Introductory Physics II

Electricity, Magnetism and Optics

by

Robert G. Brown

Duke University Physics Department
Durham, NC 27708-0305
rgb@phy.duke.edu

Copyright Notice
Copyright Robert G. Brown 1993, 2007, 2013

Notice

This physics textbook is designed to support my personal teaching activities at Duke University, in particular teaching its Physics 141/142, 151/152, or 161/162 series (Introductory Physics for life science majors, engineers, or potential physics majors, respectively). It is freely available in its entirety in a downloadable PDF form or to be read online at:

http://www.phy.duke.edu/∼rgb/Class/intro_physics_2.php

It is also available in an ***inexpensive*** (really!) print version via Lulu press here:

http://www.lulu.com/shop/product-21025164.html

where readers/users can voluntarily help support or reward the author by purchasing either this paper copy or one of the even more inexpensive electronic copies.

By making the book available in these various media at a cost ranging from free to cheap, I enable the text can be used by students all over the world where each student can pay (or not) according to their means.

Nevertheless, I am hoping that students who truly find this work useful will ***purchase a copy through Lulu or a bookseller*** (when the latter option becomes available), if only to help subsidize me while I continue to write inexpensive textbooks in physics or other subjects.

This textbook is organized for ease of presentation and ease of learning. In particular, they are hierarchically organized in a way that directly supports efficient learning. They are also remarkably ***complete*** in their presentation and contain moderately detailed derivations of many of the important equations and relations from first principles while not skimping on simpler heuristic or conceptual explanations as well.

As a "live" document (one I actively use and frequently change, adding or deleting material or altering the presentation in some way), this textbook may have errors great and small, "stub" sections where I intend to add content at some later time but haven't yet finished it, and they cover and omit topics according to *my own* view of what is or isn't important to cover in a one-semester course. Expect them to change with little warning or announcement as I add content or correct errors.

Purchasers of the paper version should be aware of its probable imperfection and be prepared to either live with it or mark up their copy with corrections or additions as need be. The latest (and hopefully most complete and correct) version is always available for free online anyway, and people who have paid for a paper copy are ***especially*** welcome to access and retrieve it.

I cherish good-hearted communication from students or other instructors pointing out errors or suggesting new content (and have in the past done my best to implement many such corrections or suggestions).

Books by Robert G. Brown

Physics Textbooks

- *Introductory Physics I and II*

  A lecture note style textbook series intended to support the teaching of introductory physics, with calculus, at a level suitable for Duke undergraduates.

- *Classical Electrodynamics*

  A lecture note style textbook intended to support the second semester (primarily the dynamical portion, little statics covered) of a two semester course of graduate Classical Electrodynamics.

Computing Books

- *How to Engineer a Beowulf Cluster*

  An online classic for years, this is the print version of the famous free online book on cluster engineering. It too is being actively rewritten and developed, no guarantees, but it is probably still useful in its current incarnation.

Fiction

- *The Book of Lilith*

  ISBN: 978-1-4303-2245-0
  Web: http://www.phy.duke.edu/∼rgb/Lilith/Lilith.php

  Lilith is the *first* person to be given a soul by God, and is given the job of giving all the things in the world souls by loving them, beginning with Adam. Adam is given the job of making up rules and the definitions of sin so that humans may one day live in an ethical society. Unfortunately Adam is weak, jealous, and greedy, and insists on being on *top* during sex to "be closer to God".

  Lilith, however, refuses to be second to Adam or anyone else. *The Book of Lilith* is a funny, sad, satirical, uplifting tale of her spiritual journey through the ancient world soulgiving and judging to find at the end of that journey – herself.

- *The Fall of the Dark Brotherhood*

  ISBN: 978-1-4303-2732-5
  Web: http://www.phy.duke.edu/∼rgb/Gods/Gods.php

  A straight-up science fiction novel about an adventurer, Sam Foster, who is forced to flee from a murder he did not commit across the multiverse. He finds himself on a primitive planet and gradually becomes embroiled in a parallel struggle against the world's pervasive slave culture and the cowled, inhuman agents of an immortal of the multiverse that support it. Captured by the resurrected clone of its wickedest agent and horribly mutilated, only a pair of legendary swords and his native wit and character stand between Sam, his beautiful, mysterious partner and a bloody death!

Poetry

- *Who Shall Sing, When Man is Gone*

  Original poetry, including the epic-length poem about an imagined end of the world brought about by a nuclear war that gives the collection its name. Includes many long and short works on love and life, pain and death.

  > Ocean roaring, whipped by storm
  > in damned defiance, hating hell
  > with every wave and every swell,
  > every shark and every shell
  > and shoreline.

- *Hot Tea!*

  More original poetry with a distinctly Zen cast to it. Works range from funny and satirical to inspiring and uplifting, with a few erotic poems thrown in.

  > Chop water, carry
  > wood. Ice all around,
  > fire is dying. Winter Zen?

All of these books can be found on the online Lulu store here:

http://stores.lulu.com/store.php?fAcctID=877977

The Book of Lilith is available on Amazon, Barnes and Noble and other online bookseller websites.

Contents

I: Preliminaries **xiii**

Preface **xiii**

Textbook Layout and Design **xv**

Getting Ready to Learn Physics **3**

See, Do, *Teach* . . . 3

Other Conditions for Learning . . . 8

Your Brain and Learning . . . 14

How to Do Your Homework Effectively . . . 21

The Method of Three Passes . . . 25

Week 0: Math Needed for Introductory E&M (and Optics) **29**

0.1: Coordinate Frames . . . 31

0.1.1: Cartesian Coordinates . . . 31

Example 0.1.1: Integrating a Function Along a Line in Cartesian Coordinates . . . 33

Example 0.1.2: Integrating a Function over an Area in Cartesian Coordinates . . . 33

Example 0.1.3: Integrating a Function over a Volume in Cartesian Coordinates . . . 35

0.1.2: Cylindrical Coordinates . . . 35

Example 0.1.4: Finding the Area of a Soup Can Label . . . 38

Example 0.1.5: Volume of a Cylinder . . . 38

Example 0.1.6: Finding the Volume of a Right Circular Cone . . . 39

Example 0.1.7: Evaluating a Volume Charge Density . . . 40

0.1.3: Spherical Polar Coordinates . . . 40

Example 0.1.8: Finding the Area of a Sphere . . . 42

Example 0.1.9: Integrating a Function of $\cos\theta$ Over a Spherical Surface . . . 43

Example 0.1.10: Integrating the Volume of a Sphere . . . 44

Example 0.1.11: Integrating a Radial Distribution over a Sphere 44
Example 0.1.12: Evaluating the Moment of Inertia of a Uniform Sphere 45

Summary **47**

II: Electrostatics **51**

Week 1: Discrete Charge and the Electrostatic Field **51**
Summary 51
1.1: Charge 55
1.1.1: Charge Quantization and Elementary Particles 58
1.1.2: Coarse-Graining and Charge Density 62
1.1.3: Insulators, Conductors, Semiconductors 64
1.2: Coulomb's Law 66
1.3: Electrostatic Field 68
1.4: The Superposition Principle 70
Example 1.4.1: Finding the Field of Two Point Charges – An 'Electric Dipole' . . 71
1.5: Electric Dipoles 74
1.5.1: Force and Torque Acting on a Dipole 76
1.5.2: Electric Field of a Dipole 79
Example 1.5.1: Find the field of a y-directed electric dipole at an arbitrary point on the x-axis *in the limit*
Homework for Week 1 82

Week 2: Continuous Charge and Gauss's Law **91**
2.1: The Field of Continuous Charge Distributions 93
2.1.1: Coarse-Graining and Charge Density Revisited 93
2.1.2: Using Calulus to Fine $\vec{E}(\vec{r})$ from a Charge Distribution 93
Example 2.1.1: Circular Loop of Charge 95
Example 2.1.2: Long Straight Line of Charge 98
Example 2.1.3: Circular Disk of Charge 100
2.2: Gauss's Law for the Electrostatic Field 104
2.3: Using Gauss's Law to Evaluate the Electric Field 110
Example 2.3.1: Spherical: A spherical shell of charge 111
Example 2.3.2: Advanced: Spherical Shell of Charge 113
Example 2.3.3: Electric Field of a Solid Sphere of Charge 117

Example 2.3.4: Cylindrical: A cylindrical shell of charge 120
Example 2.3.5: Planar: A sheet of charge 123
2.4: Gauss's Law and Conductors 124
2.4.1: Properties of Conductors 124
Example 2.4.1: Field and Charge Distribution of a Blob of Conductor 128
Example 2.4.2: Two Thick Plates Plus Wires (Capacitor) 129
Creating Charged Objects 130
Homework for Week 2 134

Week 3: Potential Energy and Potential 145
3.1: Electrostatic Potential Energy 146
3.2: Potential 147
3.3: Superposition 149
3.3.1: Deriving or Computing the Potential 149
3.4: Examples of Computing the Potential 151
Example 3.4.1: Potential of a Dipole on the x-axis 151
Example 3.4.2: Potential of a Dipole at an Arbitrary Point in Space 153
Example 3.4.3: A ring of charge 155
Example 3.4.4: Potential of a Spherical Shell of Charge 157
Example 3.4.5: Advanced: Spherical Shell of Charge 159
Example 3.4.6: Potential of a Uniform Ball of Charge 160
Example 3.4.7: Potential of an Infinite Line of Charge 162
3.4.1: Potential of an Infinite Plane of Charge 163
3.5: The Potential Energy of Charge Distributions 164
3.5.1: The Potential Energy of Multiple Point Charges 164
Example 3.5.1: The Potential Energy of Four Charges in a Square 165
3.5.2: The Potential Energy of Continuous Charge Distributions 166
Example 3.5.2: Potential Energy of a Uniform Ball of Charge 168
Example 3.5.3: Potential Energy of a Uniform Ball of Charge – Second Method . 170
Example 3.5.4: Potential Energy of a Spherical Shell of Charge 171
3.5.3: Self-energy of a 'Point Charge' 171
3.6: Conductors in Electrostatic Equilibrium 173
3.6.1: Charge Sharing 173
3.7: Dielectric Breakdown 174

Homework for Week 3 177

Week 4: Capacitance **183**

4.1: Capacitance 185

4.1.1: Computing the Capacitance: the Parallel Plate Capacitor 187

Example 4.1.1: Cylindrical Capacitor 190

Example 4.1.2: Spherical Capacitor 191

4.2: Energy of a Charged Capacitor 192

4.2.1: Energy Density 194

4.3: Adding Capacitors in Series and Parallel 195

4.4: Dielectrics 198

Example 4.4.1: The Lorentz Model for an Atom 199

4.4.1: Dielectric Response of an Insulator in an Electric Field 201

4.4.2: Dielectrics, Bound Charge, and Capacitance 206

Homework for Week 4 212

Week 5: Resistance **217**

5.1: Batteries and Voltage Sources 220

5.1.1: Chemical Batteries 220

5.1.2: The Symbol for a Battery 222

5.1.3: Batteries and Renewable Energy 222

5.2: Resistance and Ohm's Law 224

5.2.1: A Simple Linear Conduction Model 225

5.2.2: Current Density and Charge Conservation 227

5.2.3: Advanced: Differential Form and Maxwell's Equations 229

5.2.4: The Drude Model 230

5.2.5: Advanced: Details of the Drude Model 233

5.2.6: Ohm's Law 236

5.2.7: Dependence of Resistivity on Temperature 239

5.3: Resistances in Series and Parallel 241

5.3.1: Series 241

5.3.2: Parallel 242

5.4: Kirchhoff's Rules and Multiloop Circuits 243

5.4.1: Kirchhoff's Loop Rule 244

5.4.2: Kirchhoff's Junction Rule 245

Example 5.4.1: The Internal Resistance of a Battery 245
Example 5.4.2: A Multiloop Resistance Problem 247
5.5: RC Circuits 250
Example 5.5.1: Discharging Capacitor 250
Example 5.5.2: Charging Capacitor 252
Homework for Week 5 256

III: Magnetostatics 263

Week 6: Moving Charges and Magnetic Force 263
6.1: Magnetic Force versus Magnetic Field 264
6.2: Magnetic Force on a Moving Point Charge 265
Example 6.2.1: A Charged Particle Moving in a Uniform Magnetic Field 267
Example 6.2.2: The Cyclotron 268
Example 6.2.3: Cloud Chamber 269
Example 6.2.4: Region of Crossed Fields 270
Example 6.2.5: Thomson's Apparatus for measuring e/m 271
Example 6.2.6: The Mass Spectrometer 275
Example 6.2.7: The Hall Effect 277
6.3: The Magnetic Force on Continuous Currents 279
Example 6.3.1: The Magnetic Force and Torque on a Rectangular Current Loop (Magnetic D
Example 6.3.2: The Magnetic Moment of an *Arbitrary* Plane Current Loop 282
6.4: Potential Energy of a Magnetic Dipole 283
6.4.1: Advanced: Comment on Magnetic Fields and Work/Potential Energy . . . 284
6.5: The Magnetic Moments of Rotating Charged Objects 286
Example 6.5.1: The Magnetic Moment of a Rotating Ring of Mass and Charge . 287
Example 6.5.2: Magnetic Moment of a Rotating Charged Massive Disk 287
6.6: The Precession of Magnetic Moments: Magnetic Resonance 289
6.6.1: Advanced: Spin Echoes and Magnetic Resonance Imaging 293
Homework for Week 6 302

Week 7: Sources of the Magnetic Field 307
7.1: Gauss's Law for Magnetism 309
7.1.1: The Units of Magnetic Flux 312

7.2: The Biot-Savart Law 312
7.3: Examples of Using the Biot-Savart Law to Find the Magnetic Field 316
Example 7.3.1: Magnetic Field of a Straight Wire Segment 316
Example 7.3.2: Field of a Circular Loop on its Axis 318
Example 7.3.3: Field of a Revolving Ring of Charge on its Axis 320
7.4: The Magnetic Field of a Point Charge 321
7.4.1: Finite Field Propagation Speed for E and B 322
7.4.2: Violation of Newton's Third Law 323
7.5: Ampere's Law 324
7.6: Applications of Ampere's Law 328
Example 7.6.1: Cylindrical Current Density – Infinitely Long Thin Wire 329
Example 7.6.2: Cylindrical Current Density – Field of an Infinitely Long Thick Wire329
Example 7.6.3: The Solenoid 331
Example 7.6.4: Toroidal Solenoid 333
Example 7.6.5: Infinite Sheet of Current 335
7.7: Concluding Discussion 336
Homework for Week 7 338

IV: Electrodynamics 345

Week 8: Faraday's Law and Induction 345
8.1: Magnetic Forces and Moving Conductors 347
8.2: The Rod on Rails 350
8.2.1: Problem and Solution 354
8.2.2: The Magnetic Field and Work 356
8.3: Faraday's Law 359
8.4: Lenz's Law 362
8.4.1: Lenz's Law for changing C 362
8.4.2: Lenz's Law for changing B (magnitude) 363
8.4.3: Lenz's Law for changing the $\vec{B}$ and/or $\hat{n}$ direction 364
8.4.4: Lenz's Law Summary 364
Example 8.4.1: Wire and Rectangular Loop – Direction Only 365
Example 8.4.2: Rectangular Loop Pulled from Field 367
8.5: More Rod on Rails Problems 368

Example 8.5.1: Rod on Rails with Battery . . . 368
8.6: Inductance . . . 370
Example 8.6.1: The Mutual Inductance of a Wire and Rectangular Current Loop 373
8.7: Self-Induction . . . 375
Example 8.7.1: The Self-Inductance of the Solenoid . . . 376
Example 8.7.2: Toroidal Solenoid . . . 378
Example 8.7.3: Coaxial Cable . . . 379
8.8: LR Circuits . . . 380
8.8.1: Power . . . 382
8.9: Magnetic Energy . . . 383
Example 8.9.1: Energy in a Toroidal Solenoid . . . 385
8.10: Eddy Currents . . . 385
8.11: Magnetic Materials . . . 387
Diamagnetism . . . 387
8.11.1: Superconductors . . . 389
Paramagnetism . . . 390
Ferromagnetism and Antiferromagnetism . . . 391
8.11.2: The Curie Temperature and Neel Temperature . . . 393
Magnetism, Concluded . . . 394
Homework for Week 8 . . . 397

Week 9: Maxwell's Equations and Light 403

9.1: Ampere's Law and the Maxwell Displacement Current . . . 409
9.1.1: The Problem with Ampere's Law – So Far . . . 410
9.1.2: The Invariant Current through S/C . . . 413
Example 9.1.1: The Magnetic Field Inside a Parallel Plate Capacitor . . . 416
9.2: Advanced Topic: Origins of the Magnetic Field . . . 418
9.3: Maxwell's Equations for the Electromagnetic Field: The Wave Equation . . . 421
9.3.1: The Wave Equation . . . 422
9.4: Light as a Harmonic Wave . . . 426
9.5: The Poynting Vector . . . 429
9.6: Radiation Pressure and Momentum . . . 432
9.7: Sources of the Electromagnetic Field (Advanced/Optional) . . . 436
9.7.1: Larmor Radiation from Accelerating Charges . . . 437

9.7.2: Dipole Radiation 442
9.7.3: Why The Sky is Blue 443
Example 9.7.1: The Dipole Antenna 444
Homework for Week 9 446

I Optics **453**

Week 10: Light **455**
10.1: The Speed of Light 458
10.2: The Spectrum 459
10.3: The Law of Reflection 461
10.4: Snell's Law 462
10.4.1: Fermat's Principle 463
10.4.2: Total Internal Reflection, Critical Angle 467
10.4.3: Dispersion 468
10.5: Polarization 469
10.5.1: Unpolarized Light 470
10.5.2: Linear Polarization 470
10.5.3: Circularly Polarized Light 471
10.5.4: Elliptically Polarized Light 472
10.5.5: Polarization by Absorption (Malus's Law) 472
10.5.6: Polarization by Scattering 473
10.5.7: Polarization by Reflection 474
10.5.8: Polaroid Sunglasses 475
10.6: Doppler Shift 475
10.6.1: Moving Source 475
10.6.2: Moving Receiver 476
10.6.3: Moving Source and Moving Receiver 477
10.6.4: The Relativistic Doppler Shift 477
Homework for Week 10 480

Week 11: Lenses and Mirrors **485**
11.1: Vision and Plane Mirrors 488
11.2: Curved Mirrors 490

11.3: Ray Diagrams for Ideal Mirrors . . . 493
11.4: Lenses . . . 497
11.5: Multiple Lenses and Diopters . . . 499
11.5.1: Diopters . . . 500
11.6: The Eye . . . 503
11.7: Optical Instruments . . . 505
11.7.1: The Simple Magnifier . . . 505
11.7.2: Telescope . . . 507
11.7.3: Microscope . . . 510
Homework for Week 11 . . . 513

Week 12: Interference and Diffraction 519
12.1: Harmonic Waves and Superposition . . . 524
12.1.1: Hot Sources and Wave Coherence . . . 525
12.1.2: Combining Coherent Harmonic Waves . . . 529
12.2: Interference from Two Narrow Slits . . . 530
12.3: Interference from 2, 3, ... N Narrow Slits . . . 534
12.3.1: Principle Maxima, Minima, and Secondary Maxima . . . 539
12.3.2: Finding the Maxima and Minima Exactly . . . 543
12.4: The Diffraction Grating – Rayleigh's Criterion for Resolution . . . 546
12.4.1: Rayleigh's Criterion for Resolution . . . 549
12.4.2: Resolving Power . . . 549
12.5: Diffraction . . . 550
12.6: Diffraction Minima, Heuristic Rule . . . 552
12.7: Exact Solution to Diffraction by a Single Slit . . . 553
Example 12.7.1: Diffraction Pattern of a Slit of Width $a = 4\lambda$ . . . 560
12.8: Two Slits of Finite Width . . . 562
Example 12.8.1: Two Slits of Separation $d = 8\lambda$ and width $a = 4\lambda$ . . . 562
12.9: Diffraction Through Circular Apertures – Limitations on Optical Instruments . . . 563
12.10: Thin Film Interference . . . 566
12.10.1: Phase Shift Due to Path Difference *in the Thin Film!* . . . 568
12.10.2: Phase Shifts Due to Reflections at the Surfaces . . . 569
12.10.3: No Relative Phase Shift from Surface Reflections . . . 570
12.10.4: A Relative Phase Shift of π from Surface Reflections . . . 570

12.10.5: The Limits of *Very* Thin Films 571
Homework for Week 12 573

II Electronics **579**

Week 13: Alternating Current Circuits **581**
13.1: Introduction: Alternating Voltage 588
13.1.1: Electrical Distribution True Facts 589
13.1.2: The Transformer 591
13.1.3: Power Transmission 592
13.2: Passive AC Circuits 595
13.2.1: Non-driven LC circuit 596
13.2.2: Non-driven LRC circuit 597
13.3: Energy Loss in Passive LRC Circuits – Q-Factor 600
13.4: Active AC Circuits 602
13.4.1: A Harmonic AC Voltage Across a Basic Circuit Element 604
13.4.2: The Series LRC Circuit – Phasor Approach 607
13.4.3: A Universal description of the Current $I(\omega)$: Scaling into a Dimensionless Form612
13.4.4: Power in a Series LRC Circuit 615
13.4.5: The Parallel LRC Circuit 623
13.5: Filter Circuits 627
13.5.1: Low-Pass Filter $rLRC$ Circuits 628
13.5.2: High-Pass $rLRC$ Circuits 632
13.5.3: Band Pass Filters 634
13.6: The AM Radio and Bandwidth 635
13.7: Complex Representations of LRC Circuits (Advanced) 639
13.7.1: Single Circuit Elements – Complex Solution 640
13.7.2: Series LRC Circuit – Complex Solution 642
13.7.3: Parallel $rLRC$ Circuit – Complex Treatment 646
Homework for Week 13 650

I: Preliminaries

Preface

This introductory electromagnetism and optics text is intended to be used in the second semester of a two-semester series of courses teaching *introductory physics* at the college level, following a first semester course in (Newtonian) mechanics and thermodynamics. The text is intended to support teaching the material at a rapid, but *advanced* level – it was developed to support teaching introductory calculus-based physics to potential physics majors, engineers, and other natural science majors at Duke University over a period of more than twenty-five years.

