{\pi}{2}\cos\theta\right)}{\sin^2\theta}.$$

there is in Gaussian units.

[10] "The Crab Nebula: an astrophysical chimera," J. J. Hester, Annu. Rev. Astron. Astrophys. 46, 127–155 (2008), https://doi.org/10.1146/annurev.astro.45.051806.110608.

Like a short antenna, the half-wave antenna sends nothing along its own length and the most at right angles to it, but its pattern is a little narrower (Figure 28.23): the power per unit solid angle falls to half its largest value 39° from the plane $z = 0$, rather than 45°. The total power is

$$\langle P\rangle = 2\pi \int_0^\pi \frac{dP}{d\Omega}\sin\theta\, d\theta = \frac{Z_0 I_0^2}{4\pi}\int_0^\pi \frac{\cos^2\left(\frac{\pi}{2}\cos\theta\right)}{\sin\theta}\, d\theta.$$

This integral has no closed form in elementary functions. A numerical integration gives 1.2188, so

$$R = \frac{2\langle P\rangle}{I_0^2} = \frac{Z_0}{2\pi}(1.2188) = 73.1\,\Omega.$$

Check: doing both integrals numerically, without the closed form for $F(\theta)$, gives the same 73.1 Ω, and the same formulas gave back (28.13) for a short antenna with a uniform current.

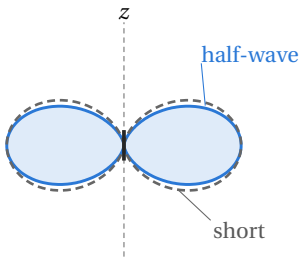


Figure 28.23 The radiation pattern of the half-wave antenna of Example 28.4, $\cos^2(\frac{\pi}{2}\cos\theta)/\sin^2\theta$ (blue), and of a short antenna, $\sin^2\theta$ (dashed), each scaled to the same largest value, in a plane through the antenna.

Half-wave and quarter-wave antennas are common. A vertical wire a quarter of a wavelength tall, fed at its base and standing on conducting ground, acts above the ground like the upper half of a half-wave antenna, because the ground supplies the image of the lower half (Section 7.2.1). It radiates only into the space above the ground, so for the same current it radiates half the power, and its radiation resistance is half, 36.5 Ω. Longer antennas have patterns with more lobes (Exercise P28.20).

28.5.2 Arrays of antennas

Antennas are often used in groups, called arrays, whose fields add. Suppose the antennas are identical and point the same way, and the nth is at the position $\mathbf{r}_n$ and is driven with the current amplitude $I_0 e^{i\alpha_n}$, so that each has the same current but its own phase α_n. By (28.28), the factor $e^{-ik\hat{\mathbf{r}}\cdot\mathbf{r}'}$ for each point of the nth antenna is $e^{-ik\hat{\mathbf{r}}\cdot\mathbf{r}_n}$ times what it would be if that antenna were at the origin, so F for the whole array is F for one antenna times a sum of phase factors:

$$F_{\text{array}}(\hat{\mathbf{r}}) = F(\theta)\sum_n e^{i(\alpha_n - k\hat{\mathbf{r}}\cdot\mathbf{r}_n)}. \tag{28.30}$$

The sum, which depends only on the positions and phases of the antennas, is called the **array factor**. With the right phases, the waves from the antennas add up in one direction and cancel in others, and changing the phases moves the beam without moving any antenna. This is how phased-array radars steer their beams (Exercises P28.21 and P28.22).

28.5.3 Tips for Mathematica and Python

Three-dimensional patterns. Mathematica's `SphericalPlot3D[f[t, p], {t, 0, Pi}, {p, 0, 2 Pi}]` draws the surface whose distance from the origin is

`f[t, p]` in the direction with the polar angle `t`, measured from the z axis as in this book, and the azimuthal angle `p`: it draws a radiation pattern directly. Matplotlib has no such function. Make a grid of θ and ϕ with `numpy.meshgrid`, compute f on it, form $x = f\sin\theta\cos\phi$, $y = f\sin\theta\sin\phi$, and $z = f\cos\theta$, and pass them to `plot_surface` on axes made with `projection='3d'`; `set_box_aspect((1, 1, 1))` keeps the shape from being stretched.

Complex integrands. Mathematica's `NIntegrate` accepts a complex integrand. SciPy's `quad` integrates real functions, so integrate the real and imaginary parts of $\tilde{I}(z')e^{-ikz'\cos\theta}$ separately, or give `quad` the option `complex_func=True`, which recent versions accept.

The ends of the θ integral. Patterns like that of Example 28.4 are $0/0$ at $\theta = 0$ and π, although they have finite limits. Adaptive integrators such as `quad` and `NIntegrate` do not evaluate the integrand at the end points, so they work, but a simple grid that includes the end points will not.

Units. Work in SI units with $Z_0 = \mu_0 c = 376.73\,\Omega$, or measure lengths in wavelengths, so that $k = 2\pi$; the radiation resistance comes out as a number times Z_0.

28.6 Summary

A source of charges and currents radiates when its moments change in time, and only the parts of its fields that fall off as $1/r$ carry power to infinity. For a source small compared with c/ω, the radiation is a sum of terms, one for each changing moment: the electric dipole first, then the magnetic dipole and the electric quadrupole together.

Small sources and what radiates

$d \ll c/\omega \ll r$	small source; the radiation zone
$\dfrac{dP}{d\Omega} = r^2\langle \mathbf{S}\rangle \cdot \hat{\mathbf{r}}$	power per unit solid angle; the radiation pattern; $\langle\cos^2\omega t\rangle = \frac{1}{2}$
Find V; $\mathbf{A}$; $\mathbf{E}$ and $\mathbf{B}$; $\mathbf{S}$ and $\langle\mathbf{S}\rangle$; $\langle P\rangle$	the five steps for a model source

Electric dipole $\mathbf{p} = p_0 \cos\omega t\, \hat{\mathbf{z}}$, with $t_0 = t - r/c$

$V = -\dfrac{p_0\omega}{4\pi\varepsilon_0 c}\dfrac{\cos\theta}{r}\sin\omega t_0,$ $\mathbf{A} = -\dfrac{\mu_0 p_0\omega}{4\pi r}\sin\omega t_0\,\hat{\mathbf{z}}$	potentials in the radiation zone; the $1/r$ in V comes from the difference in retarded times
$\mathbf{E} = -\dfrac{\mu_0 p_0\omega^2}{4\pi}\dfrac{\sin\theta}{r}\cos\omega t_0\,\hat{\boldsymbol{\theta}},$ $\mathbf{B} = \dfrac{E_\theta}{c}\hat{\boldsymbol{\phi}}$	spherical waves; $\mathbf{E}\perp\mathbf{B}\perp\hat{\mathbf{r}}$
$\langle\mathbf{S}\rangle = \dfrac{\mu_0 p_0^2\omega^4}{32\pi^2 c}\dfrac{\sin^2\theta}{r^2}\hat{\mathbf{r}}$	intensity; the $\sin^2\theta$ doughnut
$\langle P\rangle = \dfrac{\mu_0 p_0^2\omega^4}{12\pi c}$	grows as ω^4: why the sky is blue
$R = \dfrac{2\pi}{3}Z_0\left(\dfrac{d}{\lambda}\right)^2 \approx (789\,\Omega)\left(\dfrac{d}{\lambda}\right)^2,$ $Z_0 = \mu_0 c = 376.73\,\Omega$	radiation resistance, $\langle P\rangle = \frac{1}{2}I_0^2 R$

Magnetic dipole $\mathbf{m} = m_0\cos\omega t\,\hat{\mathbf{z}}$

$\mathbf{E} = \dfrac{\mu_0 m_0\omega^2}{4\pi c}\dfrac{\sin\theta}{r}\cos\omega t_0\,\hat{\boldsymbol{\phi}},$ $\mathbf{B} = -\dfrac{E_\phi}{c}\hat{\boldsymbol{\theta}}$	the roles of $\mathbf{E}$ and $\mathbf{B}$ exchanged
$\langle P\rangle = \dfrac{\mu_0 m_0^2\omega^4}{12\pi c^3}$	the electric dipole's, with m_0/c for p_0

Any small source

$\mathbf{A} \cong \dfrac{\mu_0}{4\pi r}\displaystyle\int \mathbf{J}\left(\mathbf{r}', t_0 + \frac{\hat{\mathbf{r}}\cdot\mathbf{r}'}{c}\right)d\tau'$	far away; t_0 is the retarded time of the origin
$\mathbf{B} = -\dfrac{1}{c}\hat{\mathbf{r}}\times\dfrac{\partial\mathbf{A}}{\partial t}, \quad \mathbf{E} = c\mathbf{B}\times\hat{\mathbf{r}}$	the fields from $\mathbf{A}$ alone, by the Lorenz condition
$\mathbf{B} = -\dfrac{\mu_0}{4\pi c r}\hat{\mathbf{r}}\times$ $\left[\ddot{\mathbf{p}} + \dfrac{1}{c}\ddot{\mathbf{m}}\times\hat{\mathbf{r}} + \dfrac{1}{3c}\dfrac{d^3}{dt^3}\left(\overleftrightarrow{\mathbf{Q}}\cdot\hat{\mathbf{r}}\right)\right]$	the multipole expansion; no monopole radiation, since charge is conserved
$P_{\mathrm{E1}} = \dfrac{\mu_0\ddot{p}^2}{6\pi c}$	any time dependence; $\ddot{\mathbf{p}} = q\mathbf{a}$ gives Larmor's formula
$P_{\mathrm{M1}} = \dfrac{\mu_0\ddot{m}^2}{6\pi c^3}, \quad P_{\mathrm{E2}} = \dfrac{\mu_0}{180\pi c^3}\displaystyle\sum_{i,j}\dddot{Q}_{ij}^2$	the next order; the powers add; with $Q^{\mathrm{alt}} = 2Q$, 720π (averaged, 1440π)

Antennas of any length

$$\mathbf{A} = \mathrm{Re}\left[\frac{\mu_0 e^{-i\omega t_0}}{4\pi r}\int \tilde{\mathbf{J}}(\mathbf{r}')\, e^{-ik\hat{\mathbf{r}}\cdot\mathbf{r}'}\, d\tau'\right]$$

the far field of a current at one frequency, $k = \omega/c$

$$\frac{dP}{d\Omega} = \frac{Z_0 k^2}{32\pi^2}\sin^2\theta\left|\int \tilde{I}(z')\, e^{-ikz'\cos\theta}\, dz'\right|^2$$

a straight antenna; integrate numerically

$$R = 73.1\,\Omega$$

half-wave antenna; a quarter-wave antenna on the ground has half

$$F_{\text{array}} = F\sum_n e^{i(\alpha_n - k\hat{\mathbf{r}}\cdot\mathbf{r}_n)}$$

arrays: the array factor steers the beam

Scattering

$$\langle P\rangle = \frac{\mu_0\alpha^2 E_0^2\omega^4}{12\pi c}$$

a molecule's induced dipole: blue sky, red sunsets, polarized skylight

Exercises

Exercises for 28.1 Radiation from small sources: power and angular distribution

P28.1 Which sources are small? For each of the following sources, compare its size with c/ω, say whether the formulas of this chapter for small sources can describe it, and estimate how far away the radiation zone begins.

(a) The tower of an AM radio station, 75 m tall, broadcasting at 1.0 MHz.

(b) The antenna of a Wi-Fi router, 3 cm long, at 2.4 GHz.

(c) A nitrogen molecule, about 0.3 nm across, in green sunlight of wavelength 550 nm.

(d) The coil of an MRI scanner, about 0.3 m across, which sends radio waves at 64 MHz into the body of a patient inside it. Is the patient in the radiation zone?

Exercises for 28.2 Electric dipole radiation

P28.2 Do the radiation fields satisfy Maxwell's equations? The fields (28.8) and (28.9) came from three approximations.

(a) Show that they satisfy Faraday's law, $\nabla \times \mathbf{E} = -\partial \mathbf{B}/\partial t$, to order $1/r$.

(b) Show that they satisfy the Ampère–Maxwell law in empty space, $\nabla \times \mathbf{B} = (1/c^2)\,\partial \mathbf{E}/\partial t$, to order $1/r$.

(c) One component of $\nabla \times \mathbf{B}$ does not match at order $1/r^2$. Which one, and why is that not a contradiction?

P28.3 The fields of an ideal dipole at every distance. Let $d \to 0$ with $p_0 = q_0 d$ fixed, so that the first two approximations of Section 28.2.1 become exact, but do not make the third.

(a) Show that the vector potential is then $\mathbf{A} = (\mu_0/4\pi r)\,\dot{p}(t - r/c)\,\hat{\mathbf{z}}$ at every distance, and use it with (28.5) to find $\mathbf{E}$ and $\mathbf{B}$ everywhere.

(b) Show that close to the dipole, $r \ll c/\omega$, the electric field is the static dipole field (8.15) with the dipole moment $p(t - r/c)$, and that far away it is (28.8).

(c) At what distance are the amplitudes of the parts of E_θ that fall off as $1/r$ and as $1/r^3$ equal?

P28.4 A neighbor of yours puts a small vertical dipole antenna on top of a tower of height h. It radiates the average power P at the angular frequency ω, with $c/\omega \ll h$.

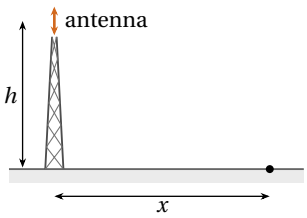


Figure 28.24 The antenna of Exercise P28.4 on top of a tower of height h, and a point on the ground a horizontal distance x from the foot of the tower.

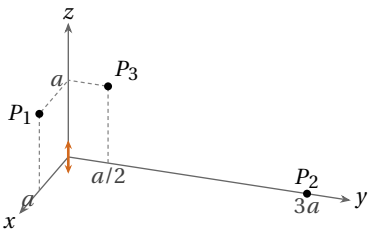


Figure 28.25 The oscillating dipole and the three detectors of Exercise P28.5.

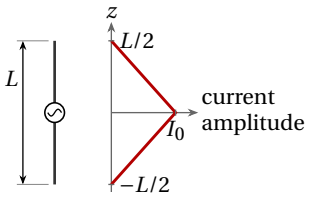


Figure 28.26 The antenna of Exercise P28.6, fed at its center, and the amplitude of its current along its length.

(a) Find the intensity at ground level, a horizontal distance x from the foot of the tower (Figure 28.24).

(b) At what distance from the tower is the intensity at ground level largest?

(c) What is that largest intensity?

(d) For $P = 10\,\text{kW}$ at $100\,\text{MHz}$ and $h = 20\,\text{m}$, evaluate it in mW/cm^2, and compare it with the limit of $0.2\,\text{mW/cm}^2$ that the US Federal Communications Commission sets for the general public at that frequency (Federal Communications Commission, "Radiofrequency radiation exposure limits," 47 CFR § 1.1310, Table 1, https://www.ecfr.gov/current/title-47/chapter-I/subchapter-A/part-1/subpart-I/section-1.1310).

P28.5 An electric dipole at the origin oscillates along the z axis. Three detectors (Figure 28.25) are at $P_1 = (a,0,a)$, $P_2 = (0,3a,0)$, and $P_3 = (0,a/2,a)$, where a is much larger than c/ω, and c/ω is much larger than the dipole.

(a) Which detector receives the smallest intensity?

(b) If that detector receives the intensity X, what do the other two receive?

P28.6 A more realistic short antenna. A straight wire of length $L \ll \lambda$ along the z axis, fed at its center, carries the current $I(z,t) = I_0(1-2|z|/L)\cos\omega t$ (Figure 28.26). The current is largest at the center and zero at the ends, since there are no spheres to hold charge.

(a) Use the continuity equation to find the charge per unit length on each half of the wire.

(b) Show that the dipole moment is $\mathbf{p}(t) = (I_0 L/2\omega)\sin\omega t\,\hat{\mathbf{z}}$.

(c) Find the average radiated power.