Students who hope to succeed in learning physics from this text will need, as a minimum prerequisite, a solid grasp of mathematics. It is strongly recommended that all students have mastered mathematics at least through single-variable differential calculus (typified by the AB advanced placement test or a first-semester college calculus course). Students should also be *taking* (or have completed) single variable integral calculus (typified by the BC advanced placement test or a second-semester college calculus course). In the text it is presumed that students are competent in geometry, trigonometry, algebra, and single variable calculus; more advanced multivariate calculus is used in a number of places but it is taught in context as it is needed and is always "separable" into two or three independent one-dimensional integrals.

Many students are, unfortunately *weak* in their mastery of mathematics at the time they take physics. This enormously complicates the process of learning for them, especially if they are years removed from when they took their algebra, trig, and calculus classes as is frequently the case for pre-medical students. For that reason, several supplemental materials including an online textbook (work in progress) on the math required *specifically* to learn introductory physics quickly and efficiently, a short collection of "One Sheet Math Review" pages that cover individual requirements such as the "needed algebra" skills or "needed calculus" skills on a single side of a single sheet of paper, and finally, a chapter in this textbook that falls somewhere in between – more than the one sheet review, but far from a textbook and concentrating more on the *new* math required for E&M specifically.

The online book is located here:

http://www.phy.duke.edu/∼rgb/Class/math_for_intro_physics.php

The "One Sheet Math Review" is located here:

http://www.phy.duke.edu/∼rgb/Class/math_for_intro_physics.php

The chapter is located at the end of this section, right before we begin actual content.

Again, I *strongly suggest* that all students who are reading these words while *preparing* to begin studying physics pause for a moment, and at least visit the One Sheet Review site,

print out its pages, put them into their physics notebook, and go over them enough to feel comfortable with their content.

Note that *Getting Ready to Learn Physics* in this Preliminaries section is not part of the course *per se*, but I usually do a quick review of this material (as well as the course structure, grading scheme, and so on) in my first lecture of any given semester, the one where students are still finding the room, dropping and adding courses, and one cannot present real content in good conscience unless you plan to do it again in the second lecture as well. Students *greatly benefit* from guidance on how to study, as most enter physics thinking that they can master it with nothing but the memorization and rote learning skills that have served them so well for their many other fact-based classes. Of course this is completely false – physics is *reason* based and *conceptual* and it requires a very different pattern of study than simply staring at and trying to memorize lists of formulae or examples.

Students, however, should not count on their instructor doing this – they need to be self-actualized in their study from the beginning. It is therefore *strongly suggested* that all students read this preliminary chapter right away as their first "assignment" whether or not it is covered in the first lecture or assigned. In fact, (if you're just such a student reading these words) you can always decide to read it *right now* (as soon as you finish this Preface). It won't take you an hour, and might make as much as a full letter difference (to the good) in your final grade. What do you have to lose?

Even if you think that you are an excellent student and learn things totally effortlessly, I strongly suggest reading it. It describes a new perspective on the teaching and learning process supported by very recent research in neuroscience and psychology, and makes very specific suggestions as to the best way to proceed to learn physics.

Finally, the *Introduction* is a rapid summary of *the entire course!* If you read it and look at the pictures *before* beginning the course proper you can get a good conceptual overview of everything you're going to learn. If you *begin* by learning in a *quick* pass the broad strokes for the whole course, when you go through each chapter in all of its detail, all those facts and ideas have a place to live in your mind.

That's the primary idea behind this textbook – in order to be easy to remember, ideas need a house, a place to live. Most courses try to build you that house by giving you one nail and piece of wood at a time, and force you to build it in complete detail from the ground up.

Real houses aren't built that way at all! First a foundation is established, then the *frame of the whole house* is erected, and then, slowly but surely, the frame is wired and plumbed and drywalled and finished with all of those picky little details. It works better that way. So it is with learning.

Textbook Layout and Design

This textbook has a design that is just about perfectly backwards compared to most textbooks that currently cover the subject. Here are its primary design features:

- All mathematics required by the student is reviewed in a standalone, cross-referenced (free) work at the *beginning* of the book rather than in an appendix that many students never find.
- There are only *thirteen substantive chapters*. The book is organized so that it can be sanely taught in a *single college semester* with at *most* a chapter a week. I teach it in a five week summer session at the Duke Marine Lab in Beaufort, NC and (at three chapters a week plus startup and wind-down) that works too!
- It *begins* each chapter with an "abstract" and chapter summary. Detail, especially lecture-note style mathematical detail, follows the summary rather than the other way around.
- This text does *not* spend page after page trying to explain in English how physics works (prose which to my experience nobody reads anyway). Instead, a terse "lecture note" style presentation outlines the main points and presents considerable mathematical detail to support solving problems.
- Verbal and conceptual understanding *is*, of course, very important. It is expected to come from verbal instruction and discussion in the classroom and recitation and lab. This textbook *relies* on having a committed and competent instructor and a sensible learning process.
- Each chapter ends with a *short* (by modern standards) selection of *challenging* homework problems that are specifically chosen to *precisely span the primary concepts and examples*, often requiring a student to rederive for themselves things that were presented as primary content or examples in lecture. A good student might well get through *all of the problems in the book*, rather than at most 10% of them as is the general rule for other texts. Students that really, really want more problems to solve to shoot for an 'A' can look at can find them in a supplementary (online) book filled with nothing but problems, but students that can do the homework perfectly will almost certainly get a 'B' or better without them.
- The homework problems are weakly sorted out by level, as this text is intended to support non-physics science and pre-health profession students, engineers, and physics majors all three. The *material* covered is of course the same for all three, but the level of detail and difficulty of the math used and required is a bit different.

- The textbook is entirely algebraic in its presentation and problem solving requirements – with *very few exceptions* no calculators should be required to solve problems. The author assumes that any student taking physics is capable of punching numbers into a calculator, but it is *algebra* that ultimately determines the formula that they should be computing. Numbers are used in problems only to illustrate what "reasonable" numbers might be for a given real-world physical situation or where the problems cannot reasonably be solved algebraically (e.g. resistance networks).

Getting Ready to Learn Physics

See, Do, *Teach*

If you are reading this, I assume that you are either taking a course in physics or wish to learn physics on your own. If this is the case, I want to begin by teaching you the importance of your personal *engagement* in the learning process. If it comes right down to it, how well you learn physics, how good a grade you get, and how much *fun* you have all depend on how enthusiastically you tackle the learning process. If you remain disengaged, detatched from the learning process, you almost certainly will do poorly and be miserable while doing it. If you can find *any degree* of engagement – or open enthusiasm – with the learning process you will very likely do well, or at least as well as possible.

Note that I use the term *learning*, not *teaching* – this is to emphasize from the beginning that learning is a choice and that *you* are in control. Learning is active; being taught is passive. It is up to you to *seize control* of your own educational process and *fully participate*, not sit back and wait for knowledge to be forcibly injected into your brain.

You may find yourself stuck in a course that is taught in a traditional way, by an instructor that lectures, assigns some readings, and maybe on a good day puts on a little dog-and-pony show in the classroom with some audiovisual aids or some demonstrations. The standard expectation in this class is to sit in your chair and watch, passive, taking notes. No real engagement is "required" by the instructor, and lacking activities or a structure that encourages it, you lapse into becoming a lecture transcription machine, recording all kinds of things that make no immediate sense to you and telling yourself that you'll sort it all out later.

You may find yourself floundering in such a class – for good reason. The instructor presents an ocean of material in each lecture, and you're going to actually retain at most a few cupfuls of it functioning as a scribe and passively copying his pictures and symbols without first extracting their sense. And the lecture *makes* little sense, at least at first, and reading (if you do any reading at all) does little to help. Demonstrations can sometimes make one or two ideas come clear, but only at the expense of twenty other things that the instructor now has no time to cover and expects you to get from the readings alone. You continually postpone going over the lectures and readings to understand the material any more than is strictly required to do the homework, until one day a *big test* draws nigh and you realize that you really don't understand anything and have forgotten most of what you did, briefly, understand. Doom and destruction loom.

Sound familiar?

On the other hand, you may be in a course where the instructor has structured the course with a balanced mix of *open* lecture (held as a freeform discussion where questions aren't just encouraged but required) and group interactive learning situations such as a carefully structured recitation and lab where discussion and doing blend together, where students teach each other and use what they have learned in many ways and contexts. If so, you're lucky, but luck only goes so far.

Even in a course like this you may *still* be floundering because you may not understand *why* it is important for you to participate with your whole spirit in the quest to learn anything you ever choose to study. In a word, you simply may not give a rodent's furry behind about learning the material so that studying is always a fight with yourself to "make" yourself do it – so that no matter what happens, *you lose*. This too may sound very familiar to some.

The importance of engagement and participation in "active learning" (as opposed to passively being taught) is not really a new idea. Medical schools were four year programs in the year 1900. They are four year programs today, where the amount of information that a physician must now master in those four years is probably *ten times greater* today than it was back then. Medical students are necessarily among the most efficient learners on earth, or they simply cannot survive.

In medical schools, the optimal learning strategy is compressed to a three-step adage: See one, do one, teach one.

See a procedure (done by a trained expert).

Do the procedure yourself, with the direct supervision and guidance of a trained expert.

Teach a student to do the procedure.

See, do, teach. Now you *are* a trained expert (of sorts), or at least so we devoutly hope, because that's all the training you are likely to get until you start doing the procedure over and over again with real humans and with limited oversight from an attending physician with too many other things to do. So you practice and study on your own until you achieve real mastery, because a mistake can *kill* somebody.

This recipe is quite general, and can be used to increase *your own* learning in almost *any* class. In fact, lifelong success in learning with or without the guidance of a good teacher is a matter of discovering the importance of *active engagement and participation* that this recipe (non-uniquely) encodes. Let us rank learning methodologies in terms of "probable degree of active engagement of the student". By probable I mean the degree of active engagement that I as an instructor have observed in students over many years and which is significantly reinforced by research in teaching methodology, especially in physics and mathematics.

Listening to a lecture as a transcription machine with your brain in "copy machine" mode is almost entirely passive and is for *most* students *probably* a nearly complete waste of time. That's not to say that "lecture" in the form of an organized presentation and review of the material to be learned isn't important or is completely useless! It serves one *very important purpose* in the grand scheme of learning, but by being passive *during* lecture *you* cause it to fail in its purpose. Its purpose is *not* to give you a complete, line by line transcription of the words of your instructor to ponder later and alone. It is to convey, for a brief shining moment, the *sense* of the *concepts* so that you *understand them*.

It is difficult to sufficiently emphasize this point. If lecture doesn't make sense *to you* when the instructor presents it, you will have to work much harder to achieve the sense of the material "later", if later ever comes at all. If you fail to identify the important concepts during the presentation and see the lecture as a string of disconnected facts, you will have to remember *each* fact as if it were an abstract string of symbols, placing impossible demands on your memory even if you are extraordinarily bright. If you fail to achieve some degree of understanding (or *synthesis* of the material, if you prefer) in lecture by asking questions and getting expert explanations on the spot, you will have to build it later out of your notes on a set of abstract symbols that made no sense to you at the time. You might as well be trying to translate Egyptian Hieroglyphs without a Rosetta Stone, and the best of luck to you with *that*.

Reading is a bit more active – at the very least your brain is more likely to be somewhat engaged if you aren't "just" transcribing the book onto a piece of paper or letting the words and symbols happen in your mind – but is still pretty passive. Even watching nifty movies or cool-ee-oh demonstrations is basically sedentary – you're still just sitting there while somebody or something *else* makes it all happen in your brain while you aren't *doing* much of anything. At best it grabs your attention a bit better (on average) than lecture, but *you* are mentally *passive*.

In all of these forms of learning, the single active thing you are likely to be doing is taking notes or moving an eye muscle from time to time. For better or worse, the human brain isn't designed to learn well in passive mode. Parts of your brain are likely to take charge and pull your eyes irresistably to the window to look outside where *active* things are going on, things that might not be so damn *boring!*

With your active engagement, with your taking charge of and participating in the learning process, things change dramatically. Instead of passively listening in lecture, you can at least *try* to ask questions and initiate discussions whenever an idea is presented that makes no intial sense to you. Discussion is an *active* process even if you aren't the one talking at the time. *You participate!* Even a tiny bit of participation in a classroom setting where students are constantly asking questions, where the instructor is constantly answering them and asking the students questions in turn makes a huge difference. Humans being social creatures, it also makes the class a lot more fun!

In summary, sitting on your ass[1] and writing meaningless (to you, so far) things down as somebody says them in the hopes of being able to "study" them and discover their meaning on your own later is *boring* and for most students, later never comes because you are busy with *many* classes, because you haven't discovered anything beautiful or exciting (which is the *reward* for figuring it all out – if you ever get there) and then there is partying and hanging out with friends and having *fun*. Even if you do find the time and really want to succeed, in a complicated subject like physics you are less likely to be *able* to discover the meaning on your own (unless you are *so bright* that learning methodology is irrelevant and you learn in a single pass no matter what). Most introductory students are swamped by the details, and have small chance of discovering the *patterns* within those details that constitute "making sense" and make the detailed information *much, much easier to learn* by enabling a compression of the detail into a much smaller set of connected ideas.

Articulation of ideas, whether it is to yourself or to others in a discussion setting, *requires* you to create tentative patterns that might describe and organize all the details you are being

[1] I mean, of course, your donkey. What did you think I meant?

presented with. Using those patterns and applying them to the details as they are presented, you naturally encounter places where your tentative patterns are wrong, or don't quite work, where something "doesn't make sense". In an "active" lecture students participate in the process, and can ask questions and kick ideas around until they *do* make sense. Participation is also *fun* and helps you pay far more attention to what's going on than when you are in passive mode. It may be that this increased attention, this consideration of many alternatives and rejecting some while retaining others with social reinforcement, is what makes all the difference. To learn optimally, even "seeing" must be an active process, one where you are not a vessel waiting to be filled through your eyes but rather part of a team studying a puzzle and looking for the patterns *together* that will help you eventually solve it.

Learning is increased still further by *doing*, the very essence of activity and engagement. "Doing" varies from course to course, depending on just what there is for you to do, but it always is the *application* of what you are learning to some sort of activity, exercise, problem. It is *not* just a recapitulation of symbols: "looking over your notes" or "(re)reading the text". The symbols for any given course of study (in a physics class, they very likely will *be* algebraic symbols for real although I'm speaking more generally here) do not, initially, mean a lot to you. If I write $\vec{F} = q(\vec{v} \times \vec{B})$ on the board, it means a great deal to *me*, but if you are taking this course for the first time it probably means zilch to *you*, and yet I pop it up there, draw some pictures, make some noises that hopefully make sense to you at the time, and blow on by. Later you read it in your notes to try to recreate that sense, but you've *forgotten* most of it. Am I describing the income I expect to make selling $\vec{B}$ tons of barley with a market value of $\vec{v}$ and a profit margin of q?

To *learn* this expression (for yes, this is a force law of nature and one that we very much must learn this semester) we have to learn what the symbols stand for – q is the charge of a point-like object in motion at velocity $\vec{v}$ in a magnetic field $\vec{B}$, and $\vec{F}$ is the resulting force acting on the particle. We have to learn that the $\times$ symbol is the *cross product of evil* (to most students at any rate, at least at first). In order to get a *gut feeling* for what this equation represents, for the directions associated with the cross product, for the trajectories it implies for charged particles moving in a magnetic field in a variety of contexts one has to *use* this expression to solve problems, *see* this expression in action in laboratory experiments that let you prove to yourself that it isn't bullshit and that the world really does have cross product force laws in it. You have to do your homework that involves this law, and be fully engaged.

The learning process isn't exactly linear, so if you participate fully in the discussion and the doing while going to even the most traditional of lectures, you have an excellent chance of getting to the point where you can score anywhere from a 75% to an 85% in the course. In most schools, say a C+ to B+ performance. Not bad, but not really excellent. A few students will still get A's – they either work extra hard, or really like the subject, or they have some sort of secret, some way of getting over that barrier at the 90's that is only crossed by those that really do understand the material quite well.

Here is the secret for getting *yourself* over that 90% hump, even in a physics class (arguably one of the most difficult courses you can take in college), even if you're *not* a super-genius (or have never managed in the past to learn like one, a glance and you're done): *Work in groups!* In fact, a *really* good course (in my opinion) is one where the entire learning process is organized around student ***teams***, basically carefully contructed, semi-permanent groups

where each member is at least partly responsible for the effective learning of all the team members, not just themselves!

That's it. Nothing really complex or horrible, just get together with your friends who are also taking the course and do your homework *together*. In a well designed physics course (and many courses in mathematics, economics, and other subjects these days) you'll have *some* aspects of the class, such as a recitation or lab, where you are *required* to work in groups/teams, and the teams and team activities may be highly structured or freeform.

"Studio" or "Team Based Learning" for teaching physics have even interleaved the lecture itself with team-based active learning, so *everything* is done in teams. This makes it it ***nearly impossible*** to be disengaged and sit passively in class waiting for learning to "happen". It also yields measureable improvements (all things being equal) on at least some objective instruments for measurement of learning, although (long story) measuring learning is a lot harder than you might think...

If you take charge of your own learning, though, you will quickly see that in *any* course, however it is formally organized and taught, *you can study in a group!* This is true even in a course where "the homework" is to be done alone by fiat of the (unfortunately ignorant and misguided) instructor. Just study "around" the actual assignment – assign *yourselves* problems "like" the actual assignment – most textbooks have plenty of extra problems and then there is the Internet and other textbooks – and do them in a group, then (afterwards!) break up and do your actual assignment alone. Note that if you use a completely different textbook to pick your group problems from and do them together before *looking* at your assignment in *your* textbook, you can't even be blamed if some of the ones you pick turn out to be ones your instructor happened to assign.

Oh, and not-so-subtly – give the instructor a (link to a) PDF copy of this book (it's as free for instructors as it is for students, after all, just a click away on the Internet). Who knows? Maybe they will give some of these ideas a try!

Let's understand in more detail *why* working on hard problems in teams often has a dramatic effect on learning. What happens when a team works together? Well, a lot of *discussion* happens, because humans working on a common problem like to talk. There is plenty of *doing* going on, presuming that the group has a common task list to work through, like a small mountain of really difficult problems that nobody can possibly solve working on their own and are *barely* within their abilities working as a group backed up by the course instructor! Finally, in team-based learning everybody has the opportunity to *teach!*

The importance of teaching – not only seeing the lecture presentation with your whole brain actively engaged and participating in an ongoing discussion so that it makes sense at the time, not only doing lots of homework problems and exercises that apply the material in some way, but *articulating* what you have discovered in this process and *answering questions* that force you to consider and reject alternative solutions or pathways (or not) cannot be overemphasized. Teaching each other in a peer setting (ideally with mentorship and oversight to keep you from teaching each other *mistakes*) is *essential!*

This problem you "get", and teach *others* (and actually learn it better from teaching it than they do from your presentation – never begrudge the effort required to teach your fellow team members even if some of them are very slow to understand). The next problem you don't get

but some *other* group member does – they get to teach *you*. In the end you all learn *far more* about every problem as a consequence of the struggle, the exploration of false paths, the discovery and articulation of the correct path, the process of discussion, resolution and agreement in teaching whereby *everybody* in the team hopefully reaches full understanding.

Note that success in this last key metric depends on ***you*** and you alone. No teaching/learning approach will help you learn if you quit halfway there. Some approaches make it easier, some harder, but in the end *you* bear the ultimate responsibility for your own active, engaged learning. When you have completed see, do, *teach*, you have achieved a critical milestone on the path to comprehension.

I would assert that it is all but *impossible* for someone to become a (halfway decent) teacher of *anything* without learning along the way that the absolute best way to learn *any* set of material deeply is to *teach* it – it is the very foundation of Academe and has been for two or three thousand years. It is, as we have noted, built right into the intensive learning process of medical school and graduate school in general. For some reason, however, we don't incorporate a teaching component in most *undergraduate* classes, which is a shame, and it is basically nonexistent in nearly all K-12 schools, which is an open tragedy.

As an engaged student *you don't have to live with that!* Put it there yourself, by incorporating group study and mutual teaching into your learning process *with or without the help or permission of your teachers!* A really smart and effective team soon learns to *iterate* the teaching – I teach you, and to make sure you got it you *immediately* use the material I taught you and try to articulate it back to me. Eventually everybody on the team understands, everybody on the team benefits, *everybody on the team gets the best possible grade on the material*. This process will actually make you (quite literally) more intelligent. You may or may not manage to lock down an A, but you will get the best grade you are capable of getting, for your given investment of effort.

This is close to the ultimate in engagement – highly active learning, with all cylinders of your brain firing away on the process. You can *see* why learning is enhanced. It is simply a bonus, a sign of a just and caring God, that it is also a lot more *fun* to work in a team, especially in a relaxed context with food and drink present. Yes, I'm encouraging you to have "physics study parties" (or history study parties, or psychology study parties). Hold contests. Give silly prizes. See. Do. Teach.

Other Conditions for Learning

Learning isn't *only* dependent on the engagement pattern implicit in the See, Do, Teach rule. Let's absorb a few more True Facts about learning, in particular let's come up with a handful of things that can act as "switches" and turn your ability to learn on and off quite independent of how your instructor structures your courses. Most of these things aren't *binary* switches – they are more like dimmer switches that can be slid up between dim (but not off) and bright (but not fully on). Some of these switches, or environmental parameters, act together more powerfully than they act alone. We'll start with the most important pair, a pair that research has shown work together to potentiate or block learning.

Instead of just telling you what they are, arguing that they are important for a paragraph

or six, and moving on, I'm going to give you an early opportunity to *practice* active learning in the context of reading a chapter on active learning. That is, I want you to participate in a tiny mini-experiment. It works a little bit better if it is done verbally in a one-on-one meeting, but it should still work well enough even if it is done in this text that you are reading.

I going to give you a string of ten or so digits and ask you to glance at it one time for a count of three and then look away. No fair peeking once your three seconds are up! Then I want you to do something else for at least a minute – anything else that uses your whole attention and interrupts your ability to rehearse the numbers in your mind in the way that you've doubtless learned permits you to learn other strings of digits, such as holding your mind blank, thinking of the phone numbers of friends or your social security number. Even rereading this paragraph will do.

At the end of the minute, try to recall the number I gave you and write down what you remember. Then turn back to right here and compare what you wrote down with the actual number.

Ready? (No peeking yet...) Set? Go!

Ok, here it is, in a footnote at the bottom of the page to keep your eye from naturally reading ahead to catch a glimpse of it while reading the instructions above[2].

How did you do?

If you are like most people, this string of numbers is a bit too long to get into your immediate memory or visual memory in only three seconds. There was very little time for rehearsal, and then you went and did something else for a bit right away that was supposed to *keep* you from rehearsing whatever of the string you *did* manage to verbalize in three seconds. Most people will get anywhere from the first three to as many as seven or eight of the digits right, but probably not in the correct order, unless...

...they are particularly smart or lucky and in that brief three second glance have time to notice that the number consists of all the digits used exactly once! Folks that happened to "see" this at a glance probably did better than average, getting all of the correct digits but maybe in not quite the correct order.

People who are downright *brilliant* (and equally lucky) realized in only three seconds (without cheating an extra second or three, you know who you are) that it consisted of the string of odd digits in ascending order followed by the even digits in descending order. Those people probably got it *all perfectly right* even without time to rehearse and "memorize" the string! Look again at the string, see the pattern now?

The moral of this little mini-demonstration is that it is *easy* to overwhelm the mind's capacity for processing and remembering "meaningless" or "random" information. A string of ten measly (apparently) random digits is too much to remember for one lousy minute, especially if you aren't given time to do rehearsal and all of the other things we have to make ourselves do to "memorize" meaningless information.

Of course things *changed radically* the instant I pointed out the pattern! At this point you could very likely go away and come back to this point in the text *tomorrow* or even *a year from now* and have an *excellent* chance of remembering this particular digit string, because it

[2] 1357986420 (one, two, three, quit and do something else for one minute...)

makes sense of a sort, and there are plenty of cues in the text to trigger recall of the particular pattern that "compresses and encodes" the actual string. You don't have to remember *ten* random things at all – only two and a half – odd ascending digits followed by the opposite (of both). Patterns rock!

This example has obvious connections to lecture and class time, and is one reason retention from lecture is so lousy. For *most* students, lecture in any nontrivial college-level course is a long-running litany of stuff they don't know yet. Since it is all new to them, it might as well be random digits as far as their cognitive abilities are concerned, at least at first. Sure, there is pattern there, but you have to *discover* the pattern, which requires *time* and a certain amount of *meditation* on all of the information. Basically, you have to have a chance for the pattern to jump out of the stream of information and punch the switch of the damn light bulb we all carry around inside our heads, the one that is endlessly portrayed in cartoons. That light bulb experience is *real* – it actually exists, in more than just a metaphorical sense – and if you study long enough and hard enough to obtain a sudden, epiphinaic realization in any topic you are studying, however trivial or complex (like the pattern exposed above) it is quite likely to be accompanied by a purely mental flash of "light". You'll know it when it happens to you, in other words, and it feels *great*.

Unfortunately, the instructor doesn't usually give students a *chance* to experience this in lecture. No sooner is one seemingly random factoid laid out on the table than along comes a new, apparently disconnected one that pushes it out of place long before we can either memorize it the hard way or make sense out of it so we can remember it with a lot less work. This isn't really anybody's fault, of course; the light bulb is quite unlikely to go off in lecture *just* from lecture no matter *what* you or the lecturer do – it is something that happens to the prepared mind at the end of a process, not something that just fires away every time you hear a new idea.

The humble and unsurprising conclusion I want you to draw from this silly little mini-experiment is that *things are easier to learn when they make sense!* A *lot* easier. In fact, things that don't make sense to you are never "learned" – they are at best memorized. Information can almost always be *compressed* when you discover the patterns that run through it, especially when the patterns all fit together into the marvelously complex and beautiful and mysterious process we call "deep understanding" of some subject.

There is one more example I like to use to illustrate how important this information compression is to memory and intelligence. I play chess, badly. That is, I know the legal moves of the game, and have no idea at all how to use them effectively to improve my position and eventually win. Ten moves into a typical chess game I can't recall how I got myself into the mess I'm typically in, and at the end of the game I probably can't remember *any* of what went on except that I got trounced, again.

A chess *master*, on the other hand, can play umpty games at once, blindfolded, against pitiful fools like myself and when they've finished winning them all they can go back and reconstruct *each one* move by move, criticizing each move as they go. Often they can remember the games in their entirety days or even years later.

This isn't just because they are *smarter* – *they* might be completely unable to derive the Lorentz group from first principles, and I can, and this doesn't automatically make me smarter

than them either. It is because chess makes *sense* to them – they've achieved a deep understanding of the game, as it were – and they've built a complex meta-structure memory in their brains into which they can poke chess moves so that they can be retrieved extremely efficiently. This gives them the *attendant* capability of searching vast portions of the game tree at a glance, where I have to tediously work through each branch, one step at a time, usually omitting some really important possibility because I don't realize that that particular knight on the far side of the board can affect things on this side where we are both moving pieces.

This sort of "deep" (synthetic) understanding of physics is very much the goal of *this* course (the one in the textbook you are reading, since I use this intro in many textbooks), and to achieve it you must *not* memorize things as if they are random factoids, you must work to abstract the beautiful intertwining of patterns that compress all of those apparently random factoids into things that you can easily remember offhand, that you can easily reconstruct from the pattern even if you forget the details, and that you can search through at a glance. But the process I describe can be applied to learning pretty much anything, as patterns and structure exist in abundance in *all* subjects of interest. There are even sensible rules that govern or describe the anti-pattern of *pure randomness!*

There's one more important thing you can learn from thinking over the digit experiment. *Some* of you reading this very likely didn't do what I asked, you didn't play along with the game. Perhaps it was too much of a bother – you didn't want to waste a *whole minute* learning something by actually *doing* it, just wanted to read the damn chapter and get it over with so you could do, well, whatever the hell else it is you were planning to do today that's more important to you than physics or learning in other courses.

If you're one of these people, you probably don't remember *any* of the digit string at this point from actually seeing it – you never even *tried* to memorize it. A very few of you may actually be so terribly jaded that you don't even remember the little mnemonic *formula* I gave above for the digit string (although frankly, people that are *that* disengaged are probably not about to do things like actually read a textbook in the first place, so possibly not). After all, either way the string is pretty damn meaningless, pattern or not.

Pattern and meaning aren't exactly the same thing. There are all sorts of patterns one can find in random number strings, they just aren't "real" (where we could wax poetic at this point about information entropy and randomness and monkeys typing Shakespeare or seeing fluffy white sheep in the clouds if this were a different course). So why bother wasting brain energy on even the *easy* way to remember this string when doing so is utterly unimportant to you in the grand scheme of all things?

From this we can learn the *second* humble and unsurprising conclusion I want you to draw from this one elementary thought experiment. *Things are easier to learn when you care about learning them!* In fact, they are damn near impossible to learn if you really *don't* care about learning them.

Let's put the two observations together and plot them as a graph, just for fun (and because graphs help one learn for reasons we will explore just a bit in a minute). If you care about learning what you are studying, and the information you are trying to learn makes sense (if only for a moment, perhaps during lecture), the chances of your learning it are quite good. This alone isn't *enough* to guarantee that you'll learn it, but it they are basically both necessary

conditions, and one of them is directly connected to degree of engagement.

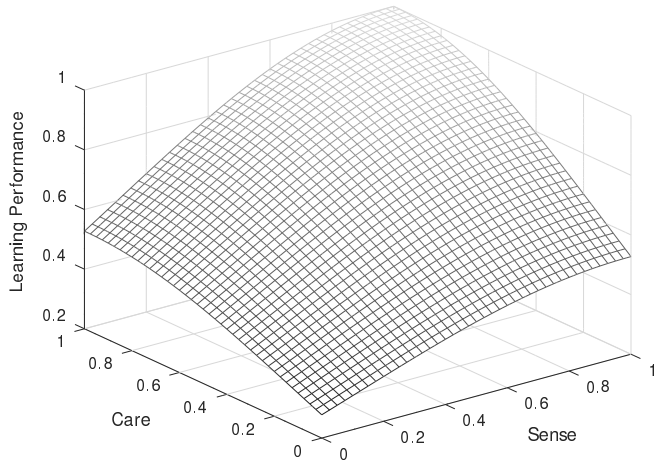


Figure 0.1: Relation between sense, care and learning

On the other hand, if you care but the information you want to learn makes no sense, or if it makes sense but you hate the subject, the instructor, your school, your life and just don't care, your chances of learning it aren't so good, probably a bit better in the first case than in the second as if you care you have a *chance* of finding someone or some way that will help you make sense of whatever it is you wish to learn, where the person who doesn't cares, well, they don't care. Why should they remember it?

If you don't give a rat's ass about the material *and* it makes no sense to you, go home. Leave school. Do something else. You basically have almost no chance of learning the material unless you are gifted with a transcendent intelligence (wasted on a dilettante who lives in a state of perpetual ennui) and are miraculously gifted with the ability learn things effortlessly even when they make no sense to you and you don't really care about them. All the learning tricks and study patterns in the world won't help a student who doesn't try, doesn't care, and for whom the material never makes sense.

If we worked at it, we could probably find other "logistic" controlling parameters to associate with learning – things that increase your probability of learning monotonically as they vary. Some of them are already apparent from the discussion above. Let's list a few more of them with explanations just so that you can see how *easy* it is to sit down to study and try to learn and have "something wrong" that decreases your ability to learn in that particular place and time.

Learning is actual work and involves a fair bit of biological stress, just like working out. Your brain needs *food* – it burns a whopping ***20-30% of your daily calorie intake all by itself*** just living day to day, even more when you are really using it or are somewhat sedentary in your physical habits so your consumption in the form of physical motion is smaller than normal or healthy. Note that your brain runs on pure, energy-rich glucose, so when your blood sugar drops your brain activity drops right along with it. This can happen (paradoxically) because you *just ate a carbohydrate rich meal*. A balanced diet containing foods with a lower glycemic

index[3] ends to be harder to digest and provides a longer period of sustained energy for your brain. A daily multivitamin (and sometimes various antioxidant or metabolic supplements such as alpha lipoic acid) can also help maintain your body's energy release mechanisms at the cellular level.

Blood sugar is typically lowest first thing in the morning, so this is a lousy time to actively study. On the other hand, a good hearty breakfast, eaten at least an hour before plunging in to your studies, is a great idea and is a far better habit to develop for a lifetime than eating no breakfast and instead eating a huge meal right before bed[4]

Learning requires adequate *sleep*. Sure this is tough to manage at college – there are no parents to tell you to go to bed, lots of things to do, and of course you're in *class* during the day and then you study, so late night is when you have fun. Unfortunately, learning is clearly correlated with engagement, activity, and mental alertness, and all of these tend to shut down when you're tired. Furthermore, the formation of *long term memory of any kind* from a day's experiences has been shown in both animal and human studies to *depend* on the brain undergoing at least a few natural sleep cycles of deep sleep alternating with REM (Rapid Eye Movement) sleep, dreaming sleep. Rats taught a maze and then deprived of REM sleep cannot run the maze well the next day; rats that are taught the *same* maze but that get a good night's of rat sleep with plenty of rat dreaming can run the maze well the next day. People conked on the head who remain unconscious for hours and are thereby deprived of normal sleep often have permanent amnesia of the previous day – it never gets turned into long term memory.

Wikipedia: http://www.wikipedia.org/wiki/Sleep Apnea is also a great undiagnosed epidemic (e.g. 24% of all males by late middle age, most of them untreated) and can seriously affect learning. Indeed, if you have any variation of Attention Deficit Disorder (ADD) ***and snore, or have any symptoms of interrupted sleep due to breathing interruption or e.g. restless legs*** you should probably read about the co-morbidity of sleep disorders and ADD[5] and talk to your doctor to make sure that you really *have* ADD and are not suffering from a sleep disorder, as the two can actually result in nearly identical daytime symptoms, including difficulty learning!

This is hardly surprising. Pure common sense and experience tell you that your brain won't work too well if it is hungry and tired or oxygen deprived. Common sense (and yes, experience) will rapidly convince you that learning generally works better if you're not stoned or drunk when you study. Learning works *much* better when you have *time* to learn and haven't put everything off to the last minute. In fact, all of Maslow's hierarchy of needs[6] are important parameters that contribute to the probability of success in learning.

There is one more set of very important variables that strongly affect our ability to learn,

[3]Wikipedia: http://www.wikipedia.org/wiki/glycemic index. t

[4]...which is, alas, my own pattern unless I'm careful, made into a habit back in college. It seemed to work a lot better at age 20 than it does at age 60...

[5]A Clinical Overview of Sleep and Attention-Deficit/Hyperactivity Disorder in Children and Adolescents

[6]Wikipedia: http://www.wikipedia.org/wiki/Maslow's_hierarchy_of_needs. In a nutshell, in order to become *self-actualized* and realize your full potential in activities such as learning you need to have your physiological needs met, you need to be safe, you need to be loved and secure in the world, you need to have good self-esteem and the esteem of others. Only then is it particularly likely that you can become self-actualized and become a great learner and problem solver.

and they are in some ways the least well understood. These are variables that describe you as an *individual*, that describe your *particular* brain and how it works. Pretty much everybody will learn better if they are self-actualized and fully and actively engaged, if the material they are trying to learn is available in a form that makes sense and clearly communicates the implicit patterns that enable efficient information compression and storage, and above all if they *care* about what they are studying and learning, if it has *value* to them.

But everybody is not the same, and the *optimal* learning strategy for one person is not going to be what works well, or even at all, for another. This is one of the things that confounds "simple" empirical research that attempts to find benefit in one teaching/learning methodology over another. Some students *do* improve, even dramatically improve – when this or that teaching/learning methodology is introduced. In others there is no change. Still others actually do worse. In the end, the beneficial effect to a selected subgroup of the students may be lost in the statistical noise of the study and the fact that no attempt is made to identify commonalities among students that succeed or fail.

The point is that finding an optimal teaching and learning strategy is *technically* an *optimization problem on a high dimensional space*. We've discussed *some* of the important dimensions above, isolating a few that appear to have a monotonic effect on the desired outcome in at least some range (relying on common sense to cut off that range or suggest trade-offs – one cannot learn better by simply discussing one idea for weeks at the expense of participating in lecture or discussing many other ideas of equal and coordinated importance; sleeping for twenty hours a day leaves little time for experience to fix into long term memory with all of that sleep). We've omitted one that is crucial, however. That is *your brain!*

Your Brain and Learning

Your brain is more than just a unique instrument. In some sense it is you. You could imagine having your brain removed from your body and being hooked up to machinary that provided it with sight, sound, and touch in such a way that "you" remain[7]. It is difficult to imagine that you still exist in any meaningful sense if your brain is taken out of your body and destroyed while your body is artificially kept alive.

Your brain, however, *is* an instrument. It has internal structure. It uses energy. It does "work". It is, in fact, a biological machine of sublime complexity and subtlety, one of the true wonders of the world! Note that this statement can be made quite independent of whether "you" are your brain per se or a spiritual being who happens to be using it (a debate that need not concern us at this time, however much fun it might be to get into it) – either way the brain itself is quite marvelous.

For all of that, few indeed are the people who bother to learn to actually *use* their brain effectively *as* an instrument. It just works, after all, whether or not we do this. Which is fine. If you want to get the most mileage out of it, however, it helps to read the manual.