(d) Show that the radiation resistance, defined with the current amplitude I_0 at the center, is $R = (\pi/6)Z_0(L/\lambda)^2 \approx (197\,\Omega)(L/\lambda)^2$.

(e) Why is this a quarter of (28.13) for the same length?

P28.7 The scattering cross section.

(a) A molecule of polarizability α sits in a plane wave whose electric field has the amplitude E_0. Show that the average power it scatters is σI, where $I = E_0^2/2\mu_0 c$ is the intensity of the wave (use $B = E/c$ for a plane wave, Section 16.5) and

$$\sigma = \frac{8\pi}{3}k^4\left(\frac{\alpha}{4\pi\varepsilon_0}\right)^2, \qquad k = \frac{\omega}{c}.$$

The **scattering cross section** σ has the units of area: the molecule removes as much power from the wave as falls on an area σ.

(b) For a nitrogen molecule, $\alpha/4\pi\varepsilon_0 = 1.71 \times 10^{-30}\,\mathrm{m}^3$ (NIST, "CCCBDB list of experimental polarizabilities," Computational Chemistry Comparison and Benchmark Database, Release 22 (2022), https://cccbdb.nist.gov/pollistx.asp). Find σ at 450, 550, and 700 nm.

(c) Air at sea level, at 101 kPa and 15°C, holds $n = P/k_B T$ molecules per cubic meter. Treating them all as nitrogen, find for each wavelength the distance $1/n\sigma$ over which a beam loses all but $1/e$ of its power to scattering.

(d) The air above each square meter of the ground weighs $1.01 \times 10^5\,\mathrm{N}$, the average mass of a molecule of air is 29 u, and $g = 9.81\,\mathrm{m/s^2}$. How many molecules are there above each square meter of the ground? What fraction of the blue light (450 nm) from the Sun overhead is scattered out of the beam before it reaches the ground, and what fraction of the red light (700 nm)?

P28.8 The polarization of skylight in any direction. Unpolarized light travels along the z axis and is scattered by a molecule at the origin toward a direction in the xz plane, at the angle Θ from the z axis. Treat the unpolarized light as two waves of equal intensity, polarized along x and along y, that do not interfere.

(a) Find the intensity that each scatters toward that direction, and its polarization.

(b) Show that the degree of polarization, $(I_\perp - I_\parallel)/(I_\perp + I_\parallel)$, is $\sin^2\Theta/(1+\cos^2\Theta)$, where $I_\perp$ is the scattered intensity polarized perpendicular to the xz plane and $I_\parallel$ the intensity polarized in it.

(c) Give two reasons why the light of the real sky, 90° from the Sun, is not completely polarized.

Exercises for 28.3 Magnetic dipole radiation

P28.9 Radiation resistance of a loop antenna.

(a) Define the radiation resistance of the loop of Section 28.3 as in Section 28.2.3, and show that $R = (8\pi^5/3)Z_0(b/\lambda)^4 \approx (3.07 \times 10^5\,\Omega)(b/\lambda)^4$.

(b) Compare it with the radiation resistance of the two-sphere antenna of Figure 28.2 with $d = 2b$, at the same wavelength. Which is larger when $b \ll \lambda$, and by what factor?

P28.10 Electric and magnetic dipoles.

(a) Show that the fields (28.19) of the magnetic dipole follow from those of the electric dipole, (28.8) and (28.9), by replacing p_0 with m_0/c, $\mathbf{E}$ with $c\mathbf{B}$, and $c\mathbf{B}$ with $-\mathbf{E}$.

(b) Show that Maxwell's equations in empty space do not change when $\mathbf{E}$ is replaced by $c\mathbf{B}$ and $c\mathbf{B}$ by $-\mathbf{E}$.

(c) Use the static dipole fields (8.15) and (12.26), which appear side by side in Chapter 14, to explain why p_0 is replaced by m_0/c rather than by m_0.

Exercises for 28.4 Radiation from an arbitrary small source: the multipole expansion

P28.11 The integral of the current density. For a localized source, show that $\int \mathbf{J}\, d\tau' = \dot{\mathbf{p}}$, the rate of change of its electric dipole moment.

(a) By the product rule, $\nabla' \cdot (x'\mathbf{J}) = x' \nabla' \cdot \mathbf{J} + J_x$. Integrate over a volume that contains the whole source, and use the divergence theorem and the continuity equation (20.5) to show that $\int J_x\, d\tau' = \dot{p}_x$. The other components follow in the same way.

(b) Check the result for the antenna of Figure 28.2, whose wire carries the current $I(t) = -q_0\omega \sin\omega t$.

P28.12 One slow charge, two ways. A charge q moves slowly near the origin along the path $\mathbf{w}(t)$, with the velocity $\mathbf{v}$ and the acceleration $\mathbf{a}$. Work far away, at order $1/r$.

(a) Show that the first two terms of the expansion of Section 28.4.2 give

$$\mathbf{A} \cong \frac{\mu_0 q}{4\pi r}\left[\mathbf{v} + \frac{1}{c}(\hat{\mathbf{r}}\cdot\mathbf{v})\,\mathbf{v} + \frac{1}{c}(\hat{\mathbf{r}}\cdot\mathbf{w})\,\mathbf{a}\right],$$

with $\mathbf{w}$, $\mathbf{v}$, and $\mathbf{a}$ evaluated at $t_0 = t - r/c$.

(b) Now start from the Liénard–Wiechert vector potential (27.2). Far away, $1/(1-\hat{\boldsymbol{\imath}}\cdot\mathbf{v}/c) \cong 1 + \hat{\mathbf{r}}\cdot\mathbf{v}/c$ and $t_r \cong t_0 + \hat{\mathbf{r}}\cdot\mathbf{w}/c$, so that $\mathbf{v}(t_r) \cong \mathbf{v}(t_0) + (\hat{\mathbf{r}}\cdot\mathbf{w}/c)\,\mathbf{a}(t_0)$. Show that, when products of two of these small corrections are dropped, the result agrees with (a), term by term.

(c) Which term of (a) carries the factor $1/(1-\hat{\boldsymbol{\imath}}\cdot\mathbf{v}/c)$ of the Liénard–Wiechert potentials, and which carries the difference between t_r and t_0?

P28.13 A pulsating sphere of charge. A spherically symmetric distribution of charge, with the total charge Q, swells and shrinks: its charge density $\rho(r,t)$ changes in time but stays spherically symmetric, and it is zero beyond some radius a.

(a) Use Gauss's law to show that outside the distribution the electric field is the field of a point charge Q at the center, at every instant.

(b) By symmetry, the magnetic field could only point along $\hat{\mathbf{r}}$ and depend only on r and t. Use $\nabla\cdot\mathbf{B}=0$ to show that it is zero everywhere.

(c) Inside the distribution, show that the current density and the displacement current density add to zero: $\mathbf{J}+\varepsilon_0\partial\mathbf{E}/\partial t=0$. (Apply the continuity equation and Gauss's law to a sphere of radius r.) Why must they, by the Ampère–Maxwell law?

(d) Does the distribution radiate? Explain.

P28.14 A charge near a conducting plane. A charge q of mass m moves slowly along a line perpendicular to a large grounded conducting plane, at the height z above it, with the acceleration a (Figure 28.27). It stays close to the plane: $z \ll c/\omega$ for the highest frequency in its motion.

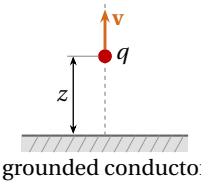


Figure 28.27 The charge q of Exercise P28.14 at the height z above a grounded conducting plane, moving along the perpendicular to the plane.

(a) Above the plane, the fields are those of the charge and its image (Section 7.2.1). Find the electric dipole moment of the charge and its image, and its second derivative.

(b) Find the radiated power. (Hint: the fields exist only above the plane.) Compare it with Larmor's formula for the charge alone.

(c) Suppose that the only force on the charge is the attraction of its image. Find the radiated power in terms of q, m, and z. (Do not worry about effects of the retarded time on the image force.)

P28.15 A dipole switched on. The dipole moment of a small source grows smoothly from 0 to p_0 as $\mathbf{p}(t)=\frac{1}{2}p_0[1+\tanh(t/\tau)]\,\hat{\mathbf{z}}$.

(a) Sketch $p(t)$ and $\ddot{p}(t)$. When does the source radiate most strongly?

(b) Find the total energy radiated. (Hint: substitute $u=\tanh(t/\tau)$, with $du=\operatorname{sech}^2(t/\tau)\,dt/\tau$ and $\operatorname{sech}^2=1-\tanh^2$.)

(c) How does the energy depend on τ? Why does a dipole that is switched on quickly radiate more?

P28.16 Why the powers add.

(a) Each of the three terms of (28.25) either changes sign or does not when $\hat{\mathbf{r}}$ is replaced by $-\hat{\mathbf{r}}$, with the moments held fixed. Which do, and which do not?

(b) Use (a) to show that the cross terms between the electric dipole term and the other two integrate to zero over all directions.

(c) The cross term between the other two does not change sign. Show that it integrates to zero anyway. Use the symmetry of Q_{ij} and the integral

$$\oint \hat{r}_j \hat{r}_l \, d\Omega = \frac{4\pi}{3} \delta_{jl}.$$

(d) Show that for a symmetric quadrupole moment whose diagonal components add to zero,

$$\oint \left| \hat{\mathbf{r}} \times \left(\overleftrightarrow{\mathbf{Q}} \cdot \hat{\mathbf{r}} \right) \right|^2 d\Omega = \frac{4\pi}{5} \sum_{i,j} Q_{ij}^2.$$

Use also $\oint \hat{r}_i \hat{r}_j \hat{r}_k \hat{r}_l \, d\Omega = (4\pi/15)(\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})$.

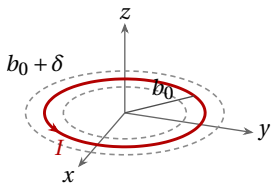


Figure 28.28 The ring of Exercise P28.17, whose radius oscillates between $b_0 - \delta$ and $b_0 + \delta$ (dashed).

P28.17 A ring that swells and shrinks. A ring in the xy plane, centered at the origin, has the radius $b(t) = b_0 + \delta \cos \omega t$, with $\delta \ll b_0 \ll c/\omega$ (Figure 28.28).

(a) The ring is a wire that carries the constant current I while its radius changes. Find its magnetic dipole moment to first order in δ, and the average power that it radiates.

(b) Instead, the ring is a thin ring of charge Q and mass M that spins about its axis. Its radius oscillates in the same way while its angular momentum L stays fixed. Show that its magnetic dipole moment is $QL/2M$, which does not change. Does it emit magnetic dipole radiation?

(c) Find the quadrupole moment of the spinning ring of (b), and the average power of its electric quadrupole radiation, to first order in δ.

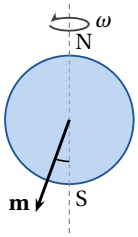


Figure 28.29 The Earth of Exercise P28.18, its axis of rotation, and its magnetic dipole moment **m**, tilted from the axis (the angle is exaggerated).

P28.18 Does the Earth radiate? The Earth's magnetic dipole moment (Figure 28.29) is about $8.0 \times 10^{22}\,\mathrm{A \cdot m^2}$ (Section 12.6.2), tilted 9.21° from the axis of rotation in 2025 ("Wandering geomagnetic poles," NOAA National Centers for Environmental Information, https://www.ncei.noaa.gov/products/wandering-geomagnetic-poles). The Earth rotates at the angular velocity $7.292115 \times 10^{-5}\,\mathrm{rad/s}$ (National Geospatial-Intelligence Agency, "World Geodetic System 1984," https://www.unoosa.org/pdf/icg/2012/template/WGS_84.pdf).

(a) Treat the dipole as fixed in the Earth, which is surrounded by empty space. How much power does it radiate?

(b) How far from the Earth does the radiation zone begin? Compare that distance with the Earth's radius, $6.371 \times 10^6\,\mathrm{m}$.

P28.19 How old is the Crab pulsar?

(a) Suppose that a pulsar loses rotational energy only by magnetic dipole radiation, with $\mathbf{m}$ and ψ fixed. Use the results of Example 28.3 to show that its rotation frequency obeys $\dot{\nu} = -K\nu^3$, and find the constant K.

(b) Integrate to find how long ago the pulsar rotated at the frequency ν_i. Show that if $\nu_i \gg \nu$, the answer is $\nu/2|\dot{\nu}|$, which is called the pulsar's *characteristic age*.

(c) Evaluate the characteristic age of the Crab pulsar with the data of Example 28.3. The supernova whose remnant is the Crab Nebula was seen in 1054 ("The Crab Nebula: an astrophysical chimera," J. J. Hester, Annu. Rev. Astron. Astrophys. 46, 127–155 (2008), https://doi.org/10.1146/annurev.astro.45.051806.110608). Compare.

(d) The *braking index* of a pulsar is $n = \nu\ddot{\nu}/\dot{\nu}^2$. Show that $n = 3$ for a pulsar slowed by magnetic dipole radiation alone.

Exercises for 28.5 Numerical methods: antennas of any length

P28.20 (Numerical) Longer antennas. A thin antenna of length L, along the z axis and fed at its center, carries to a good approximation a standing wave of current that is zero at its ends, $I(z,t) = I_m \sin[k(L/2 - |z|)]\cos\omega t$ (Figure 28.30).

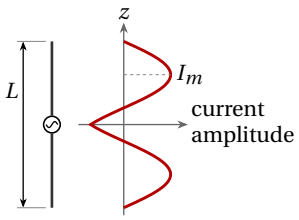


Figure 28.30 An antenna of Exercise P28.20, fed at its center, and the amplitude of its current along its length, a sine wave that is zero at the ends.

(a) Show that the half-wave antenna of Example 28.4 is the case $L = \lambda/2$.

(b) Show that $F(\theta) = (2I_m/k)[\cos(\frac{1}{2}kL\cos\theta) - \cos(\frac{1}{2}kL)]/\sin^2\theta$.

(c) Plot the radiation patterns for $L = \lambda$ and $L = 3\lambda/2$, in a plane through the antenna and in three dimensions. How do they differ from the half-wave antenna's?

(d) Compute the radiation resistance of each, defined with the current amplitude I_m at the largest point of the standing wave.

(e) In this model, what is the radiation resistance of the antenna with $L = \lambda$ as seen at its feed point, its center? Why?

P28.21 Two antennas. Two identical short dipole antennas along the z axis, one at the origin and one at $z = d$, carry currents of the same amplitude (Figure 28.31).

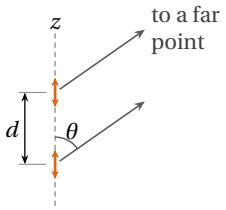


Figure 28.31 The two antennas of Exercise P28.21, a distance d apart on the z axis, and a far direction at the angle θ from the axis.

(a) Plot the radiation pattern of one short dipole, $\sin^2\theta$, in three dimensions, with SphericalPlot3D or with matplotlib (Section 28.5.3).

(b) The two antennas are driven in phase. Show that far away $\langle S\rangle \propto \sin^2\theta\cos^2[(\pi d/\lambda)\cos\theta]$, and plot the pattern for $d = \lambda$.

(c) Now the current in the upper antenna lags that in the lower one by the phase Δ: $I_{\text{upper}} = I_0 \cos(\omega t - \Delta)$, $I_{\text{lower}} = I_0 \cos \omega t$. Show that $\langle S \rangle \propto \sin^2\theta \cos^2[(\Delta - kd\cos\theta)/2]$, and plot the pattern for $d = \lambda$ and $\Delta = \pi/2$. What symmetry has it lost?

(d) For $d = \lambda/2$ and the two antennas in phase, is the total power twice that of one antenna? Compute it numerically.

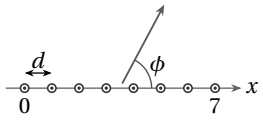


Figure 28.32 The row of antennas of Exercise P28.22, seen from above, and a direction in the horizontal plane at the angle ϕ from the x axis.