So here's at least *one* user manual for your brain. It is by no means complete or authoritative, but it should be enough to get you started, to help you discover that you are actually a lot

[7] Imagine very easily if you've ever seen *The Matrix* movie trilogy...

smarter than you think, or that you've been in the past, once you realize that you can *change* the way you think and learn and experience life and gradually *improve* it.

In the spirit of the learning methodology that we eventually hope to adopt, let's simply itemize in no particular order the various features of the brain[8] that bear on the process of learning. Bear in mind that such a minimal presentation is more of a *metaphor* than anything else because simple (and extremely common) generalizations such as "creativity is a right-brain function" are not strictly true as the brain is far more complex than that.

- The brain is *bicameral*: it has two *cerebral hemispheres*[9] , right and left, with brain functions *asymmetrically* split up between them.
- The brain's hemispheres are connected by a networked membrane called the *corpus callosum* that is how the two halves talk to each other.
- The human brain consists of *layers* with a structure that recapitulates evolutionary phylogeny; that is, the core structures are found in very primitive animals and common to nearly all vertebrate animals, with new layers (apparently) added by evolution on top of this core as the various phyla differentiated, fish, amphibian, reptile, mammal, primate, human. The outermost layer where most actual thinking occurs (in animals that think) is known as the *cerebral cortex*.
- The *cerebral cortex*[10] – especially the outermost layer of *it* called the *neocortex* – is where "higher thought" activities associated with learning and problem solving take place, although the brain is a very complex instrument with functions spread out over many regions.
- An important brain model is a *neural network*[11] . Computer simulated neural networks provide us with insight into how the brain can remember past events and process new information.
- The fundamental operational units of the brain's information processing functionality are called *neurons*[12] . Neurons receive electrochemical signals from other neurons that are transmitted through long fibers called *axons*[13] *Neurotransmitters*[14] are the actual chemicals responsible for the triggered functioning of neurons and hence the neural network in the cortex that spans the halves of the brain.
- Parts of the cortex are devoted to the senses. These parts often contain a *map* of sorts of the world as seen by the associated sense mechanism. For example, there exists a topographic map in the brain that roughly corresponds to points in the retina, which in turn are stimulated by an image of the outside world that is projected onto the retina by your eye's lens in a way we will learn about later in this course! There is thus a *representation of your visual field* laid out inside your brain!

[8]Wikipedia: http://www.wikipedia.org/wiki/brain.
[9]Wikipedia: http://www.wikipedia.org/wiki/cerebral_hemisphere.
[10]Wikipedia: http://www.wikipedia.org/wiki/Cerebral_cortex.
[11]Wikipedia: http://www.wikipedia.org/wiki/Neural_network.
[12]Wikipedia: http://www.wikipedia.org/wiki/Neurons.
[13]Wikipedia: http://www.wikipedia.org/wiki/axon. .
[14]Wikipedia: http://www.wikipedia.org/wiki/neurotransmitters.

- Similar maps exist for the other senses, although sensations from the right side of your body are generally processed in a laterally inverted way by the *opposite* hemisphere of the brain. What your right eye sees, what your right hand touches, is ultimately transmitted to a sensory area in your left brain hemisphere and vice versa, and volitional muscle control flows from these brain halves the other way.

- Neurotransmitters require biological resources to produce and consume bioenergy (provided as glucose) in their operation. You can *exhaust* the resources, and *saturate* the receptors for the various neurotransmitters on the neurons by overstimulation.

- You can also block neurotransmitters by chemical means, put neurotransmitter analogues into your system, and alter the chemical trigger potentials of your neurons by taking various drugs, poisons, or hormones. The *biochemistry of your brain* is extremely important to its function, and (unfortunately) is not infrequently a bit "out of whack" for many individuals, resulting in e.g. attention deficit or mood disorders that can greatly affect one's ability to easily learn while leaving one otherwise highly functional.

- Intelligence[15] , learning ability, and problem solving capabilities are not fixed; they can vary (often improving) over your whole lifetime! Your brain is highly *plastic* and can sometimes even reprogram itself to full functionality when it is e.g. damaged by a stroke or accident. On the other hand neither is it infinitely plastic – any given brain has a range of accessible capabilities and can be improved only to a certain point. However, for people of supposedly "normal" intelligence and above, it is by no means clear what that point is! Note well that *intelligence is an extremely controversial subject* and you should not take things like your own measured "IQ" too seriously.

- Intelligence is not even fixed within a population over time. A phenomenon known as "the Flynn effect"[16] (after its discoverer) suggests that IQ tests have increased almost six points a decade, on average, over a timescale of tens of years, with most of the increases coming from the lower half of the distribution of intelligence. This is an active area of research (as one might well imagine) and some of that research has demonstrated fairly conclusively that individual intelligences can be improved by five to ten points (a significant amount) by environmentally correlated factors such as nutrition, education, complexity of environment.

- The best time for the brain to learn is right before sleep. The process of sleep appears to "fix" long term memories in the brain and things one studies right before going to bed are retained much better than things studied first thing in the morning. Note that this conflicts directly with the party/entertainment schedule of many students, who tend to study early in the evening and then amuse themselves until bedtime. It works much better the other way around.

- Sensory memory[17] corresponds to the roughly 0.5 second (for most people) that a sensory impression remains in the brain's "active sensory register", the sensory cortex. It can typically hold less than 12 "objects" that can be retrieved. It quickly decays and

[15]Wikipedia: http://www.wikipedia.org/wiki/intelligence.
[16]Wikipedia: http://www.wikipedia.org/wiki/flynn_effect.
[17]Wikipedia: http://www.wikipedia.org/wiki/memory. Several items in a row are connected to this page.

cannot be improved by rehearsal, although there is some evidence that its object capacity can be improved over a longer term by practice.

- Short term memory is where *some* of the information that comes into sensory memory is transferred. Just which information is transferred depends on where one's "attention" is, and the mechanics of the attention process are not well understood and are an area of active research. Attention acts like a filtering process, as there is a *wealth* of parallel information in our sensory memory at any given instant in time but the thread of our awareness and experience of time is serial. We tend to "pay attention" to one thing at a time. Short term memory lasts from a few seconds to as long as a minute without rehearsal, and for nearly all people it holds $4-5$ objects[18]. However, its capacity can be increased by a process called "chunking" that is basically the information compression mechanism demonstrated in the earlier example with numbers – grouping of the data to be recalled into "objects" that permit a larger set to still fit in short term memory.

- Studies of chunking show that the ideal size for data chunking is three. That is, if you try to remember the string of letters:

  FBINSACIAIBMATTMSN

  with the usual three second look you'll almost certainly find it impossible. If, however, I insert the following spaces:

  FBI NSA CIA IBM ATT MSN

  It is suddenly much easier to get at least the first four. If I parenthesize:

  (FBI NSA CIA) (IBM ATT MSN)

  so that you can recognize the first three are all government agencies in the general category of "intelligence and law enforcement" and the last three are all market symbols for information technology mega-corporations, you can once again recall the information a day later with only the most cursory of rehearsals. You've taken eighteen "random" objects that were meaningless and could hence be recalled only through the most arduous of rehearsal processes, converted them to six "chunks" of three that can be easily tagged by the brain's existing long term memory (note that you are *not learning* the string FBI, you are building an *association* to the already existing memory of what the string FBI *means*, which is *much easier* for the brain to do), and chunking the chunks into *two* objects.

  Eighteen objects without meaning – difficult indeed! Those *same* eighteen objects *with* meaning – umm, looks pretty easy, doesn't it...

  Short term memory is still that – short term. It typically decays on a time scale that ranges from minutes for nearly everything to order of a day for a few things unless the information can be transferred to *long* term memory. Long term memory is the big payoff – *learning* is associated with formation of long term memory.

- Now we get to the really good stuff. Long term is memory that you form that lasts a long time in human terms. A "long time" can be days, weeks, months, years, or a lifetime. Long term memory is encoded *completely differently* from short term or sensory/immediate

[18] From this you can see why I used ten digits, gave you only a few seconds to look, and blocked rehearsal in our earlier exercise.

memory – it appears to be encoded *semantically*[19] , that is to say, *associatively* in terms of its *meaning*. There is considerable evidence for this, and it is one reason we focus so much on the importance of meaning in the previous sections.

To miraculously transform things we try to remember from "difficult" to learn random factoids that have to be brute-force stuffed into disconnected semantic storage units created as it were one at a time for the task at hand into "easy" to learn factoids, all we have to do is *discover* meaning associations with things we already know, or *create* a strong memory of the global meaning or *conceptualization* of a subject that serves as an associative home for all those little factoids.

A characteristic of this as a successful process is that when one works systematically to learn by means of the latter process, learning gets *easier* as time goes on. Every factoid you add to the semantic structure of the global conceptualization strengthens it, and makes it even easier to add new factoids. In fact, the mind's extraordinary rational capacity permits it to interpolate and extrapolate, to *fill in* parts of the structure on its own *without effort* and in many cases without even being exposed to the information that needs to be "learned"!

- One area where this extrapolation is particularly evident and powerful is in *mathematics*. Any time we can learn, or discover from experience a *formula* for some phenomenon, a mathematical *pattern*, we don't have to actually see something to be able to "remember" it. Once again, it is easy to find examples. If I give you data from sales figures over a year such as January = $1000, October = $10,000, December = $12,000, March=$3000, May = $5000, February = $2000, September = $9000, June = $6000, November = $11,000, July = $7000, August = $8000, April = $4000, at first glance they look quite difficult to remember. If you organize them temporally by month and look at them for a moment, you recognize that sales increased *linearly* by month, starting at $1000 in January, and suddenly you can reduce the whole series to a simple mental formula (straight line) and a couple pieces of initial data (slope and starting point). One amazing thing about this is that if I asked you to "remember" something that you *have not seen*, such as sales in February in the *next* year, you could make a very plausible guess that they will be $14,000!

  Note that this isn't a memory, it is a guess. Guessing is what the mind is designed to do, as it is part of the process by which it "predicts the future" even in the most mundane of ways. When I put ten dollars in my pocket and reach in my pocket for it later, I'm basically guessing, on the basis of my memory and experience, that I'll find ten dollars there. Maybe my guess is wrong – my pocket could have been picked[20], maybe it fell out through a hole. My *concept* of object permanence plus my *memory* of an initial state permit me to make a *predictive guess* about the Universe!

  This is, in fact, physics! This is what physics is all about – coming up with a set of rules (like conservation of matter) that encode observations of object permanence, more rules (equations of motion) that dictate how objects move around, and allow me to conclude that "I put a ten dollar bill, at rest, into my pocket, and objects at rest remain at rest. The matter the bill is made of cannot be created or destroyed and is bound together in

[19] Wikipedia: http://www.wikipedia.org/wiki/semantics.

[20] With three sons constantly looking for funds to attend movies and the like, it isn't as unlikely as you might think!

a way that is unlikely to come apart over a period of days. Therefore the ten dollar bill is still there!" Nearly anything that you do or that happens in your everyday life can be formulated as a predictive physics problem.

- The *hippocampus*[21] appears to be partly responsible for both forming spatial maps or visualizations of your environment and also for forming the *cognitive map* that organizes what you know and transforms short term memory into long term memory, and it appears to do its job (as noted above) *in your sleep*. Sleep deprivation *prevents the formation of long term memory*. Being rendered unconscious for a long period often produces *short term amnesia* as the brain loses short term memory before it gets put into long term memory. The hippocampus shows evidence of plasticity – taxi drivers who have to learn to navigate large cities actually have larger than normal hippocampi, with a size proportional to the length of time they've been driving. This suggests (once again) that it is possible to *deliberately increase the capacity* of your *own* hippocampus through the exercise of its functions, and consequently *increase your ability to store and retrieve information*, which is an important component (although not the only component) of intelligence!

- Memory is improved by *increasing the supply of oxygen to the brain*, which is best accomplished by *exercise*. Unsurprisingly. Indeed, as noted above, having good general health, good nutrition, good oxygenation and perfusion – having all the biomechanism in tip-top running order – is perfectly reasonably linked to being able to perform at your best in anything, mental activity included.

- Finally, the *amygdala*[22] is a brain organ in our *limbic system* (part of our "old", reptile brain). The amygdala is an important part of our *emotional* system. It is associated with primitive survival responses, with sexual response, and appears to play a *key role* in modulating (filtering) the process of turning short term memory into long term memory. Basically, any sort term memory associated with a powerful emotion is much more likely to make it into long term memory.

  There are clear evolutionary advantages to this. If you narrowly escape being killed by a saber-toothed tiger at a particular pool in the forest, and then forget that this happened by the next day and return again to drink there, chances are decent that the saber-tooth is still there and you'll get eaten. On the other hand, if you come upon a particular fruit tree in that same forest and get a free meal of high quality food and forget about the tree a day later, you might starve.

  We see that both negative and positive emotional experiences are strongly correlated with learning! *Powerful* experiences, especially, are correlated with learning. This translates into learning strategies in two ways, one for the instructor and one for the student. For the instructor, there are two general strategies open to helping students learn. One is to create an atmosphere of *fear, hatred, disgust, anger* – powerful negative emotions. The other is to create an atmosphere of *love, security, humor, joy* – powerful positive emotions. In between there is a great wasteland of bo-ring, bo-ring, bo-ring where students plod along, struggling to form memories because there is nothing "exciting" about

[21] Wikipedia: http://www.wikipedia.org/wiki/hippocampus.
[22] Wikipedia: http://www.wikipedia.org/wiki/amygdala.

the course in either a positive or negative way and so their amygdala degrades the memory formation process in favor of other more "interesting" experiences.

Now, in my opinion, negative experiences in the classroom do indeed promote the formation of long term memories, but they aren't the memories the instructor intended. The student is likely to remember, and loath, the instructor for the rest of their life but is *not* more likely to remember the material except sporadically in association with particularly traumatic episodes. They may well be *less* likely, as we naturally avoid negative experiences and will study less and work less hard on things we can't stand doing.

For the instructor, then, positive is the way to go. Creating a warm, nurturing classroom environment, ensuring that the students know that you *care* about their learning and about them as individuals helps to promote learning. Making your lectures and teaching processes *fun* – and *funny* – helps as well. Many successful lecturers make a powerful *positive* impression on the students, creating an atmosphere of amazement or surprise. A classroom experience should really be a *joy* in order to optimize learning in so many ways.

For the student, be aware that *your attitude matters!* As noted in previous sections, *caring* is an essential component of successful learning because you have to attach *value* to the process in order to get your amygdala to do its job. However, you can do *much more*. You can see how *many* aspects of learning can be enhanced through the simple expedient of making it a positive experience! Working in groups is *fun*, and you learn more when you're having fun (or quavering in abject fear, or in an interesting mix of the two). Attending an interesting lecture is fun, and you'll retain more than average. Participation is fun, especially if you are "rewarded" in some way that makes a moment or two special to you, and you'll remember more of what goes on.

From all of these little factoids (presented in a way that I'm hoping helps you to build at least the beginnings of a working conceptual model of your own brain) I'm hoping that you are coming to realize that *all of this is at least partially under your control!* Even if your instructor is scary or boring, the material at first glance seems dry and meaningless, and so on – all the negative-neutral things that make learning difficult, *you* can decide to make it fun and exciting, *you* can ferret out the meaning, *you* can adopt study strategies that focus on the formation of cognitive maps and organizing structures *first* and *then* on applications, rehearsal, factoids, and so on, *you* can learn to study right before bed, get enough sleep, become aware of your brain's learning biorhythms.

Finally, you can learn to *increase your functional learning capabilities* by a *significant* amount. Solving puzzles, playing mental games, doing crossword puzzles or sudoku, working homework problems, writing papers, arguing and discussing, just plain *thinking* about difficult subjects and problems even when you don't *have* to all increase your active intelligence in initially small but cumulative ways. You too can increase the size of your hippocampus, learn to engage your amygdala by *choosing* in a self-actualized way what you value and learning to discipline your emotions accordingly, and create more conceptual maps within your brain that can be shared as components across the various things you wish to learn. The more you know about *anything*, the easier it is to learn *everything* and vice versa! This is the pure biology underlying the value of the liberal arts education.

Use your whole brain, exercise it often, don't think that you "just" need math and not spatial

relations, visualization, verbal skills, a knowledge of history, a memory of performing experiments with your hands or mind or both – you need it all! Remember, just as is the case with physical exercise (which you should get plenty of), *mental* exercise gradually makes you mentally stronger, so that you can eventually do easily things that at first appear insurmountably difficult. You can learn to learn *three to ten times as fast* as you did in high school, to have more fun while doing it, and to gain tremendous reasoning capabilities along the way just by *trying* to learn to learn more efficiently instead of continuing to use learning strategies that worked (possibly indifferently) back in elementary and high school.

The next section, at long last, will make a very specific set of suggestions for *one* very good way to study physics (or nearly anything else) in a way that maximally takes advantage of your own volitional biology to make learning as efficient and pleasant as it is possible to be.

How to Do Your Homework Effectively

By now in your academic career (and given the information above) it should be very apparent just where homework exists in the grand scheme of (learning) things. Ideally, you attend a class where a warm and attentive professor clearly explains some abstruse concept and a whole raft of facts in some moderately interactive way that encourages engagement and "being earnest". Alas, there are *too many* facts to fit in short term/immediate memory and *too little time* to move most of them through into long term/working memory before finishing with one and moving on to the next one. The material may appear to be boring and random so that it is difficult to pay full attention to the *patterns* being communicated and remain emotionally enthusiastic all the while to help the process along. As a consequence, by the end of lecture you've already *forgotten* many if not most of the facts, but if you were paying attention, asked questions as needed, and really cared about learning the material you *would* remember a handful of the most important ones, the ones that made your brief understanding of the material hang (for a brief shining moment) together.

This conceptual overview, however initially tenuous, is the skeleton you will eventually clothe with facts and experiences to transform it into an entire system of associative memory and reasoning where you can work intellectually at a high level with little effort and usually with a great deal of pleasure associated with the very act of thinking. But you aren't there yet.

You now know that you are not terribly likely to retain a lot of what you are shown in lecture without engagement. In order to actually learn it, you must *stop* being a passive recipient of facts. You must *actively* develop your understanding, by means of *discussing* the material and kicking it around with others, by *using* the material in some way, by *teaching* the material to peers as you come to understand it.

To help facilitate this process, associated with lecture your professor almost certainly gave you an *assignment*. Amazingly enough, its purpose is not to torment you or to be the basis of your grade (although it may well do both). It is to give you some concrete stuff to *do* while thinking about the material to be learned, while discussing the material to be learned, while using the material to be learned to accomplish specific goals, while teaching some of what you figure out to others who are sharing this whole experience while being taught by them in turn. The assignment is *much more important* than lecture, as it is entirely participatory, where real

learning is *far more likely to occur*. You could, once you learn the trick of it, blow off lecture and do fine in a course in all other respects. If you fail to do the assignments *with your entire spirit engaged*, you are doomed.

In other words, to learn you must *do your homework*, ideally at least partly in a *group* setting. The only question is: *how* should you do it to both finish learning all that stuff you sort-of-got in lecture and to re-attain the moment(s) of clarity that you then experienced, until eventually it becomes a permanent characteristic of your awareness and you *know* and *fully understand* it all on your own?

There are two general steps that need to be *iterated* to finish learning anything at all. They are a lot of work. In fact, they are far *more* work than (passively) attending lecture, and are *more important* than attending lecture. You can learn the material with these steps without *ever* attending lecture, as long as you have access to what you need to learn in some media or human form. You in all probability will *never* learn it, lecture or not, without making a few passes through these steps. They are:

a) Review the whole (typically lecture, textbooks and/or notes, the Internet, videos...)

b) Work on the parts (***do homework***, and otherwise try to ***use*** what you are learning for something)

(iterate until you thoroughly understand whatever it is you are trying to learn).

Let's examine these steps.