P28.22 (Numerical) Steering a beam. Eight half-wave antennas, all vertical, stand in a row along the x axis, a distance $d = \lambda/2$ apart (Figure 28.32). Each carries the same current amplitude, and the nth ($n = 0, 1, \ldots, 7$) is driven with the phase $n\alpha$.

(a) Use (28.30) to write the radiation pattern in the horizontal plane, $\theta = 90°$, as a function of the angle ϕ from the x axis.

(b) Plot the pattern for $\alpha = 0$. In which direction is the beam, and how wide is it, measured between the directions where the power per unit solid angle is half its largest value?

(c) Find the direction of the beam for $\alpha = 45°$, $90°$, and $135°$, and show that it obeys $\cos\phi_0 = \alpha/kd$. How does the width of the beam change as it is steered?

(d) For $\alpha = 0$, how does the width of the beam depend on the number of antennas? Try 4 and 16.

Exercises for 28.6 Summary

P28.23 True or false? Explain your reasoning.

(a) A sphere of charge whose radius oscillates, while it keeps its spherical symmetry, radiates at the frequency of the oscillation.

(b) The static fields of a source can carry energy all the way to infinity.

(c) An oscillating electric dipole has the potential $p(t - r/c)\cos\theta/4\pi\varepsilon_0 r^2$.

(d) A charge that oscillates back and forth along a line radiates.

(e) A small loop antenna radiates as much power as a two-sphere antenna of the same size with the same current.

(f) In the radiation zone of any small source, $\mathbf{E}$, $\mathbf{B}$, and $\hat{\mathbf{r}}$ are perpendicular to one another.

(g) A source whose electric dipole moment never changes cannot radiate.

(h) If the frequency of an oscillating electric dipole is doubled and its amplitude kept the same, it radiates four times the power.

(i) A half-wave antenna, like a short one, sends no radiation along its own length.

(j) The sky is blue because the molecules of the air absorb red light.

P28.24 AI programming. Use AI resources to do something computational beyond what you would be able to do yourself, to help you better understand one of the concepts from this chapter, such as the exact field lines of an oscillating dipole, radiation patterns in three dimensions, the far field of any distribution of current, half-wave and longer antennas, phased arrays that steer a beam, the multipole moments of a source and the power radiated by each, or the color of the sky computed from Rayleigh scattering and the spectrum of sunlight.

Part IX

Appendices

Appendix A

Convolutions

Appendix B

Green's Functions?

Appendix C

The Kramers–Kronig Relations

Portrait Credits

The portraits in the margins come from Wikimedia Commons or the sources named below. Most are in the public domain. Credits are given where a license asks for them and otherwise as a courtesy; where a creator is not named, none is recorded.

d'Alembert. Pastel by Maurice Quentin de La Tour, 1753.

Ampère. Engraving by Ambroise Tardieu, 1825. Dibner Collection, Smithsonian Libraries.

Arago. Portrait by Ary Scheffer, 1842.

Argand. No portrait of Argand is known; the picture is a copy of his *Essai* (Paris, 1806), inscribed to Gergonne. Bibliothèque universitaire de Bordeaux, Fonds patrimonial; digitized by the Université de Bordeaux.

Bessel. Portrait by Johann Eduard Wolff, 1834.

Biot. Engraving by J. Quenedey after a physionotrace. Wellcome Collection, CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

Bohr. Photograph by AB Lagrelius & Westphal, about 1922. Nobel Foundation.

Brewster. Photograph by H. J. Whitlock, before 1868. Dibner Library, Smithsonian Libraries.

Cauchy. Portrait by Jean Roller, about 1840. Château de Sceaux.

Clairaut. Engraving by Louis-Jacques Cathelin after Charles-Nicolas Cochin, 18th century.

Clausius. Photograph, photographer not recorded.

Compton. Press photograph, 1927 (Mondadori).

Coulomb. Portrait by Louis Hierle, 1894.

Marie and Pierre Curie. Photograph by Science Service, 1903. Smithsonian Institution Archives, no known copyright restrictions.

Debye. Photograph, Nobel Foundation, 1936.

Descartes. Portrait after Frans Hals. Musée du Louvre, Paris.

Dirac. Photograph, Nobel Foundation, 1933.

Doppler. Portrait, artist and date not recorded.

Drude. Photograph, photographer and date not recorded. Dibner Library, Smithsonian Libraries.

Earnshaw. Photograph of a bust by William Ellis, from W. S. Porter, *Sheffield Literary and Philosophical Society: A Centenary Retrospect, 1822–1922* (Sheffield, 1922); the bust was then at the University of Sheffield; cropped.

Einstein. Photograph by Ferdinand Schmutzer, Vienna, 1921; restoration by Adam Cuerden; cropped.

Euler. Portrait by Jakob Emanuel Handmann, 1753. Kunstmuseum Basel.

Ewing. Sir (James) Alfred Ewing by Bassano Ltd, whole-plate glass negative, 13 January 1915. © National Portrait Gallery, London, CC BY-NC-ND 3.0 (https://creativecommons.org/licenses/by-nc-nd/3.0/).

Faraday. Photograph, photographer not recorded, about the 1850s.

Feynman. Photograph from *The Big T*, the yearbook of the California Institute of Technology, 1986; cropped.

Fourier. Engraving by A. F. B. Geille after a portrait by Julien-Léopold Boilly, early 19th century.

Franklin. Portrait by Joseph Siffred Duplessis, about 1785. National Portrait Gallery, Smithsonian Institution.

Fresnel. Engraving by E. Rosette after a portrait by A. Tardieu, from the frontispiece of Fresnel's *Œuvres complètes* (1866).

Galileo. Oil by Justus Sustermans, 1636–40. Royal Museums Greenwich.

Gauss. Portrait by Christian Albrecht Jensen, 1840.

Gibbs. Photograph, photographer not recorded, before 1903; frontispiece of *The Scientific Papers of J. Willard Gibbs* (1906).

Gilbert. Oil on wood, artist not recorded, 16th century. Wellcome Collection.

Green. No portrait of Green is known; the picture is his windmill, Green's Mill, Sneinton, Nottingham. Photograph by Alex Foster, 2005 (geograph.org.uk). CC BY-SA 2.0 (https://creativecommons.org/licenses/by-sa/2.0/).

Hall. Photograph, photographer not recorded; from *Popular Science Monthly*, vol. 64 (1903–04).

Hartree. Photograph, photographer not recorded, about 1940.

Heaviside. Photograph, photographer not recorded, about 1900. IET Archives; cropped.

Helmholtz. Portrait, creator not recorded.

Henry. Photograph by Mathew Brady or Levin Handy, between 1865 and 1880. Library of Congress, Brady-Handy Collection.

Hertz. Photograph by Robert Krewaldt, Bonn, about 1890; cropped.

Jacobi. Portrait, creator not recorded.

Kelvin. Photograph by T. & R. Annan & Sons, about 1900; restored by Adam Cuerden. National Galleries of Scotland.

Kirchhoff. Photograph, photographer and date not recorded. Smithsonian Libraries.

Kramers. Photograph by van Duinen (Anefo), 1952. Nationaal Archief, the Netherlands (CC0); cropped.

Kronecker. Photograph, photographer not recorded, 1865.

Kronig. Photograph by the Fotografische Dienst TU Delft, about 1962. © Delft University of Technology, CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/); cropped.

Lagrange. Engraving, engraver not recorded. Dibner Collection, Smithsonian Libraries.

Langevin. Photograph by Henri Manuel. Wellcome Collection, CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

Laplace. Engraving by James Posselwhite, about 1850.

Larmor. Sir Joseph Larmor by Walter Stoneman, bromide print, 1920. © National Portrait Gallery, London, CC BY-NC-ND 3.0 (https://creativecommons.org/licenses/by-nc-nd/3.0/).

Lawrence. Photograph, Nobel Foundation, 1939.

Legendre. Sketch from Georges Barral, *Le Panthéon scientifique de la tour Eiffel* (1892).

Lenz. Engraving, engraver not recorded. Dibner Collection, Smithsonian Libraries.

Liénard. Photograph, 1887, as a student at the École polytechnique. Collections École polytechnique; cropped.

Lorentz. Photograph, photographer not recorded, 1902. Nationaal Archief, the Netherlands (public domain mark).

Lorenz. Portrait from H. Valentiner, ed., *Œuvres scientifiques de L. Lorenz*, vol. 2 (1896–1904); scan provided by the Niels Bohr Archive, Copenhagen; cropped and retouched by Pbrks (Wikimedia Commons file "Ludvig Valentin Lorenz (cropped).jpg"). Free Art License (http://artlibre.org/licence/lal/en).

Maxwell. Photograph, photographer not recorded, about 1870.

Michelson. Photograph, Bain News Service (glass negative; date not recorded). Library of Congress; cropped.

Minkowski. Frontispiece of *Raum und Zeit* (1909); cropped (as on Wikimedia Commons).

Morley. Photograph, about 1885. Digital Public Library of America.

Mossotti. Portrait, creator not recorded, before 1863.

Néel. Press photograph by Farabola, 1970.

Neumann. Portrait by Carl Steffeck, 1886.

Ohm. Painting, painter and date not recorded. Bavarian Academy of Sciences and Humanities.

Ørsted. Portrait by Christian Albrecht Jensen, 1832–33. Statens Museum for Kunst, Copenhagen.

Peregrinus. Drawing of a compass in a fourteenth-century copy of his letter on the magnet. Bodleian Library, Oxford, MS Ashmole 1522, fol. 186r.

Poincaré. Photograph by Henri Manuel; print by Ch. Wittmann. KU Leuven Libraries; cropped.

Poisson. Lithograph by François-Séraphin Delpech after Nicolas Eustache Maurin, before 1840.

Poynting. Photograph by Lafayette, Dublin, reproduced in *The Illustrated London News*, 11 April 1914. Library of Congress; cropped.

Rayleigh. Photograph from "Obituary notices of Fellows deceased," Proc. R. Soc. Lond. A 98, No. 695 (1921).

Rodrigues. Portrait, creator not recorded, 19th century.

Savart. Photograph of his bust at the Institut de France, Paris, by Moonik, 2012 (Wikimedia Commons file "Interior of Institut de France, Paris 6th 013.jpg"). CC BY-SA 3.0 (https://creativecommons.org/licenses/by-sa/3.0/).

Snellius. Engraving, workshop of de Passe, from Joannes Meursius, *Athenae Batavae* (Leiden, 1625). Rijksmuseum, Amsterdam; cropped.

Steinmetz. Photograph by the White Studio, Schenectady, New York, about 1900. Collections of The Henry Ford.

Stokes. Photograph, photographer not recorded, 1860s.

Tait. Photograph by Dr Adamson of St Andrews, 1870.

Taylor. Portrait by Louis Goupy, about 1720. National Portrait Gallery, London.

Thomson. Photograph published by the Bain News Service, about 1920–25. Library of Congress, George Grantham Bain Collection; cropped.

Weiss. Photograph by Johannes Meiner, 1909. ETH-Bibliothek, Zürich.

Wessel. No portrait of Wessel survives; the picture is his map of the southeast quarter of Zealand, 1770. Bibliothèque nationale de France.

Wiechert. Photograph, photographer and date not recorded.

Index

3 dB point, 589

aberration of light, 773
absorption coefficient, 674
acceleration field, 852
advanced potentials, 736
d'Alembert, Jean le Rond, 732
Ampère's law, 17, 338
Ampère's law for H, 417
Ampère, André-Marie, 333
Ampère–Maxwell law, 504
Amperian loop, 337
amplitude, 636
angular frequency, 637
anomalous dispersion, 680
antiferromagnet, 426
apparent length, 735
Arago, François, 480
Argand, Jean-Robert, 548
array factor, 907
atomic polarizability, 235

back EMF, 476
band-pass filter, 589
base units, 533
beats, 706
Bessel functions, 162
Bessel, Friedrich Wilhelm, 162
Biot, Jean-Baptiste, 305
Biot–Savart law, 305
Bode plot, 589
Bohr magneton, 405
Bohr radius, 539
Bohr, Niels, 405
Boris method, 297
    relativistic, 789
bound charges, 241
bound currents, 409
bound-charge method, 242
bound-current method, 411
boundary conditions, 172
boundary value problem, 171
bremsstrahlung, 863
Brewster's angle, 653
Brewster, David, 655

capacitance, 127
capacitor, 127
Cauchy's formula, 681
Cauchy, Augustin-Louis, 681
causality, 684
center-of-momentum frame, 783
central difference, 124
characteristic impedance, 716
charge density, 11
circular polarization, 640
    right- and left-handed, 640
Clairaut, Alexis Claude, 7
Clausius, Rudolf, 246
Clausius–Mossotti relation, 246
coercive field, 425
coercivity, 425
collisions, 782
Colton's Commandments
    First, 61
    Second, 84
    Third, 339
Colton's trick, 161, 193
complex amplitude, 551
complex conductivity, 678
complex conjugate, 549
complex index of refraction, 674
complex number, 547
complex permittivity, 676

complex plane, 547
complex wave number, 555
Compton scattering, 783
Compton, Arthur H., 784
conductivity, 571
conductor, 91
conservative field, 106
conserved quantity, 781
constitutive relations, 569
continuity equation, 303
coordinates
    cylindrical, 4
    spherical, 4
corona discharge, 127
Coulomb gauge, 359, 730
Coulomb's law, 53
Coulomb, Charles-Augustin de, 51
coupling coefficient, 579
critical angle, 651
critical frequency, 862
crossed fields, 829
Curie temperature, 426
Curie's law, 406
Curie, Marie Skłodowska, 406
Curie, Pierre, 406
curl, 15
curl theorem, 22
curl-free field, 106
current, 300
cutoff frequency, 702
cyclotron, 295
cyclotron frequency, 295

D field, 248
d'Alembertian, 732
damped wave equation, 671
damping time, 574
DC conductivity, 574
Debye, Peter, 234
decibel, 589
Descartes, René, 12
diamagnetism, 407
dielectric constant, 249
dielectrics, 231
dipole, 212
    physical, 213
    pure, 213
dipole layer, 112
dipole moment, 216
Dirac delta function, 24
Dirac, Paul, 24
dispersion, 637, 680
dispersion relation, 702
displacement current, 505, 511
displacement current density, 505
displacement field, 248
divergence, 14
divergence theorem, 21
domain walls, 425
domains, 407
dominant mode, 703
Doppler effect, 826
Doppler, Christian, 826
drift velocity, 829
    of conduction electrons, 571
driven oscillator, 552
Drude model, 574
Drude's scattering model, 574
Drude, Paul, 574
dual tensor, 815

Earnshaw's theorem, 174
Earnshaw, Samuel, 175
eddy currents, 479
eigenvalue problem, 711
Einstein, Albert, 759
electrets, 238
electric dipole radiation, 897
electric displacement, 248
electric field, 54
electric flux, 82
electric potential, 103, 107
electric quadrupole radiation, 899
electric susceptibility, 239
electromotive force, 470
electron volt, 114
electrostatic unit of charge, 536
elliptical polarization, 640
elsewhere, 769
EMF, 470

energy
relativistic, 780
energy density, 131
equipotential surfaces, 110
equipotentials, 110
ether, 757
Euler's formula, 548
Euler, Leonhard, 8
evanescent wave, 666
even function, 156
event, 759
Ewing, James Alfred, 426
extinction coefficient, 674

farad, 127
Faraday cage, 93
Faraday current density, 478
Faraday's law, 475
Faraday, Michael, 474
ferrimagnet, 426
ferroelectrics, 238
ferromagnet, 424
ferromagnetism, 407
Feynman, Richard P., 476
field lines, 57
field point, 53
field tensor, 814
fine-structure constant, 540
flux method, 371
flux rule, 472
four-acceleration, 777
four-current, 816
four-divergence, 817
four-gradient, 817
four-momentum, 780
four-potential, 818
four-tensor, 776
four-vector, 774
four-velocity, 777
four-wavevector, 827
Fourier coefficients, 153
Fourier series, 157
Fourier's trick, 155
Fourier, Joseph, 155
frame of reference, 756
franklin (unit), 536
Franklin, Benjamin, 52
free charge, 241
free current, 415
frequency, 637
Fresnel equations, 652
Fresnel, Augustin-Jean, 652
fundamental frequency, 154
future (of an event), 769