The first is pretty obvious. You generally don't "get it" (where "it" is almost anything nontrivial you are trying to learn) from one lecture, from reading one textbook one time. There is too much material, and it doesn't initially make sense to you. If you are *lucky* and well prepared and blessed with a good instructor, perhaps you grasp *some* of it for a *moment* (and if your instructor is poor or you are particularly poorly prepared you may not manage even that) but what you do momentarily understand is fading, flitting further and further away with every moment that passes. You need to review the entire topic, as a whole, as well as all its parts. A set of good summary notes might contain all the relative factoids, but there are *relations* between those factoids – a temporal sequencing, mathematical derivations connecting them to other things you know, a topical association with other things that you know. They tell a *story*, or part of a story, and you need to know that story in *broad* terms, not try to memorize it word for word.

Reviewing the material should be done in layers, skimming the textbook and your notes, creating a *new* set of notes out of the text in combination with your lecture notes, maybe reading in more detail to understand some particular point that puzzles you, reworking a few of the examples presented. Lots of increasingly deep passes through it (starting with the merest skim-reading or reading a summary of the whole thing) are *much* better than trying to work through the whole text one line at a time and not moving on until you understand it. Many things you might want to understand will only come clear from things you are exposed to *later*, as it is not the case that all knowledge is ordinal, hierarchical, and derivatory.

You especially do *not* have to work on *memorizing* the content. In fact, it is *not* desireable to try to memorize content at this point – you want the big picture *first* so that facts have a place to live in your brain. If you build them a house, they'll move right in without a fuss, where

if you try to grasp them one at a time with no place to put them, they'll (metaphorically) slip away again as fast as you try to take up the next one. Let's understand this a bit.

As we've seen, your brain is fabulously efficient at storing information in a *compressed associative* form. It also tends to remember things that are *important* – whatever that means – and forget things that aren't important to make room for more important stuff, as your brain structures work together in understandable ways on the process. Building the cognitive map, the "house", is what it's all about. But as it turns out, building this house *takes time*.

This is the goal of your iterated review process. At first you are memorizing things the hard way, trying to connect what you learn to very simple hierarchical concepts such as this step comes before that step. As you do this over and over again, though, you find that absorbing new information takes you less and less time, and you remember it much more easily and for a longer time without additional rehearsal. Sometimes your brain even *outruns* the learning process and "discovers" a missing part of the structure before you even read about it! By reviewing the whole, well-organized structure over and over again, you gradually build a greatly compressed representation of it in your brain and tremendously reduce the amount of work required to flesh out that structure with increasing levels of detail *and remember them and be able to work with them* for a long, long time.

Now let's understand the second part of doing homework – working problems. As you can probably guess on your own at this point, there are good ways and bad ways to do homework problems. The worst way to do homework (aside from not doing it at all, which is *far too common* a practice and a *bad idea* if you have any intention of learning the material) is to do it all in one sitting, right before it is due, and to never again look at it.

Doing your homework in a single sitting, working on it just one time *fails to repeat and rehearse the material* (essential for turning short term memory into long term in nearly all cases). It *exhausts the neurons in your brain* (quite literally – there is metabolic energy consumed in thinking) as one often ends up working on a problem far too long in one sitting just to get done. It *fails to incrementally build up* in your brain's long term memory the *structures* upon which the more complex solutions are based, so you have to constantly go back to the book to get them into short term memory long enough to get through a problem. Even this simple bit of repetition does *initiate* a learning process. Unfortunately, by not repeating the steps associated with the solution to this kind of problem after this one sitting they soon fade, often without a discernable trace in long term memory.

Just as was the case in our experiment with memorizing the number above, the problems almost invariably are *not* going to be a matter of random noise. They have certain key facts and ideas that are the basis of their solution, and those ideas are used over and over again. There is plenty of pattern and meaning there for your brain to exploit in information compression, and it may well be *very cool stuff to know* and hence *important* to you once learned, but it takes time and repetition and a certain amount of meditation for the "gestalt" of it to spring into your awareness and burn itself into your conceptual memory as "high order understanding".

You have to *give* it this time, and perform the repetitions, while maintaining an optimistic, philosophical attitude towards the process. You have to do your best to have *fun* with it. You don't get strong by lifting light weights a single time. You get strong lifting weights repeatedly, starting with light weights to be sure, but then working up to the *heaviest weights you can*

manage. When you *do* build up to where you're lifting hundreds of pounds, the fifty pounds you started with seems light as a feather to you.

As with the body, so with the brain. Repeat broad strokes for the big picture with increasingly deep and "heavy" excursions into the material to explore it in detail as the overall picture emerges. Intersperse this with sessions where you *work on problems* and try to *use* the material you've figured out so far. Be sure to *discuss* it and *teach it to others* as you go as much as possible, as articulating what you've figured out to others both uses a different part of your brain than taking it in (and hence solidifies the memory) and it helps you articulate the ideas to *yourself!* This process will help you learn more, better, faster than you ever have before, and to have fun doing it!

Your brain is more complicated than you think. You are very likely used to *working hard* to try to *make* it figure things out, but you've probably observed that this doesn't work very well. A lot of times you simply *cannot* "figure things out" because your brain doesn't yet know the key things required to do this, or doesn't "see" how those parts you do know fit together. Learning and discovery is not, alas, "intentional" – it is more like trying to get a bird to light on your hand that flits away the moment you try to grasp it.

People who do really hard crossword puzzles (one form of great brain exercise) have learned the following. After making a pass through the puzzle and filling in all the words they can "get", and maybe making a couple of extra passes through thinking hard about ones they can't get right away, looking for patterns, trying partial guesses, they arrive at an impasse. If they continue working hard on it, they are unlikely to make further progress, no matter how long they stare at it.

On the other hand, if they *put the puzzle down* and *do something else for a while* – especially if the something else is go to bed and sleep – when they come back to the puzzle they often can *immediately see* a dozen or more words that the day before were absolutely invisible to them. Sometimes one of the *long theme answers* (perhaps 25 characters long) where they have no more than *two letters* just "gives up" – they can simply "see" what the answer must be.

Where do these answers come from? The person has not "figured them out", they have "recognized" them. They come all at once, and they don't come about as the result of a logical sequential process.

Often they come from the person's *right brain*[23]. The left brain tries to use logic and simple memory when it works on crosswork puzzles. This is usually good for some words, but for many of the words there are *many possible answers* and without any insight one can't even recall *one* of the possibilities. The clues don't suffice to connect you up to a word. Even as letters get filled in this continues to be the case, not because you don't *know* the word (although in really hard puzzles this can sometimes be the case) but because you don't know how to *recognize* the word "all at once" from a cleverly nonlinear clue and a few letters in this context.

The right brain is (to some extent) responsible for *insight* and *non-linear thinking*. It sees

[23]Note that this description is at least partly metaphor, for while there is some hemispherical specialization of some of these functions, it isn't always sharp. I'm retaining them here (oh you brain specialists who might be reading this) because they are a *valuable* metaphor.

patterns, and *wholes*, not sequential relations between the parts. It isn't intentional – we can't "make" our right brains figure something out, it is often the other way around! Working hard on a problem, then "sleeping on it" (to get that all important hippocampal involvement going) is actually a *great* way to develop "insight" that lets you solve it *without really working terribly hard* after a few tries. It also utilizes more of your brain – left and right brain, sequential reasoning and insight, and if you articulate it, or use it, or make something with your hands, then it exercieses these parts of your brain as well, strengthening the memory and your understanding still more. The learning that is associated with this process, and the problem solving power of the method, is *much greater* than just working on a problem linearly the night before it is due until you hack your way through it using information assembled a part at a time from the book.

The following "Method of Three Passes" is a *specific* strategy that implements many of the tricks discussed above. It is known to be effective for learning by means of doing homework (or in a generalized way, learning anything at all). It is ideal for "problem oriented homework", and will pay off big in learning dividends should you adopt it, especially when supported by a *group oriented recitation* with *strong tutorial support* and *many opportunities for peer discussion and teaching*.

The Method of Three Passes

Pass 1 Three or more nights before recitation (or when the homework is due), make a *fast* pass through all problems. Plan to spend 1-1.5 hours on this pass. With roughly 10-12 problems, this gives you around 6-8 minutes per problem. Spend *no more* than this much time *per problem* and if you can solve them in this much time fine, otherwise move on to the next. Try to do this the last thing before bed at night (seriously) and *then go to sleep*.

Pass 2 After at least one night's sleep, make a *medium speed* pass through all problems. Plan to spend 1-1.5 hours on this pass as well. Some of the problems will already be solved from the first pass or nearly so. *Quickly* review their solution and then move on to concentrate on the still unsolved problems. If you solved 1/4 to 1/3 of the problems in the first pass, you should be able to spend 10 minutes or so per problem in the second pass. Again, do this right before bed if possible and then go immediately to sleep.

Pass 3 After at least one night's sleep, make a *final* pass through all the problems. Begin as before by quickly reviewing all the problems you solved in the previous two passes. Then spend fifteen minutes or more (as needed) to solve the remaining unsolved problems. Leave any "impossible" problems for recitation – there should be no more than three from any given assignment, as a general rule. Go immediately to bed.

This is an *extremely powerful* prescription for deeply learning nearly *anything*. Here is the motivation. Memory is formed by repetition, and this obviously contains a lot of that. Permanent (long term) memory is actually formed in your sleep, and studies have shown that whatever you study right before sleep is most likely to be retained. Physics is actually a "whole brain" subject – it requires a synthesis of both right brain visualization and conceptualization and left brain verbal/analytical processing – both geometry and algebra, if you like, and you'll often find that problems that stumped you the night before just solve themselves "like magic" on the second or third pass if you work hard on them for a short, intense, session and then

sleep on it. This is your right (nonverbal) brain participating as it develops intuition to guide your left brain algebraic engine.

Other suggestions to improve learning include working in a study group for that third pass (the first one or two are best done alone to "prepare" for the third pass). Teaching is one of the best ways to learn, and by working in a group you'll have opportunities to both teach and learn more deeply than you would otherwise as you have to articulate your solutions.

Make the learning *fun* – the *right* brain is the key to forming long term memory and it is the seat of your *emotions*. If you are happy studying and make it a positive experience, you will increase retention, it is that simple. Order pizza, play music, make it a "physics homework party night".

Use your whole brain on the problems – draw lots of pictures and figures (right brain) to go with the algebra (left brain). Listen to quiet music (right brain) while thinking through the sequences of events in the problem (left brain). Build little "demos" of problems where possible – even using your hands in this way helps strengthen memory.

Avoid memorization. You will learn physics far better if you learn to *solve* problems and *understand* the concepts rather than attempt to *memorize* the umpty-zillion formulas, factoids, and specific problems or examples covered at one time or another in the class. That isn't to say that you shouldn't learn the important formulas, Laws of Nature, and all of that – it's just that the learning should generally *not* consist of putting them on a big sheet of paper all jumbled together and then trying to memorize them as abstract collections of symbols out of context.

Be sure to review the problems *one last time* when you get your graded homework back. Learn from your mistakes or you will, as they say, be doomed to repeat them.

If you follow this prescription, you will have seen *every assigned homework problem* a minimum of five or six times – three original passes, recitation itself, a final write up pass after recitation, and a review pass when you get it back. At least three of these should occur after you have solved *all* of the problems correctly, since recitation is devoted to ensuring this. When the time comes to study for exams, it should really be (for once) a *review* process, not a cram. Every problem will be like an old friend, and a very brief review will form a *seventh* pass or *eighth* pass through the assigned homework.

With this methodology (enhanced as required by the physics resource rooms, tutors, and help from your instructors) there is no reason for you do poorly in the course and every reason to expect that you will do well, perhaps very well indeed! And you'll still be spending only the 3 to 6 hours per week on homework that is expected of you in any college course of this level of difficulty!

This ends our discussion of course preliminaries (for nearly *any* serious course you might take, not just physics courses) and it is time to get on with actual material for *this* course. The first "chapter" of course material is still placed in the Preliminaries section, for good reason. The topic of this chapter is math, not physics. Math is a very important component of learning physics, and while you can, and should, significantly improve your math skills while taking physics and actually *using* them for something instead of drilling in them, it is still quite important that you enter the course with a certain amount of competence. The following "Week 0" chapter, then, is intended to help guide you in both finding resources for review and math support for the course and to include a self-contained description of a math topic only lightly

passed over to now – integrating over two and three dimensional distributions in cartesian, cylindrical and spherical polar coordinates, the “big three” as far as electricity and magnetism (and a lot of other physics!) are concerned.

Week 0: Math Needed for Introductory E&M (and Optics)

Physics in general, as was noted in the preface, requires a solid knowledge of all mathematics through calculus. Newton *invented* calculus *so* that he could invent physics. Yes, there are "algebraic physics" textbooks out there, but I don't think much of them, just as I don't think much of memorization of physics formulae (as opposed to learning them at a much deeper level than memorization). To use a (possibly poor) metaphor – if mathematics is the language of physics, algebraic physics is the equivalent of memorizing a book of phrases in a foreign language for a traveller who expects to spend only a couple of weeks using it and then never use that language again. Of course *some* stuff will stick – a few phrases here or here – and you may even remember in detail what those words mean. You won't be able to engage in an actual *conversation* in the language, however, even if you *can* ask where the train station is, or order a beer and say please and thank you while doing so because you remember those particular very important phrases.

To engage in a meaningful "conversation" in physics – in order to master both the concepts and apply them in solving real problems to the point where you could read a scientific or medical paper involving physics and not be utterly lost – you really do have to be competent in a wide range of mathematics so that you can understand the physical principles *expressed* in the language of mathematics, and *understand* the mathematical derivations and relations that connect them and tell you what they *mean*. Meaning in physics is often very difficult to convey in words, and requires a lot of them to make things precise, where a mathematical formula that is linked to many *other* formulas in a meaningful way condenses it all into a single short statement.

Here's some of what you need to get started in introductory *mechanics* (the first semester course):

a) A bit of **Number Theory** – what integers, rational numbers, irrational numbers, real numbers, and complex numbers *are*, an understanding of their arithmetical operations, and how to represent them and these operations with *symbols* and manipulate them using...

b) **Algebra**. Physics with calculus is still mostly algebra, it just includes the algebraic operations not only of symbols representing physical quantities that can have numerical values but those of...

c) **Geometry**. This includes traditional plane geometry – knowing various true facts about opposite interior angles and the number of degrees/radians inside a triangle to *analytic*

geometry and trigonometry – coordinate systems and sine, cosine and tangents. This is used mostly to understand...

d) **Vectors**. Many of the quantities of greatest interest in physics are vectors (or if you prefer, tensors with a range greater than zero). Finally, you use *all* of this in physics problems ultimately based on...

e) **Simple Differential and Integral Calculus**. You don't have to know a *lot* of calculus to take the course. Five derivatives and their corresponding integrals will do (seven if you want to master a few "advanced" topics). In addition to the derivatives themselves, you of course need to know the chain rule and its linked cousin, u-substitution, and the derivative of a product and *its* linked cousin, integration by parts. Finally, you need to be at least a little bit familiar with the Taylor series expansion of a smooth function and the binomial expansion.

For second semester E& M and Optics, we still need all of this but we also need to add a few items specific to E& M. The most important of these are:

f) **Coordinate Systems**. One can "get by" in mechanics with cartesian coordinates plus the plane polar coordinate for 2D problems. Only a few problems emphasize integration over more than one dimension, cylindrical symmetry, or spherical symmetry, and that is mostly to prepare you for E& M. Here we will really *need* to understand cylindrical coordinate frames and spherical polar coordinate frames because many problems have precisely those symmetries. Spherical coordinates, especially, are the "natural" coordinates for much of E& M (and for gravitation in the mechanics semester as well, although this isn't really emphasized beyond radial dependencies). This leads us to...

g) **Integration over Two and Three Dimensions**. This *sounds* like multivariate calculus, and of course it is, but the textbook itself only presents problems where the integration effectively *separates* so integrating over (say) a three dimensional charge density distribution is equivalent to doing three *independent* one-dimensional integrals. In this way one can encounter and solve much more "interesting" and realistic problems but still use nothing but one dimensional integrals covered on the One Sheet Math Review pages available on the internet[24].

Let's go over the crucial new math content (only) and take a quick look at cartesian, cylindrical, and spherical polar coordinate systems and see how to set up *separable* integrals like the ones you will encounter in this course (with two examples that are *not* separable, one easily enough solvable, the other very difficult indeed to solve, so you can see what the fuss is about and why – eventually – you might have to take multivariate calculus to go on in physics as a major or tackle simularly multivariate problems in other disciplines).

[24] http://www.phy.duke.edu/~rgb/Class/one-sheet-math-review.php As noted above, students are *strongly encouraged* to download and print out a copy of these review sheets and stick them in their notes for quick reference, although their general content will be covered in this chapter.

0.1: Coordinate Frames

Before we dig in to the electric field produced by continuous charge distributions, we need to review the three most useful coordinate systems, or frames, that we will use to evaluate those fields. Up to now we have mostly used ordinary ***cartesian coordinates*** (x, y) or (x, y, z), sometimes used plane polar coordinates (r, θ), and in the first semester course have even used the radial coordinate r of ***spherical polar coordinates*** (r, θ, ϕ) or ***cylindrical coordinates*** (r, θ, z), but only in contexts where the other two coordinates don't really matter.

This will no longer do. We will be doing integrals and expressing relations in all three of these coordinate frames! This means that we have to *both* know what the coordinates are, *and* be able to go back and forth between the coordinate systems and do *calculus* – specifically integrating over continuous charge distributions – using them! We will stop far short of invoking the full power and range of multivariate calculus – the integrals we will ultimately do will still be *one-dimensional* integrals over *one coordinate at a time* with *each coordinate integral fully independent of the others* (which is why you don't need to have taken multivariate calculus to take this course). We will therefore develop concepts like "length elements", "area elements", and "volume elements" informally at a level sufficient to support their intelligent use in this course, without the use of partial derivatives or jacobians.

Let's start with the easiest and most familiar of the three: cartesian cooordinates!

0.1.1: Cartesian Coordinates

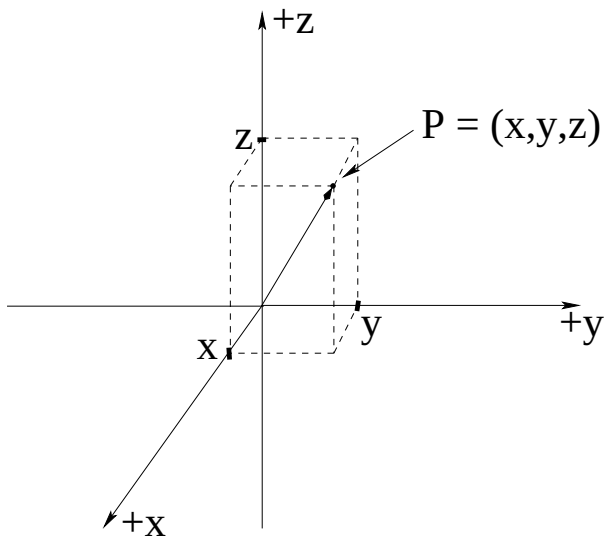


Figure 0.1: The classic rectilinear, orthogonal***cartesian coordinate system***, with a point $P = (x, y, z)$ illustrated.

In figure 0.1 the well-known cartesian coordinate frame is used to represent the position in 3-dimensional space of a single point, $P = (x, y, z)$. This point P can also be thought of as the tip of *position vector* from the origin of the particular coordinate frame to the point P:

$$\vec{\boldsymbol{r}} = x\hat{\boldsymbol{x}} + y\hat{\boldsymbol{y}} + z\hat{\boldsymbol{z}} \quad (\text{ or }) \quad \vec{\boldsymbol{r}} = x\hat{\boldsymbol{\imath}} + y\hat{\boldsymbol{\jmath}} + z\hat{\boldsymbol{k}} \tag{0.1}$$

(depending on how you learned to represent your unit vectors in the x, y, and z directions). Since we use i, j, and k for many things in physics and engineering – the imaginary unit, currents, constants such as the spring constant or electric and magnetic field constants – I prefer to use $\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{z}}$ as the unit vectors in the three orthogonal directions.

If you are taking this course, you have learned to do one dimensional integrals – integrals along lines of functions expressed in linear coordinates, and if you took the preceding mechanics course you were required to integrate over mass distributions in one or two dimensions to evaluate centers of mass or moments of inertia.

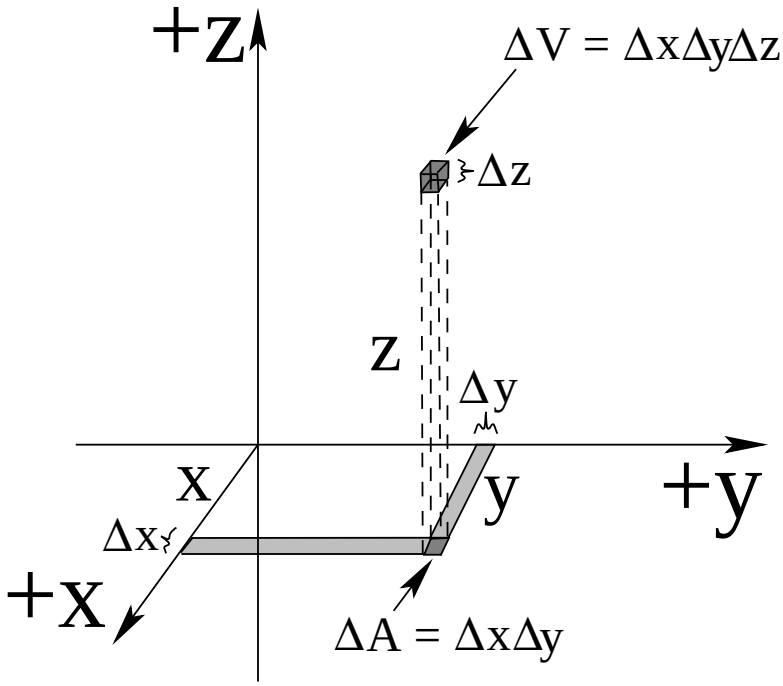


Figure 0.2: The differential elements appropriate to cartesian coordinates, expressed as Δ's instead of d's. Physicists often draw out these finite sized (but small) line, area, or volume segments to help visualization *before* mentally taking the differential limit and converting the Δ's all to d's.

In figure 0.2 the differential elements in cartesian coordinates are displayed as finite-sized "chunks" of each coordinate. In your mind you should look at e.g. the Δx as a finite chunk from x to $x + \Delta x$ on the x axis, and mentally shrink it (at the end) to "differential" sized – dx. In this way we can *see* the length, area, and volume elements as actual (small) lengths, areas, and volumes where of course the 'size' of an infinitesimal is, well, infinitesimal, a point. It's hard to visualize points squared or points cubed, be easy to visualize smaller and smaller square areas or cubic volumes.

From this figure, you can immediately see:

Three length elements: $\Delta x \Rightarrow dx$, $\Delta y \Rightarrow dy$, $\Delta z \Rightarrow dz$. Using these three differential lengths, we can integrate functions of x (but not y or z), y (but not x or z), and z (but not x or y).

One (of three) area elements: $\Delta A = \Delta x \Delta y \Rightarrow dA = dx\,dy$. Note that there are area elements for *each* of the orthogonal planes: $dA = dx\,dz$ and $dA = dy\,dz$ allow surfaces

parallel to the x-z plane or y-z plane to be integrated over as well, again assuming that we have no dependence on the omitted coordinate in each case.

One volume element: $\Delta V = \Delta x \Delta y \Delta z \Rightarrow dV = dx\, dy\, dz$. There is only one volume element in a three dimensional cartesian frame.

A convenient way to visualize the volume element is to go to the point $P = x, y, z$, and "push out" from x to $x + \Delta x$ to make a short line segment there in the x direction. Then "push out" this line segment in the y direction to $y + \Delta y$, generating a little square with area (as illustrated) of $\Delta A = \Delta x \Delta y$. Finally "push up" this square in the z direction from z to $z + \Delta z$ to make a tiny cube with volume $\Delta V = \Delta x \Delta y \Delta z$ with one corner at (x, y, z) and the opposite corner at $(x + \Delta x, y + \Delta y, z + \Delta z)$.

Example 0.1.1: Integrating a Function Along a Line in Cartesian Coordinates

This is pretty obvious if you've taken calculus *at all*, so I'll just write down a few trivial examples. A power law integral:

$$\int_{x_1}^{x_2} ax^2\, dx = \left.\frac{ax^3}{3}\right|_{x_1}^{x_2} = \frac{a}{3}\left(x_2^3 - x_1^3\right) \tag{0.2}$$

An exponential (using u-substitution with $u = x'/x_0$, $du = dx'/x_0$):

$$\int_0^x Ae^{-x'/x_0}\, dx' = -x_0 A \int_0^{-x/x_0} e^u du = x_0 A\left(1 - e^{-x/x_0}\right) \tag{0.3}$$

A harmonic function (again using u-substitution with $u = kx$, $du = k\, dx$):

$$\int_0^\lambda \sin(kx + \phi)\, dx = -\left.\frac{1}{k}\cos(kx + \phi)\right|_0^\lambda = \frac{1}{k}\left(1 - \cos(kx + \phi)\right) \tag{0.4}$$

Example 0.1.2: Integrating a Function over an Area in Cartesian Coordinates

In order for us to proceed, we have to assume that the x-coordinate itself appearing in a function does not depend on the y-coordinate. Usually this is *true* in the expression of the function itself in x and y coordinates, but it may well *not be true* at the *boundary of integration*.

For example, it is very easy to find the area inside a rectangle that runs from $(0, 0)$ to $(a, 0)$, from $(a, 0)$ to (a, b), from (a, b) to $(0, b)$, and finally from $(0, b)$ to $(0, 0)$:

$$A_{\text{rectangle}} = \int_{\text{rectangle}} dx\, dy = \int_0^a dx \int_0^b dy = ab \tag{0.5}$$

It is a bit more difficult, but still pretty easy, to integrate over a right triangle in cartesian coordinates provided one lines the sides up with the coordinate system. For a triangle that runs from $(0, 0)$ to $(a, 0)$, then to (a, b), then back to $(0, 0)$ one can determine its area using the function $y(x) = \frac{b}{a}x$ as its upper limit of integration in y. Then:

$$A_{\text{triangle}} = \int_{\text{triangle}} dx\, dy = \int_0^a dx \int_0^{y(x)} dy = \int_0^a y(x) dx = \int_0^a \frac{b}{a} x\, dx = \frac{1}{2}\frac{b}{a}a^2 = \frac{1}{2}ab \tag{0.6}$$

as expected. The complication of this integral is that it no longer *separates* into two *independent* integrals – the x-integral depends on doing the y-integral *first*, and the result of the integration is y as a *function of* x and not a (possibly dimensioned) *number*.

However, it is rather *absurdly difficult* to integrate the area under (say) a semicircle in cartesian coordinates. Here's what it looks like. Remember that the cartesian formula for a circle of radius R is: $R^2 = x^2 + y^2$. If we try what we just did for a triangle, and solve this for $y(x) = +\sqrt{R^2 - x^2}$:

$$A_{\text{semicircle}} = \int_{\text{semicircle}} dx\, dy = \int_{-R}^{R} dx \int_0^{y(x)} dy = \int_0^a y(x)\, dx = \int_{-R}^{R} (R^2 - x^2)^{\frac{1}{2}} dx \tag{0.7}$$

and we're stuck. There is no *simple* u-substitution to give us the answer! This is the kind of nasty integral that requires both integration by parts and trigonometric substitution to evaluate[25]. Assuming, I think safely enough, that most of my readers cannot evaluate this integral without help, they are left with the somewhat unsatisfying method called *looking up the answer* (nowadays, on the internet). While remarkably efficient, this is hardly useful on a quiz or exam unless you are presented with an integral table along with the exam, and besides, at some point this stops testing your understanding of *basic physics* and starts simply connecting your grade to how good you are at *moderately advanced calculus!*

In a bit, we'll do this integral another way entirely that makes it *trivial* to evaluate – at the cost of *changing coordinate systems* to one where the boundaries of integration are "natural" ones and the integral separates into two *independent* integrals. However, just to finish the example, if we use the look-it-up method, we continue to get:

$$A_{\text{semicircle}} = \int_{-R}^{R} (R^2 - x^2)^{\frac{1}{2}} dx = x\frac{\sqrt{R^2 - x^2}}{2} + \frac{R^2}{2}\sin^{-1}\left(\frac{x}{R}\right)\Bigg|_{-R}^{R} = 2 \times \frac{R^2}{2}\frac{\pi}{2} = \frac{1}{2}\pi R^2 \tag{0.8}$$

or *half the area inside a circle of radius* R, as expected. But ouch!

We'll conclude with an integral over a function – one we might shortly consider to be a *surface charge distribution* function – of x and y *inside the rectangular boundary* used in the first example. The result of such an integral is the total charge of the surface, so it means something useful physically. Let $\dfrac{dQ}{dA} = \sigma(x, y) = \sigma_0\, xy$. Then $dQ = \sigma dA = \sigma(x, y) dx\, dy$ and:

$$Q = \int dQ = \int_0^a \int_0^b \sigma_0\, xy\, dx\, dy = \sigma_0 \int_0^a dx\, \frac{xy^2}{2}\Bigg|_0^b = \sigma_0 \frac{b^2}{2}\int_0^a x\, dx = \sigma_0 \frac{a^2 b^2}{4} \tag{0.9}$$

This integral still separates! The result is the *product* of an integral over x and an integral over y! We will learn to write this sort of integral more or less automatically as:

$$Q = \sigma_0 \int_0^a x\, dx \int_0^b y\, dy = \sigma_0 \frac{a^2}{2}\frac{b^2}{2} = \sigma_0 \frac{a^2 b^2}{4} \tag{0.10}$$

as it is now the product of two *trivial* integrals, ones we can do in our heads!

[25] https://www.emathzone.com/tutorials/calculus/integration-of-square-root-of-a2-x2.html As of the time of my writing this, this page had a fairly concise listing of the steps used to evaluate this integral, all ten or twelve of them. This course does *not* require integration at this level of skill!

Example 0.1.3: Integrating a Function over a Volume in Cartesian Coordinates

By this point, you should be ready for an example that uses *exactly* the kind of reasoning that will suffice for *nearly* all of the integrals we will need to do in this course. We may still need to do some work (largely, choosing the right coordinate frame and doing some math to express the problem in that frame) to get integrals that separate, but once we do they are all going to be 1-3 one dimensional, independent integrals, easily done using the methods of *ordinary* calculus covered on the One Sheet Math Review pages provided for the course[26].

Separable volume integrals in cartesian coordinates must therefore have rectilinear boundaries – for example from $(0,0,0)$ to (a,b,c) enclosing a total volume of abc – and must be over functions or distributions where x is independent of y is independent of z. We'll do a single example, just like the previous one – we'll find the total charge in exactly this volume given a volume charge distribution:

$$\rho(x,y,z) = \frac{dQ}{dV} = \rho_0 e^{-\kappa x}\cos(ky)z^2 \tag{0.11}$$

which corresponds to no realistic physics problem I can think of but which *does* let us practice writing down rectilinear separable integrals and doing them:

$$Q = \int dQ = \int \rho\, dV = \int \rho(x,y,z)\, dxdydz = \rho_0 \int_0^a e^{-\kappa x}dx \int_0^b \cos(ky)\, dy \int_0^c z^2\, dz \tag{0.12}$$

I get:

$$Q = \frac{1}{3\kappa k}\left(1 - e^{-\kappa a}\right)\sin(kb)c^3 \tag{0.13}$$

How about you?

0.1.2: Cylindrical Coordinates

In figure 0.3 the *cylindrical* coordinate system is illustrated, including a typical point $P = (r,\phi,z)$. The angle ϕ is swept out, by convention, counterclockwise (or in the $\hat{z}$ direction using the ***right hand rule*** in this ***right handed coordinate frame***) from the positive x-axis. r is the perpendicular distance of P from the z axis – the radius of the cylinder on which P lies. z is the usual cartesian z component of P.

An important note about the symbols chosen: Note well, some math and physics textbooks (more math than physics) use θ instead of ϕ for the azimuthal angle in cylindrical coordinates. Other textbooks, both math and physics, may well use ρ instead of r (to avoid colliding with r as defined for spherical polar coordinates, covered next). I cannot guess the symbols you, dear reader, might have used in the course(s) you took that hopefully covered coordinate systems, so please take a few minutes now to make sure you know what *we* will use in *this* course.

We will not use ρ as the radius of the cylinder, as ρ is already getting plenty of a workout in the textbook as both a charge density ρ and as conductivity ρ_c of a material, and we would rather not have to make sense of expressions like $\rho(\rho,\phi,z)$. We will use ϕ for the azimuthal

[26] http://www.phy.duke.edu/~rgb/Class/one_sheet_math_review.php

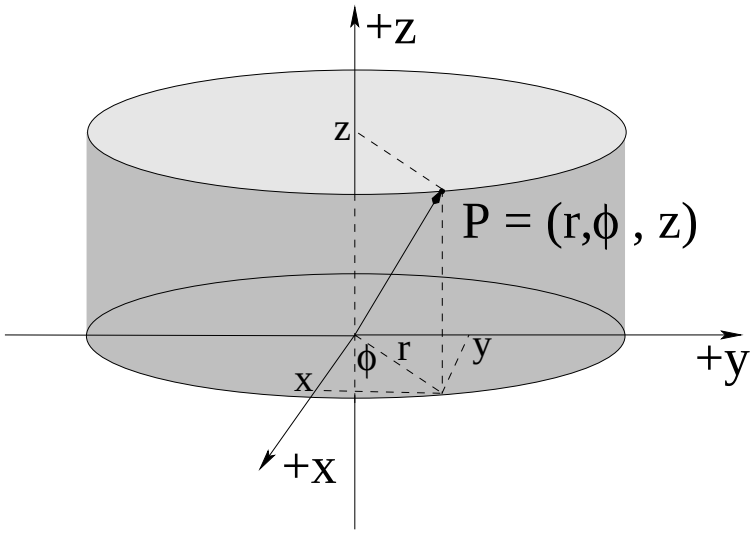


Figure 0.3: ***Cylindrical coordinates*** are basically plane polar coordinates in the x-y plane, plus the regular cartesian z-axis perpendicular to that plane. The point $P = (r, \theta, z)$ can be found by rotating a point at $x = r$ around z in the z direction (RHR) through the angle ϕ and then lifting it straight up a distance z.

angle around the z-axis because it then matches the *same* azimuthal angle used in spherical polar coordinates most commonly used in physics (as opposed to math) textbooks and research papers. However, nearly everybody learns plane polar coordinates as (r, θ), not (r, ϕ) or ρ, θ, in high school because of those pesky math textbooks, so I'm left trading off the possibility of confusion now against the certainty of confusion later for *some* students no matter what I do.

Defining the angle to be ϕ fortunately *agrees* with wikipedia article on the subject[27] but it does use ρ instead of r, which we will not do for the reason given above. I think that the spherical vs cylindrical contexts of physics problems are so obviously different that less confusion will result than that which can follow from overloading the symbol for ρ. Note well, Wikipedia *itself* isn't even consistent on this – it uses (r, ϕ) for plane polar coordinates, which (again, fortunately) corresponds with our usage, allowing us to define cylindrical coordinates as "Wikipedia's plane polar coordinates *plus a z-axis!*" If you want to have a better idea of the confusion and lack of consistency among even primary University mathematical physics textbooks, there is a nice table in Wolfram's Mathworld article on the subject that says it all[28].

There are unit vectors $(\hat{\boldsymbol{r}}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{z}})$ defined in cylindrical coordinates, and it is possible to write a vector in terms of them, but this *notation* is comparatively uncommon because the unit vectors themselves are *functions of the coordinates* in everything but the cartesian frame and this is "real multivariate calculus" stuff – we won't need it. In introductory physics it is much more common to just give the $\vec{\boldsymbol{A}} = (A_r, A_\phi, A_z)$ cylindrical components of the vector (which is also perfectly acceptable in both cartesian and spherical polar frames, note well).

[27] Wikipedia: http://www.wikipedia.org/wiki/Cylindrical coordinate system. I highly recommend that you look at this article while reading this section, especially if you *have* had multivariate calculus.

[28] https://mathworld.wolfram.com/CylindricalCoordinates.html This table *still* doesn't do the subject justice. If math is the language of physics and engineering, then there are distinct *dialects* of math used in different contexts and by different authors, which can easily lead to confusion if you aren't aware of it.

With the coordinates defined in the picture, we can see how to go back and forth between cylindrical coordinates and cartesian coordinates. To go from cylindrical to cartesian:

$$x = r\cos\phi \qquad y = r\sin\phi \qquad z = z \tag{0.14}$$

(just like plane polar coordinates in the x-y plane and z is unchanged) and to go from cartesian to cylindrical:

$$r = \sqrt{x^2 + y^2} \qquad \phi = \tan^{-1}\frac{y}{x} \qquad z = z \tag{0.15}$$

Note well: When evaluating ϕ from x and y one has to *pay attention to the quadrant in which P lies*, as the range of the inverse tangent is typically $(-\pi/2, \pi/2]$ and you may have to use the unit circle and a bit of thought to shift the result returned by a calculator for ϕ into the correct quadrant!

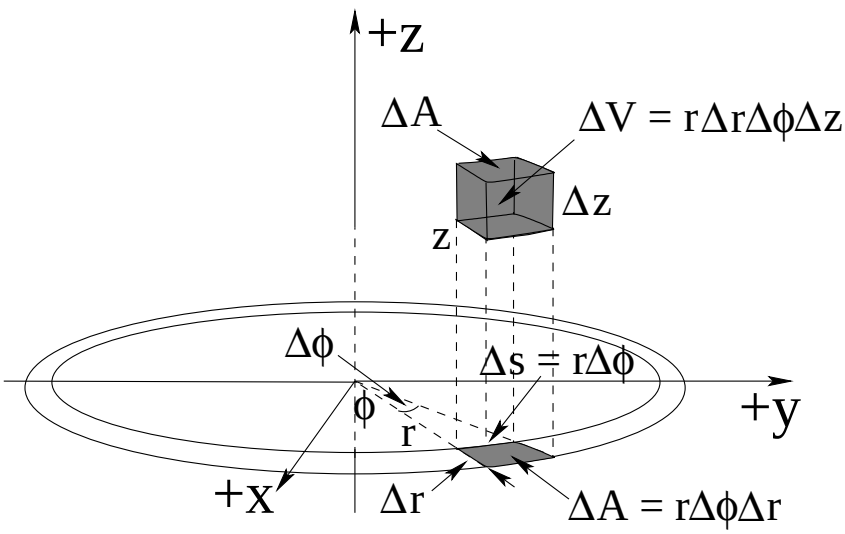


Figure 0.4: The differential elements appropriate to cylindrical coordinates, expressed as Δ's.

In figure 0.4 the differential elements in cylindrical coordinates are displayed as finite-sized "chunks" of each coordinate (as before). Again we can think of these as going to the point P and then swinging around the z-axis through an angle $\Delta\phi$ to create a small arc of length $\Delta s = r\Delta\phi$ as one length element, then "pushing out" that arc by a second length element Δr to make an area element $\Delta A = r\Delta\phi\Delta r$, and then lifting this surface up a length element Δz in the z-direction to make a volume element $\Delta V = r\Delta r\Delta\phi\Delta z$

We will summarize this as:

Three length elements: $\Delta r \Rightarrow dr$, $\Delta s \Rightarrow ds = r\,d\phi$, $\Delta z \Rightarrow dz$.