Galilean transformation, 756
Galilei, Galileo, 756
gauge, 359
gauge transformation, 359
gauss (unit), 291, 537
Gauss's law, 17, 80
Gauss's law for D, 250
Gauss's law for H, 417
Gauss, Carl Friedrich, 79
Gauss–Seidel method, 181
Gaussian surface, 82
generating function, 215
Gibbs, J. Willard, 45
Gilbert, William, 401
good conductor, 671
gradient theorem, 19
Green, George, 107
group index, 707
group velocity, 706

H field, 415
Hall voltage, 304
Hall, Edwin, 304
hard magnetic materials, 426
hartree (unit), 539
Hartree, Douglas, 539
headlight effect, 773
Heaviside, Oliver, 44
Heaviside–Lorentz units, 540
helicity, 641
Helmholtz decomposition, 42
Helmholtz equation, 698
Helmholtz pair, 329
Helmholtz theorem, 42
Helmholtz, Hermann von, 42

henry (unit), 384
Henry, Joseph, 478
Hertz, Heinrich, 520
Hertzian dipole, 885
high-pass filter, 588
hyperbolic motion, 779
hysteresis loop, 425

image charge, 177
imaginary part, 547
impedance, 581
impedance matching, 587
impedance of free space, 643
index notation, 211
index of refraction, 636
induced charge, 92
induced dipole moment, 235
inductance, 384
inductor, 576
inertial frame, 756
inhomogeneous wave equation, 732
instantaneous rest frame, 778
integration by parts, 29
intensity, 609, 888
interval, 768
invariant quantity, 781
irrotational field, 106

Jacobi method, 181
Jacobi, Carl Gustav Jacob, 9
Jacobian, 10
Jefimenko's equations, 738
junction law, 580

Kelvin, Lord (William Thomson), 177
Kirchhoff's laws, 580
Kirchhoff, Gustav Robert, 581
Kramers' law, 865
Kramers, Hendrik Anthony, 864
Kramers–Kronig relations, 684
Kronecker, Leopold, 211
Kronig, Ralph de Laer, 684

Lagrange, Joseph-Louis, 135
Langevin, Paul, 407
Laplace's equation, 135
Laplace, Pierre-Simon, 17
Laplacian, 17
Larmor's formula, 856
Larmor, Joseph, 857
law of reflection, 651
law of refraction, 651
Lawrence, Ernest, 298
Legendre function of the second kind, 159
Legendre polynomials, 158
Legendre's differential equation, 159
Legendre, Adrien-Marie, 158
length contraction, 763
Lenz's law, 476
Lenz, Emil, 477
Liénard's formula, 860
Liénard, Alfred-Marie, 860
Liénard–Wiechert potentials, 848
light clock, 761
light cone, 759
lightlike, 769
line charge density, 58
linear dielectric, 239
linear magnetic material, 415
linear polarization, 640
longitudinal wave, 640
longitudinal-to-transverse equations, 697
loop law, 580
Lorentz factor, 758
Lorentz force law, 291
Lorentz oscillator model, 679
Lorentz scalar, 775
Lorentz transformation, 765
Lorentz, Hendrik Antoon, 291
Lorenz condition, 731
Lorenz gauge, 731
Lorenz, Ludvig, 247
loss tangent, 679
low-pass filter, 589

macroscopic field, 245
magnetic dipole, 357
magnetic dipole moment, 379
magnetic dipole radiation, 899

magnetic field, 289
magnetic field strength, 415
magnetic flux, 334
magnetic forces, 289
magnetic monopole, 334
magnetic scalar potential, 429
magnetic susceptibility, 415
magnetic vector potential, 357, 358
magnetization, 406
magnetostatics, 289
magnitude of a complex number, 547
matched load, 717
Maxwell stress tensor, 612
Maxwell's equations, 41
  in matter, 569
  tensor form, 817
Maxwell's term, 504
Maxwell, James Clerk, 41
Maxwellian surface, 613
method of images, 177
Michelson, Albert A., 757
Minkowski force, 787
Minkowski, Hermann, 764
mode (of a waveguide), 699
momentum
  relativistic, 780
momentum density of fields, 614
monopole, 212
monopole moment, 216
Morley, Edward W., 758
Mossotti, Ottaviano-Fabrizio, 245
motional EMF, 471
MST theorem, 612
multipole expansion, 209
mutual inductance, 384

Néel temperature, 426
Néel, Louis, 428
Neumann boundary condition, 573
Neumann, Franz Ernst, 384
normal dispersion, 680
normal incidence, 647
null point, 57

octopole, 221
odd function, 156
oersted (unit), 415
Ohm's law, 571
Ohm, Georg Simon, 571
ohmic heating, 573
ohmic material, 571
Ørsted, Hans Christian, 289
orthogonality, 153
oscillator strength, 680

p-polarization, 652
paramagnetism, 406
past (of an event), 769
Peregrinus de Maricourt, Petrus, 43
perfect conductor, 570, 695
period, 637
permeability, 416
permeability of free space, 305
permittivity, 249
permittivity of free space, 53
phase constant, 637
phase of a complex number, 547
phase velocity, 637
phase-lead convention, 550
pitch (of a helix), 295
plane of incidence, 651
plane wave, 642
plasma frequency, 679
Poincaré, Henri, 768
Poisson's equation, 134
  for **A**, 362
Poisson, Siméon Denis, 134
polar form, 547
polar molecules, 234
polarizability tensor, 237
polarization, 237
polarization current, 568
polarization current density, 506
polarization vector, 640
polarized material, 231
poor conductor, 671
postulates of relativity, 756
power factor, 585
Poynting vector, 606

Poynting's theorem, 606
Poynting, John Henry, 606
principal root, 550
principal value, 684
principle of relativity, 756
product rules, 28
projected position, 852
proper acceleration, 778
proper charge density, 816
proper length, 763
proper time, 761, 776

quadrupole, 212
quadrupole moment, 220
quality factor, 715
quasistatic approximation, 517
quasistatic fields, 517

radiation, 855
radiation pattern, 884
radiation pressure, 645
radiation resistance, 890
radiation zone, 884
Rayleigh scattering, 893
Rayleigh, Lord (John William Strutt), 893
reactance, 581
real part, 547
rectangular form, 547
reference point, 108
reflectance, 639
reflection coefficient, 639
relative permeability, 416
relative permittivity, 249
relativistic dynamics, 786
relativity of simultaneity, 760
relaxation method, 179
    for **A**, 375
    with dielectrics, 262
remanence, 425
remanent magnetization, 425
resistance, 572
resistivity, 571
resonance
    of a series RLC circuit, 584
resonant cavity, 712
resonant frequency (of a cavity), 713
rest energy, 780
rest length, 763
retarded potentials, 735
retarded time, 733
    numerical solution, 741
ridge waveguide, 711
right-hand rule, 290
Rodrigues' formula, 159
Rodrigues, Olinde, 163
root-mean-square value, 556
roots of a complex number, 549
rotation matrix, 210
Rydberg atomic units, 540

s-polarization, 652
saddle point, 124
saturation magnetization, 407
Savart, Félix, 306
scalar, 209
scalar potential, 358
scalar product
    of four-vectors, 775
scattering cross section, 913
screening, 237
script-r method, 60
self-energies, 133
self-inductance, 384
separation of variables, 182
separation vector, 53
shadow (of a moving source), 733
sheet resistance, 573
siemens (unit), 571
sifting property, 25
sign convention for complex numbers, 557
skin depth, 671
skin effect, 485
Snell's law, 651
Snellius, Willebrord, 651
soft magnetic materials, 426
solenoid, 300
solid angle, 855
source charges, 52

source point, 53
spacelike, 769
spacetime, 760
spacetime diagram, 759
spherical wave, 642
spin, 405
standing wave, 638
statcoulomb, 536
statvolt, 536
steady currents, 303
steady state, 581
Steinmetz, Charles Proteus, 558
Stokes' theorem, 22
Stokes, George Gabriel, 22
stress–energy tensor, 828
summation convention, 775
superposition, 52
surface charge density, 58
surface current density, 301
symmetry
    cylindrical, 85
    planar, 85, 339, 481
    spherical, 85
    toroidal, 340
synchronized clocks, 759
synchrotron radiation, 861