Three area elements: $\Delta A = \Delta r\Delta s \Rightarrow dA_1 = r\,dr\,d\phi$ is shown. There is a *second* area element (not shown) on the surface of the cylinder containing P consisting of arc length element ds being "pushed up" in the z-direction by dz, so that $dA_2 = ds\,dz = r\,d\phi\,dz$. We will integrate this below to obtain the area of the label of a soup can (for example). Finally, there is a third area element on an angled flat surface with constant angle ϕ ("where the knife cuts" when cutting a piece of cake, for example): $dA_3 = d\rho\,dz$.

One volume element: $\Delta V = \Delta r\Delta s\Delta z \Rightarrow dV = rdr\,d\phi\,dz$. We can view this as *either* dA in the polar plane times dz *or* as the area dA on the cylinder pushed out by the distance dr.

Note that I am increasingly jumping from Δ notation to d notation without warning. This is deliberate. Eventually you should view them as different ways of writing the same thing – physicists are as likely to label a length element in a figure ds instead of Δs even though it is obviously not differential in length! Note also that *in* the limits that the Δ's become differential d's, the area element dA becomes rectangular in shape – we neglect by convention the "extra" length on the outside of $\Delta s = (r+\Delta r)\Delta\phi \Rightarrow r\Delta\phi + \Delta r\Delta\phi$ because it contains two differentials, not just one. The term $\Delta r\Delta\phi$, in other words, is "infinitely smaller" than the leading order piece $r\Delta\phi$ when both Δr and $\Delta\phi$ have become differentially small. This is the Tao of Calculus, if you will recall...

Example 0.1.4: Finding the Area of a Soup Can Label

Here is an example we will frequently encounter in Gauss's Law problems with cylindrical symmetry. The electric field flux through a tiny chunk of the the surface of a cylinder of radius r in these problems turns out to be $d\Phi_e = E_r\, dA$ where E_r is constant everywhere on the cylinder, so all we have to do to find the total flux is evaluate $A = \int dA$ for a cylinder of radius r and length (say) ℓ.

While we can do this one in our heads (see below) let's do it "the hard way" with explicit integration. We need to integrate:

$$dA(= dA_2) = r\, d\phi\, dz$$

where ϕ goes from 0 to 2π and z from 0 to H. Even the hard way is trivial:

$$A = \int dA = \int_0^{2\pi}\int_0^{\ell} r\, d\phi\, dz = r\left(\int_0^{2\pi} d\phi\right) \times \left(\int_0^{\ell} dz\right) = 2\pi r\ell \tag{0.16}$$

This is obviously correct! It is the area you would measure if you imagine that the area in question is the area of the label of a soup can with radius r and height ℓ! We can use our "mental scissors" to snip this label right off of the can and *unroll* it into a rectangle with one side equal to $2\pi r$ (the circumference of the can) and the other equal to ℓ in length. The area is then just width times height or $A = 2\pi r\ell$!

We don't have a lot of occasions to integrate over dA_1 or dA_2 explicitly, but dA_1 is most convenient one to use to form the (one) volume element so I gave it precedence in figure 0.4 above.

Example 0.1.5: Volume of a Cylinder

Again we can get this result heuristically easily enough, but let's integrate it the hard way and use the result to inspire a more difficult example following. To find the volume of a cylinder of radius R and height H, we start with:

$$dV = r\, dr\, d\phi\, dz \tag{0.17}$$

and integrate it over suitable limits:

$$V = \int dV = \int_0^{R}\int_0^{2\pi}\int_0^{H} r\, dr\, d\phi\, dz \tag{0.18}$$

The integral separates, as the limits of integration of each coordinate is not a function of the other two, and because there is no other function inside which one coordinate is a function of the others. Hence:

$$V = \left(\int_0^R r\,dr\right) \times \left(\int_0^{2\pi} d\phi\right)\left(\int_0^H dz\right) = \frac{R^2}{2} \times 2\pi \times H = \pi R^2 H \tag{0.19}$$

Again this is obviously the area of the base of the cylinder times its height, so we worked way harder than we had to. Note that two of these integrals *are* evaluating the area of the base as $\int dA_1 = \pi R^2$ above; our example includes finding the area inside a circle!

Example 0.1.6: Finding the Volume of a Right Circular Cone

This is a tricky one! Indeed, it is the cylindrical equivalent of finding the area of a right triangle, a ***non***-separable example! To find the volume of a right circular cone of radius R and height H, we will mentally place its apex at the origin and its base at the height H. This may seem strange at first, but it makes it very easy to find the radius of the cone as a function of z.

$$r(z) = \frac{R}{H}z$$

To see this, note that when $z = 0$, $r = 0$ (apex at the origin) and when $z = H$, $r = R$ (circular base at height H). The volume is then the integral of dV as before, but the upper limit of integration for dr is a function of z!

We need to do the integrals in a certain order, then, and differentiate $r(z)$ the limit of integration at height z from r', a variable we integrate over. Let's write it down, as that makes it clear enough:

$$V = \int dV = \int_0^H \int_0^{r(z)} \int_0^{2\pi} r'\,dr' dz d\phi = \left(\int_0^H dz\right) \times \left(\int_0^{r(z)} r'\,dr'\right) \times \left(\int_0^{2\pi} d\phi\right) \tag{0.20}$$

The first (ϕ) integral just gives us 2π. The next two are tricksy! Let's do the r' integral next:

$$\int_0^{r(z)} r'\,dr' = \left.\frac{r'^2}{2}\right|_0^{r(z)} = \frac{r(z)^2}{2} = \frac{R^2}{2H^2}z^2$$

Now we can do the z integral, which is no longer as trivial as it first appeared!

$$\int_0^H \frac{R^2}{2H^2}z^2 dz = \frac{R^2}{2H^2} \times \frac{H^3}{3}$$

Note that this *one* integral is the result of *both* of the remaining integrals – we had to integrate (over z) the result of integrating over r' because it was a *function* of z.

We combine all of the pieces to get:

$$\boxed{V = 2\pi \times \frac{R^2}{2H^2} \times \frac{H^3}{3} = \frac{\pi R^2 H}{3}} \tag{0.21}$$

You *might* even remember that this is the correct answer from a previous geometry or calculus class – the volume of a right circular cone is a third of the volume of the cylinder with the same radius and height!

Example 0.1.7: Evaluating a Volume Charge Density

Finally, we'll do one non-trivial *separable* example of integrating a radial volume charge density *distribution* to find the total charge in a cylinder of radius R and height H. Suppose the charge density varies *only* with the radius r such that $\rho(r,\phi,z) = \frac{A}{r}e^{-\kappa r}$ where A has appropriate dimensions to make the volume integral have units of charge. This is a physically plausible model of charge that is concentrated along the z axis but dies off exponentially the further you are away from the axis. Then (letting $u = -\kappa r$, $du = -\kappa\, dr$ and recalling that $dV = r dr\, d\phi\, dz$):

$$\begin{aligned} Q &= \int_{\text{cylinder}} \rho(r,\phi,z)\, dV \\ &= \int_0^R \int_0^{2\pi} \int_0^H \frac{A}{r} e^{-\kappa r} \left(r\, dr\, d\phi\, dz\right) \\ &= A \int_0^R e^{-\kappa r} dr \int_0^{2\pi} d\phi \int_0^H dz = \left[A \int_0^R e^{-\kappa r}\right] [2\pi]\, [H] \\ &= -\frac{2\pi H A}{\kappa} \int_0^{-\kappa R} e^u\, du = \boxed{\frac{2\pi H A}{\kappa}\left(1 - e^{-\kappa R}\right)} \end{aligned} \tag{0.22}$$

0.1.3: Spherical Polar Coordinates

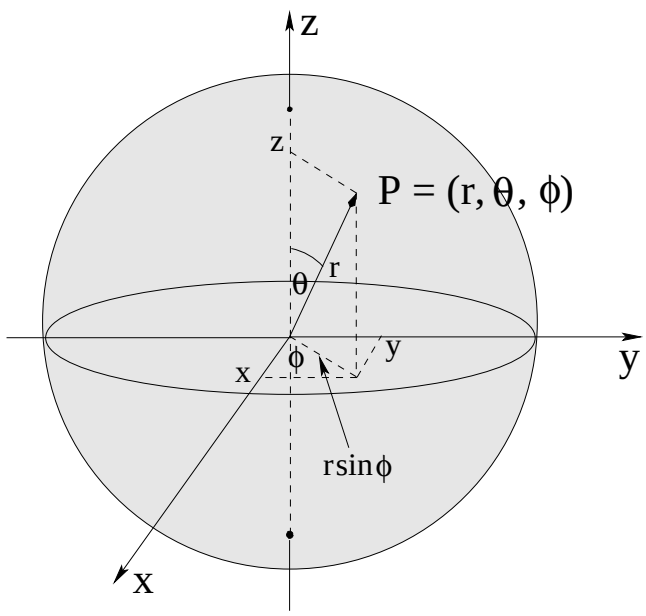


Figure 0.5: ***Spherical polar coordinates*** represent an arbitrary point as $P = (r,\theta,\phi)$. Note that ϕ is the same as in cylindrical coordinates when the coordinates of point P are projected onto the x-y plane, but r is quite different. θ is the angle between the positive z-axis and the line from the origin to the point P.

In figure 0.5 the *spherical* coordinate system is illustrated, including a typical point $P = (r,\theta,\phi)$. In this coordinate frame, r is the distance of the point P from the origin – basically the radius of the *sphere* that contains P. The remaining two angles are used to locate P on the surface of that sphere. If you start with a point at the "north pole", where $z = r$, and then rotate the point around the y axis in the direction of the x axis by the angle $\theta \in [0,\pi]$, you locate the

circle concentric to the z-axis on which the point P lies. One then rotates the point aximuthally around the z axis through the angle ϕ (counterclockwise, or "in the z direction" according to the ***right-hand rule***) to end up with it as the point P.

Alternatively (as figure 0.5 most clearly illustrates), you can visualize *projecting* the point P into the x-y plane so that ϕ is the usual plane polar angle of the projection, but the "cylindrical" radius of its projection is now $r\sin\theta$. From the picture it is also obvious that $z = r\cos\theta$, that $r = \sqrt{x^2+y^2+z^2}$, and so on, see below.

An important note about the symbols chosen: Some math and physics textbooks (more math than physics) ***switch θ and ϕ***. Other textbooks, both math and physics, may well use ρ instead of r (to avoid colliding with r as defined for spherical polar coordinates, but in the other direction). The convention I'm using matches that used in the Wikipedia article on spherical coordinates[29] and again there is a table in Wolfram's related Mathworld article[30] that is *even longer* than the table for cylindrical coordinates!

As before, there are unit vectors $(\hat{\boldsymbol{r}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$ defined in spherical polar coordinates that are again because the unit vectors themselves are functions of the coordinates and beyond our scope. We will just give the spherical polar components $\vec{\boldsymbol{A}} = (A_r, A_\theta, A_\phi)$ cylindrical components of the vector when necessary and avoid expressing e.g. $\hat{\boldsymbol{r}}$ in cartesian components or as a function of θ and ϕ as it will not be needed at this time.

From the picture in figure 0.5 we can easily see how to go back and forth between spherical polar coordinates and cartesian coordinates and vice versa:

$$x = r\sin\theta\cos\phi \qquad y = r\sin\theta\sin\phi \qquad z = r\cos\theta \tag{0.23}$$

and:

$$r = \left(x^2+y^2+z^2\right)^{\frac{1}{2}} \qquad \phi = \tan^{-1}\left(\frac{y}{x}\right) \qquad \theta = \cos^{-1}\left(\frac{z}{(x^2+y^2+z^2)^{\frac{1}{2}}}\right) \tag{0.24}$$

Hopefully it is obvious how to go from spherical polar to cylindrical coordinates as well, although you will likely not need to do this in this class.

In figure 0.6 the differential elements in cylindrical coordinates are displayed as finite-sized "chunks" of each coordinate, but this time I didn't bother to label them as Δ's, I went straight to differential form with d's. The easiest way to visualize and remember the elements is to go to the point $P = (r, theta, phi)$ and rotate it *out* from the z-axis through an angle $d\theta$ on the great circle of radius r to create a small arc of length $r\,d\theta$ as one length element. The second length element is obtained by again starting at the point P and rotating it around the z-axis by the angle $d\phi$ to make a length element $r\sin\theta\,d\phi$ (note the length is obtained from the *projection* of this arc onto the x-y plane as lies in a plane parallel to this plane). The third length element comes from "pushing out" P radially by dr.

We can then assemble three area elements by taking these orthogonal length elements two at a time, and one volume element from the product of all three:

We will summarize this as:

[29] Wikipedia: http://www.wikipedia.org/wiki/Spherical coordinate system. I highly recommend that you look at this article while reading this section, especially if you *have* had multivariate calculus.

[30] https://mathworld.wolfram.com/SphericalCoordinates.html ***Sheesh!*** *Seven* of them!

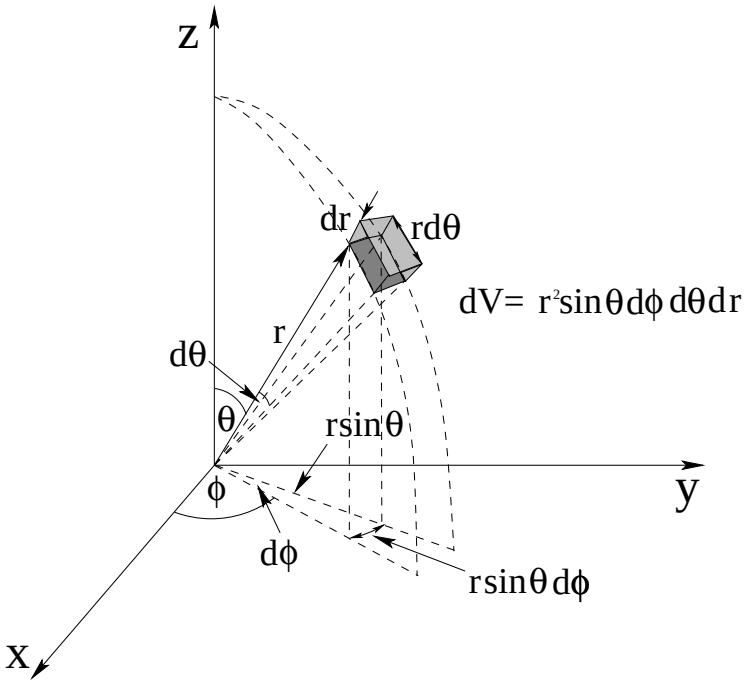


Figure 0.6: The differential elements appropriate to cylindrical coordinates, expressed as Δ's.

Three length elements: dr, $r\,d\theta$, $r\sin\theta\, and d\phi$.

Three area elements: $dA_1 = r\,dr\,d\theta$, $dA_2 = r\sin\theta\,dr\,d\phi$, and $dA_3 = r^2\sin\theta\,d\theta\,d\phi$. Of these three, dA_3 is the most important/useful (see example below).

One volume element: $\Delta V = r^2\sin\theta\,dr\,d\theta\,d\phi$ (basically, the area element dA_3 pushed out by a radial distance dr to form our differentially rectilinear volume).

Example 0.1.8: Finding the Area of a Sphere

We can use the differential elements described in detail above to easily find e.g. the surface area of a sphere of (fixed) radius R as follows. We start with the area element of the constant $r = R$ spherical surface (dA_3 above) and integrate the θ and ϕ contributions independently. This seems as if it is pretty trivial (and it is) but note well the trick I use to do the θ integral as it will *often* be the only or best way to proceed for more difficult spherical integrals that involve θ explicitly!

So:

$$\begin{aligned} dA &= R^2\sin\theta\,d\theta\,d\phi \\ A = \int dA &= R^2\int_0^{\pi}\int_0^{2\pi}\sin\theta\,d\theta\,d\phi \\ A &= R^2\left(\int_0^{\pi}\sin\theta\,d\theta\right)\times\left(\int_0^{2\pi}d\phi\right) \\ A &= 2\pi R^2\left(\int_0^{\pi}\sin\theta\,d\theta\right) \end{aligned} \tag{0.25}$$

Note Well that the limits of the θ integral are 0 to $\boldsymbol{\pi}$, not 2π. If you integrate both θ and ϕ from 0 to 2π you will double count every point (look until you see why).

At this point we could easily do the remaining integral using the standard formula to get "2" as the result, but instead I'm going to use u-substitution. This is overkill for *this* integral, but in the next example, it will be the only game in town! Let:

$$u = \cos\theta \quad \text{so that} \quad du = -\sin\theta d\theta \tag{0.26}$$

Then we can transform the integral into u-form:

$$\int_0^\pi \sin\theta\, d\theta = -\int_{\cos 0}^{\cos\pi} du = -\int_1^{-1} du = \int_{-1}^1 du = u|_{-1}^1 = 2 \tag{0.27}$$

In the future, I will often use this trick to go from:

$$\int_0^\pi f(\sin\theta \text{ or } \cos\theta)\sin\theta d\theta \qquad \Longrightarrow \qquad \int_{-1}^1 f(\sqrt{1-u^2} \text{ or } u)du$$

in a single step, be warned!

Either way, we do that last integral to get:

$$\boxed{A = 4\pi R^2} \tag{0.28}$$

which, of course, you should already know.

To conclude: A very useful thing to remember is that the ***double integral over θ and ϕ by themselves for a sphere is 4π***. This is called "the integral over the solid angle of the sphere". The units in this integral are dimensionless (angles are always dimensionless) but just as planar angles are expressed in dimensionless *radians* in the SI, solid angles are expressed in the SI as "stereo radians", or ***steradians***. The area of a sphere is thus:

$$A = R^2 \text{ (meters}^2) \times 4\pi \text{ (dimensionless steradians)} = 4\pi R^2 \text{ (meters}^2)$$

where we simply drop the "steradians" in the final expression as they are dimensionless!

Example 0.1.9: Integrating a Function of $\cos\theta$ Over a Spherical Surface

It isn't always the case that you integrate only *constants* over the surface of a sphere. Sometimes you have to integrate a function of θ! Arbitrary functions of θ can easily be challenging, but a very common case you will encounter (more in later courses than in this one, but a bit even in this one) is integrating a function of not θ, but $\cos\theta$, over the surface of a sphere. This example embraces integrals over $\sin\theta$ as well, because one can express $\sin\theta$ in terms of $\cos\theta$.

Let's pick a case you are actually likely to encounter sooner or later. Suppose:

$$f(\theta,\phi) = \sin^2\theta$$

and we would like to integrate:

$$\int_{\text{sphere}} f(\theta,\phi)\, dA = \int_0^\pi \int_0^{2\pi} \sin^2\theta \times R^2 \sin\theta\, d\theta\, d\phi \tag{0.29}$$

over the spherical surface with (constant) radius R. We *immediately* use the u-substitution trick illustrated above to convert this into a *simple one dimensional integral*. Recall $u = \cos\theta$, $du = -\sin\theta\, d\theta$, and observe that

$$\sin^2\theta = 1 - \cos^2\theta = 1 - u^2$$

from the familiar trig identity $\sin^2\theta + \cos^2\theta = 1$ to get:

$$\int_{\text{sphere}} f(\theta,\phi)\, dA = 2\pi R^2 \int_{-1}^{1} (1-u^2)\, du = 2\pi R^2 \left(u - \frac{u^3}{3}\right)\Bigg|_{-1}^{1} = 2\pi R^2 \frac{4}{3} = \frac{8\pi R^2}{3} \tag{0.30}$$

Observe that I simply 'did' the ϕ integral over its appropriate limits to get the 2π right away (since f was independent of ϕ) and in a single step converted the $\sin^3$ integral into u-form, where it was a trivial power law integral! This trick isn't *always* the best one to use – a *different* trick is used to integrate:

$$\int_0^{\pi} \sin^2\theta\, d\theta = \frac{\pi}{2}$$

because integrating

$$\int_{-1}^{1} \left(1-u^2\right)^{\frac{1}{2}} du$$

doesn't get us very far in terms of our five simple integral forms!