Tait, Peter Guthrie, 13
Taylor, Brook, 214
telegrapher's equations, 716
tensor, 211
tesla, 291
test charge, 52
Thomson, J. J., 297
time average, 556
time constant
    of an LR circuit, 577
time dilation, 761
time-delay convention, 550
timelike, 768
total internal reflection, 665
transfer function, 587
transformer, 579
transient, 581
transmission coefficient, 639
transmission line, 715
transmittance, 639
transverse Doppler effect, 827
transverse electric (TE) mode, 699
transverse electromagnetic (TEM) mode, 699
transverse magnetic (TM) mode, 699
transverse wave, 640
triangle diagram, 135
    for electrodynamics, 743
triple products, 28
twin paradox, 770

uniqueness theorems
    first, 174
    second, 175
unpolarized light, 641

vacuum wavelength, 637
vector, 210
vector area, 378
vector area element, 11
vector potential, 358
velocity addition, 772
velocity field, 852
velocity selector, 813
volt, 109
voltage, 109
volume charge density, 58
volume current density, 301

wave equation, 520
wave impedance, 643
wave number, 636
wave vector, 642
waveguide, 695
wavelength, 637
weber, 334
weight function, 163
Weiss, Pierre, 424
Wessel, Caspar, 548
Wiechert, Emil, 849
world line, 759

Vector Derivatives

Cylindrical (s, ϕ, z):

$$\nabla f = \frac{\partial f}{\partial s}\hat{\mathbf{s}} + \frac{1}{s}\frac{\partial f}{\partial \phi}\hat{\boldsymbol{\phi}} + \frac{\partial f}{\partial z}\hat{\mathbf{z}},$$

$$\nabla \cdot \mathbf{A} = \frac{1}{s}\frac{\partial (sA_s)}{\partial s} + \frac{1}{s}\frac{\partial A_\phi}{\partial \phi} + \frac{\partial A_z}{\partial z},$$

$$\nabla \times \mathbf{A} = \left(\frac{1}{s}\frac{\partial A_z}{\partial \phi} - \frac{\partial A_\phi}{\partial z}\right)\hat{\mathbf{s}} + \left(\frac{\partial A_s}{\partial z} - \frac{\partial A_z}{\partial s}\right)\hat{\boldsymbol{\phi}} + \frac{1}{s}\left(\frac{\partial (sA_\phi)}{\partial s} - \frac{\partial A_s}{\partial \phi}\right)\hat{\mathbf{z}},$$

$$\nabla^2 f = \frac{1}{s}\frac{\partial}{\partial s}\left(s\frac{\partial f}{\partial s}\right) + \frac{1}{s^2}\frac{\partial^2 f}{\partial \phi^2} + \frac{\partial^2 f}{\partial z^2}.$$

Spherical (r, θ, ϕ):

$$\nabla f = \frac{\partial f}{\partial r}\hat{\mathbf{r}} + \frac{1}{r}\frac{\partial f}{\partial \theta}\hat{\boldsymbol{\theta}} + \frac{1}{r\sin\theta}\frac{\partial f}{\partial \phi}\hat{\boldsymbol{\phi}},$$

$$\nabla \cdot \mathbf{A} = \frac{1}{r^2}\frac{\partial (r^2 A_r)}{\partial r} + \frac{1}{r\sin\theta}\frac{\partial (\sin\theta\, A_\theta)}{\partial \theta} + \frac{1}{r\sin\theta}\frac{\partial A_\phi}{\partial \phi},$$

$$\nabla \times \mathbf{A} = \frac{1}{r\sin\theta}\left(\frac{\partial (\sin\theta\, A_\phi)}{\partial \theta} - \frac{\partial A_\theta}{\partial \phi}\right)\hat{\mathbf{r}} + \frac{1}{r}\left(\frac{1}{\sin\theta}\frac{\partial A_r}{\partial \phi} - \frac{\partial (rA_\phi)}{\partial r}\right)\hat{\boldsymbol{\theta}} + \frac{1}{r}\left(\frac{\partial (rA_\theta)}{\partial r} - \frac{\partial A_r}{\partial \theta}\right)\hat{\boldsymbol{\phi}},$$

$$\nabla^2 f = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial f}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta\frac{\partial f}{\partial \theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 f}{\partial \phi^2}.$$

These are the formulas of Section 1.4.6. Like the Cartesian ones, they assume that the field is finite and differentiable at the point in question; at a point charge, for example, they can give the wrong answer (Section 1.6.3).

Physical Constants

Constant	Symbol	Value
Speed of light in vacuum	c	$2.99792458 \times 10^{8}\,\mathrm{m/s}$
Permittivity of free space	ε_0	$8.8542 \times 10^{-12}\,\mathrm{C^2/(N\,m^2)}$
	$1/4\pi\varepsilon_0$	$8.9876 \times 10^{9}\,\mathrm{N\,m^2/C^2}$
Permeability of free space	μ_0	$1.2566 \times 10^{-6}\,\mathrm{N/A^2}$ $\quad(\approx 4\pi \times 10^{-7}\,\mathrm{N/A^2})$
Elementary charge	e	$1.602176634 \times 10^{-19}\,\mathrm{C}$
Electron mass	m_e	$9.1094 \times 10^{-31}\,\mathrm{kg}$
Proton mass	m_p	$1.6726 \times 10^{-27}\,\mathrm{kg}$
Planck's constant	h	$6.62607015 \times 10^{-34}\,\mathrm{J\,s}$
	$\hbar$	$1.054571817\ldots \times 10^{-34}\,\mathrm{J\,s}$
Boltzmann's constant	k_B	$1.380649 \times 10^{-23}\,\mathrm{J/K}$
Avogadro's number	N_A	$6.02214076 \times 10^{23}\,\mathrm{mol^{-1}}$
Fine-structure constant	α	$7.2974 \times 10^{-3} = 1/137.04$
Bohr radius	a_0	$5.2918 \times 10^{-11}\,\mathrm{m}$
Bohr magneton	μ_B	$9.2740 \times 10^{-24}\,\mathrm{J/T}$

The values of c, e, h, k_B, and N_A are exact: since 2019 they define the SI units (Section 17.4). So is $\hbar = h/2\pi$, shown here truncated. The other values are measured; they are from CODATA 2022,[1] rounded to five significant figures.

Integrals that come up in script-r problems ($a > 0$; constants of integration omitted)

$$\int \frac{dx}{(x^2+a^2)^{3/2}} = \frac{x}{a^2\sqrt{x^2+a^2}}, \qquad \int \frac{x\,dx}{(x^2+a^2)^{3/2}} = -\frac{1}{\sqrt{x^2+a^2}},$$

$$\int \frac{dx}{\sqrt{x^2+a^2}} = \ln\!\left(x+\sqrt{x^2+a^2}\right), \qquad \int \frac{x\,dx}{\sqrt{x^2+a^2}} = \sqrt{x^2+a^2},$$

$$\int \frac{dx}{x^2+a^2} = \frac{1}{a}\tan^{-1}\!\left(\frac{x}{a}\right).$$

For a spherical shell of radius R and a point on its axis a distance $z > 0$ from its center, with $u = \cos\theta'$,

$$\int_{-1}^{1} \frac{du}{\sqrt{R^2+z^2-2Rzu}} = \begin{cases} 2/z, & z > R, \\ 2/R, & z < R, \end{cases} \qquad \int_{-1}^{1} \frac{(z-Ru)\,du}{(R^2+z^2-2Rzu)^{3/2}} = \begin{cases} 2/z^2, & z > R, \\ 0, & z < R. \end{cases}$$

[1] "CODATA recommended values of the fundamental physical constants: 2022," P. J. Mohr, D. B. Newell, B. N. Taylor, and E. Tiesinga, Rev. Mod. Phys. 97, 025002 (2025). https://doi.org/10.1103/RevModPhys.97.025002.