Example 0.1.10: Integrating the Volume of a Sphere

Evaluating the volume of a sphere of radius R is now straightforward. The spherical element is:

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

so we perform the usual u-substitution for $\sin\theta\, d\theta$ and sort this into three independent integrals:

$$V = \int dV = \left(\int_0^R r^2\, dr\right)\left(\int_0^{\pi} \sin\theta\, d\theta\right)\left(\int_0^{2\pi} d\phi\right) = \left(\frac{R^3}{3}\right)(2)\,(2\pi) = \boxed{\frac{4\pi R^3}{3}} \tag{0.31}$$

A second (and faster!) way to visualize and express this is to *start* with the known result that the *area* of a sphere of radius r is $A = 4\pi r^2$ so that dV is the **volume of a spherical shell of radius r and thickness dr**:

$$V = \int dV = \int_0^R A\, dr = 4\pi \int_0^R r^2\, dr = \boxed{\frac{4\pi R^3}{3}} \tag{0.32}$$

Example 0.1.11: Integrating a Radial Distribution over a Sphere

A very common problem you will encounter is evaluating, for example, the total charge in a sphere of radius R when it is distributed according to some function of r on the interior. Let's pick a simple/integrable charge distribution function:

$$\rho(r) = \frac{dQ}{dV} = Ar^2$$

where A has the appropriate dimensions to produce a total charge Q from the integral:

$$Q = \int dQ = \int_{\text{sphere}} Ar^2 dV = \int_0^R \int_0^\pi \int_0^{2\pi} \left(Ar^2\right) \times \left(r^2 \sin\theta \, dr \, d\theta \, d\phi\right) \tag{0.33}$$

We perform the usual u-substitution for $\sin\theta \, d\theta$ and sort this into three independent integrals:

$$Q = A \left(\int_0^R r^4 \, dr\right) \left(\int_{-1}^1 du\right) \left(\int_0^{2\pi} d\phi\right) \tag{0.34}$$

or

$$\boxed{Q = \frac{4\pi R^5}{5}} \tag{0.35}$$

Example 0.1.12: Evaluating the Moment of Inertia of a Uniform Sphere

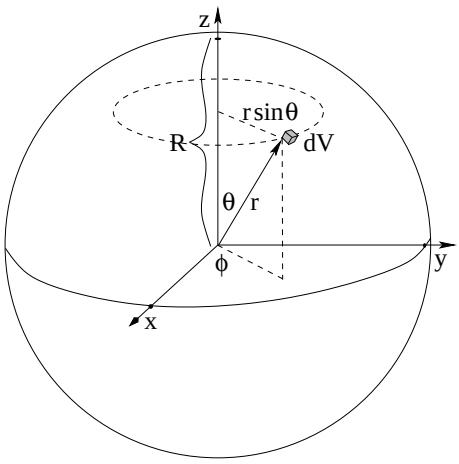


Figure 0.7: A differential chunk of mass (or charge!) in a solid sphere of mass of volume dV moves in a circle of radius $r \sin\theta$ as the sphere is rotated around the z-axis.

In figure 0.7 a tiny chunk of mass (or charge!) with volume dV is drawn. The mass of this chunk is (using the litany we learned in mechanics and will learn in the context of electromagnetism as well herein) "the mass of the chunk is the mass per unit volume times the volume of the chunk, or (for uniformly distributed mass M in a sphere of radius R as shown):

$$dm = \rho_M \, dV = \left(\frac{3M}{4\pi R^3}\right) \times \left(r^2 \sin\theta \, dr \, d\theta \, d\phi\right) \tag{0.36}$$

This chunk moves in a circle of radius $r \sin\theta$ if the solid sphere is rotated around its z-axis. This figure has enough symmetry that this is a "principle axis", so we can find e.g. its total moment of inertia around the z-axis by summing up:

$$dI = dm \, (r \sin\theta)^2 = \left(\frac{3M}{4\pi R^3}\right) r^4 \sin^3\theta \, dr \, d\theta \, d\phi \tag{0.37}$$

for the entire sphere.

This is now a problem that combines *both* of the nontrivial examples we just did into one! We have to do the integral:

$$\begin{aligned} I = \int dI &= \frac{3M}{4\pi R^3}\int_0^R r^4\,dr\,\int_{-1}^{1}(1-u^2)du\,\int_0^{2\pi}d\phi \\ &= \frac{3M}{4\pi R^3}\frac{R^5}{5}\times\frac{4}{3}\times 2\pi \\ &= \frac{3M}{4\pi R^3}\frac{8\pi R^5}{3\times 5} \end{aligned}$$

or:

$$\boxed{I_{\text{sphere}} = \frac{2}{5}MR^2} \quad (!) \tag{0.38}$$

This is how we actually get the result we used extensively in mechanics! We will encounter ***exactly this integral*** in the chapter where we are evaluating the magnetic moments of spinning uniform balls of charge!

Summary

That's enough preliminary stuff. At this point, if you've read all of this "week"'s material and vowed to adopt the method of three passes in all of your homework efforts, if you've bookmarked the math help or downloaded it to your personal ebook viewer or computer, if you've realized that your brain is actually something that you can *help and enhance* in various ways as you try to learn things, then my purpose is well-served and you are as well-prepared as you can be to tackle physics.

There isn't really any homework for the preliminary part of the book other than to read over it to accomplish these goals, but here are a few things you could do on your own – or even just think about *how* you would do them – if you want to put it into practice:

- Skim read the ***How to Learn Physics*** section, then then read it like a novel, front to back. Think about the connection between engagement and learning and how important it is to try to have *fun* in a physics course. Think about at least one time in the past where you were extremely engaged in a course you were taking, had lots of fun in the class, and had a really great learning experience. Contrast it to a course where were lost and hated it. What made the two experiences so different? Sure, maybe the material was boring, maybe the book was terrible, maybe the teacher was awful – none of these are directly under your contro – but were there things that *were* under your control that you could have used to *both* have more fun *and* do better?

- Skim-read the entire content of the ***Math Needed for Introductory E&M*** section above. Identify things that it covers that you *don't* remember or *never* learned, and *don't* understand (yet).

- Apply the *Method of Three Passes* to learning the things you just identified. Over a few days, you can learn each coordinate system well enough to draw the essential pictures defining its coordinates, how to go back and forth between it and cartesian coordinates, and its differential elements. See if you can get to where you can work the examples (at least) without looking and without real pain.

  Note well: You may well have found the content *boring* on the third pass because it was so familiar to you, but that's *not a bad thing!* If you learn physics and its requisite math so thoroughly that its laws become *boring*, not because they confuse you and you'd rather play World of Warcraft but because you know them so well that reviewing them isn't adding anything to your understanding, well *damn* you'll do well on the exams testing the concept, won't you?

II: Electrostatics

Week 1: Discrete Charge and the Electrostatic Field

Summary

- **Charge**

  Objects can carry a (net) charge q when "electrified" various ways. This charge comes in two flavors, + and -. Like charges exert a long range (action at a distance) repulsive force on one another. Unlike charges attract. The SI unit of charge is called the *Coulomb* (C).

- **Charge Quantization**

  Charge is discrete and quantized in units of $e/3$, where $e = 1.6 \times 10^{-19}$ C, but we can never directly observe bare particles with the thirds (quarks). All charges we can directly measure on independent particles come in units of e, the charge of the electron or proton.

- **Approximate Continuous Charge Distributions**

  When we study charge distributions in actual matter (with many many charged atoms in even a tiny chunk) we will often be able to *approximate* the average distribution of charge as being *continuous*, so that we can use calculus and integration instead of discrete summations over absurdly large numbers of charges. To facilitate the treatment of continuous charge distributions next week, we will go ahead and define the following *charge densities*:

$$\begin{aligned} \rho &= \frac{dq}{dV} \\ \sigma &= \frac{dq}{dA} \\ \lambda &= \frac{dq}{dx} \end{aligned}$$

- **Charge Conservation**

  Net charge is a conserved quantity in nature. Later we will learn to write the conservation law mathematically in terms of the flux of the current density, but we don't yet have the mathematical tools to do this with.

- **Mobility of Charge in Matter**

  Matter comes in three distinct forms:

  - **Insulators**
  - **Conductors**
  - **Semiconductors**

- **Coulomb's Law**

  From performing many careful experiments directly measuring the forces between static charges and from the consistent observations of many other things such as the electric structure of atoms, the conductivity of metals, the motion of charged particles, and much, much more, we infer that for any two stationary charges, the *experimentally verified* electrostatic force acting *on* charge 1 *due to* charge 2 is:

$$\vec{\boldsymbol{F}}_{12} = k_e q_1 q_2 \frac{(\vec{\boldsymbol{r}}_1 - \vec{\boldsymbol{r}}_2)}{|\vec{\boldsymbol{r}}_1 - \vec{\boldsymbol{r}}_2|^3}$$

  Note that it acts on a line *from* charge 2 *to* charge 1, is proportional to both charges, and is inversely proportional to the distance that separates them squared.

- **The Electrostatic Constant k_e**

  The electrostatic constant k_e sets the scale; it is a *very important number* (as we shall see) – a genuine constant of nature as was G for the gravitational field. It is often expressed in terms of a related quantity called the *permittivity of free space*, ϵ_0, which is more useful for advanced treatments of electrodynamics. We will often/generally use k_e instead in this course (because it is very easy to remember), but I would like you to know the relationship between this quantity and ϵ_0 so that you can easily calculate the latter if you should ever need it or care.

$$k_e = \frac{1}{4\pi\epsilon_0} = 9 \times 10^9 \ \ \frac{\mathrm{N-m}^2}{\mathrm{C}^2}$$

  This is accurate to something like 3 significant figures, which is plenty for our purposes. Note also that you don't have to *remember* the units of k_e per se, you can figure them out by just remembering Coulomb's Law (which you have to know anyway). Newtons on the left, coulombs squared on top and meters squared on the bottom on the right.

- **Electrostatic Field**

  The fundamental definition of electrostatic field produced by a charge q at position $\vec{\boldsymbol{r}}$ is that it is the electrostatic force per unit charge on a small test charge q_0 placed at each point in space $\vec{\boldsymbol{r}}_0$ in the limit that the test charge vanishes:

$$\vec{\boldsymbol{E}} = \lim_{q_0 \to 0} \frac{F}{q_0}$$

  or

$$\vec{\boldsymbol{E}}(\vec{\boldsymbol{r}}_0) = k_e q \frac{(\vec{\boldsymbol{r}}_0 - \vec{\boldsymbol{r}})}{|\vec{\boldsymbol{r}}_0 - \vec{\boldsymbol{r}}|^3}$$

  If we locate the charge q at the origin and relabel $\vec{\boldsymbol{r}}_0 \to \vec{\boldsymbol{r}}$, we obtain the following simple expression for the electrostatic field of a point charge:

$$\vec{\boldsymbol{E}}(\vec{\boldsymbol{r}}) = \frac{k_e q}{r^2}\hat{\boldsymbol{r}}$$

- **Superposition Principle**

  Given a collection of charges located at various points in space, the total electric field at a point is the sum of the electric fields of the individual charges:

$$\vec{\boldsymbol{E}}(\vec{\boldsymbol{r}}) = \sum_i \frac{k_e q_i(\vec{\boldsymbol{r}} - \vec{\boldsymbol{r}}_i)}{|\vec{\boldsymbol{r}} - \vec{\boldsymbol{r}}_i|^3}$$

  To find the electrostatic field produced by a charge density distribution, we use the superposition principle in *integral* form:

$$\vec{\boldsymbol{E}}(\vec{\boldsymbol{r}}) = k_e \int \frac{\rho(\vec{\boldsymbol{r}}_0)(\vec{\boldsymbol{r}} - \vec{\boldsymbol{r}}_0)d^3r_0}{|\vec{\boldsymbol{r}} - \vec{\boldsymbol{r}}_0|^3}$$

  Because one has to integrate over the vectors, this integral is remarkably difficult. We'll revisit it in a much more similar form when we get to electrostatic *potential*, a scalar quantity.

- **Electric Dipoles**

  When two electric charges of equal magnitude and opposite sign are bound together, they form an *electric dipole*. The *dipole moment* of this arrangement is the source of a characteristic electrostatic field, the *dipole field*. The dipole moment of the two charges is defined to be:

$$\vec{\boldsymbol{p}} = q\vec{\boldsymbol{l}}$$

  where q is the magnitude of the charge and $\vec{\boldsymbol{l}}$ is the vector that points from the negative charge to the positive charge.

  When an electric dipole $\vec{\boldsymbol{p}}$ is placed in a *uniform* electric field $\vec{\boldsymbol{E}}$, the following expressions describe the force and torque acting on the dipole (which tries to align itself with the applied field):

$$\begin{aligned} \vec{\boldsymbol{F}} &= 0 \\ \vec{\boldsymbol{\tau}} &= \vec{\boldsymbol{p}} \times \vec{\boldsymbol{E}} \end{aligned}$$

  Associated with this torque is the following potential energy:

$$U = -\vec{\boldsymbol{p}} \cdot \vec{\boldsymbol{E}}$$

  and from this, we can see that the force on the dipole in a more general (non-uniform) field should be:

$$\vec{\boldsymbol{F}} = -\vec{\boldsymbol{\nabla}}U = \vec{\boldsymbol{\nabla}}(\vec{\boldsymbol{p}} \cdot \vec{\boldsymbol{E}})$$

  which is actually nontrivial to compute.

This completes the chapter/week summary. The sections below illuminate these basic facts and illustrate them with examples.

1.1: Charge

In nature we can readily observe electromagnetic forces. In fact, we can do little else. In a very fundamental sense, we *are* electromagnetism. Electromagnetic forces bind electrons to atomic nuclei, bond atoms together to form molecules, mediate the interactions between molecules that allow them to change and organize and, eventually, live. The energy that is used to support life processes is electromagnetic energy. The objects that we touch, or hear, or taste, or smell, the light that we see, the organized pattern of neural impulses that we use to think about the input from our senses – all are electromagnetic.

Given its ubiquity, it should come as no surprise that the directed observation and study of electricity is quite ancient. It was studied, and written about, at least 3000 years ago, and artifacts that may have been primitive electrical batteries have been discovered in the Middle East that date back to perhaps 250 BCE. It is revealing that the very *word* electricity and the name of the elementary particle most visibly responsible for its transport is derived from the greek word for amber, *electron*. One of the first *recorded* observations of electrical force was the static electrical force created between amber, charged by rubbing it with wool, and small bits of wool or hair.

However, it took until the Enlightenment (roughly 1600) and the invention of physics and calculus for the scientific method to develop to where systematic studies of the phenomenon could occur, and it wasn't until the middle 1700s that the correct model for *electrical charge*[31] was proposed. From that point rapid progress was made over a period of 250 years, culminating in our contemporary understanding of electromagnetic forces as one aspect of a unified field theory.

As pointed out above, even our prehistoric ancestors no doubt knew about "charge". The experience of rubbing one's body against fur on a cold, dry day and thereby picking up enough charge to generate a spark is probably tens of thousands of years old. By the historic time of the Greeks, it was known that rubbing amber with wool or fur would charge the amber, and the term electricity is derived from the Greek word for amber, *elektron*. We now know that the charge produced on the amber is negative.

During the Enlightenment much more systematic studies were made of this phenomenon. It is possible to charge *many* objects by rubbing them against other objects. For example, if one rubs glass with silk, one literally rubs electrons off of the molecules that make up the glass and transfer them to the silk. The silk becomes negatively charged and the glass becomes positively charged. The study of this continues today where this sort of charge transfer due to friction is called the Triboelectric effect[32] . Recall that the study of friction is called "Tribology", so that this makes sense.

In order to do the experimental work that led to the identification of the two kinds of charge and our ability to manipulate electrostatic charges and measure forces quantitatively, it was necessary to find ways of *systematically* charging up conductors with specific increments of charge. One could use the triboelectric effect to charge up a piece of glass or amber or bone or metal, but the amount and even the sign of the charge produced was not always consistent. Charge also has a habit of "leaking away" from anything that is charged because same-sign

[31] Wikipedia: http://www.wikipedia.org/wiki/electric charge.
[32] Wikipedia: http://www.wikipedia.org/wiki/Triboelectric effect.

charge is always repulsive.

It is difficult to properly and completely summarize all of the people that contributed to the formal discoveries. Otto von Guericke almost by accident built the first triboelectric electrostatic generator. Charge generated in this way could be stored in *Leyden Jars*[33] .

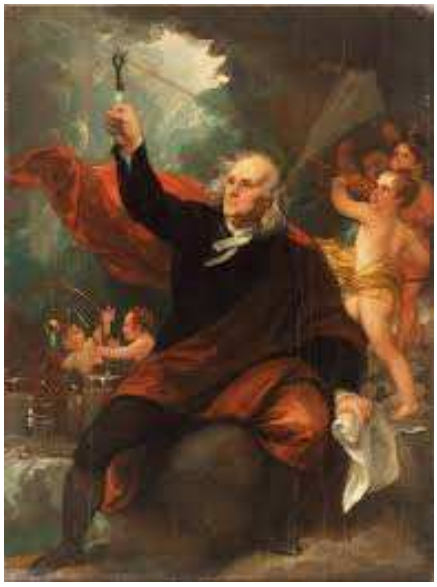


Figure 1.1: Kids! Don't try this at home! The angels in this figure are simply waiting for lightning to follow the graphite covered string down and fry Benjamin Franklin so they can escort him to heaven!

.

One of the premier figures in the earliest days of the study of electricity was Benjamin Franklin, an individual who can only be described as a "polymath" – physicist, inventor, publisher, politician, diplomat. Franklin conducted a series of experiments in the mid-1700's (long before the American revolution!) that determined that lightning was electrical in nature, that charging an object generally involved moving charge of a single sign (an invisible electrical "fluid") from one object that otherwise contained equal, balanced amounts of both signs of charge, to another, leaving behind a surplus of the other sign.

Figure 1.1 is an apocryphal illustration of one of Franklin's experiments with charge – flying a kite in a thunderstorm using string that had been rubbed with graphite to make it conductive to charge up a Leyden jar and demonstrate that lightning itself is a triboelectric static electrical phenomenon. Note that this is *incredibly risky* as it provides an easy path to ground for the massive charge collected on an overhead cloud, making it not at all unlikely to *attract* a lightning strike, which would of course then kill anyone holding onto the string (and quite possibly any nearby onlookers).

Franklin's discoveries were a tremendous achievement, as they set the stage for investigating electricity in the context of Newtonian mechanics. Unfortunately, he *misguessed the sign of the mobile charge*, thinking it to be the one that he named *positive*, but as it happens, mobile charge in solid conductors is almost always electrons, which are *negative*. This mistake

[33]Wikipedia: http://www.wikipedia.org/wiki/Leyden Jar. A Leyden Jar is a primitive capacitor, which we will study in more detail in three more weeks.

persists today as generations of physics students have had to draw arrows indicating the motion of *electrical current* one way, associated with the movement of (negative) *electrical charge* the opposite way. Sorry about that, but be warned, when we get to the chapters on electrical current and the Hall effect (used to *determine* the sign of mobile charge in a conductor).

In 1756 Franklin was elected as a Fellow of the Royal Society in England, which in some ways was the "heart" of the Enlightenment, and remained engaged in natural philsophy (as science was then called) for most of the rest of his life, but his energies from then on were largely diverted to politics, revolution, and ultimately, diplomacy.

Once charge was correctly identified as the "source" of electrical force, many natural philosophers of the time felt strongly that electric charge would follow the inverse square law Newton guessed and then demonstrated for the gravitational field (possibly influenced by other contemporary researchers in the late 17th century such as Robert Hooke). In the end, Charles-Augustin de Coulomb[34] , the inventor of a very sensitive ***torsional balance***, was able to use the balance and his ability to precisely divide charges to demonstrate the correctness of the inverse square law hypothesis and make electrostatics quantitative. He published his results over the period from 1785 to 1789, thirty full years after Franklin first demostrated the existence of two opposed electrical charges.

The primary way one can use charge generated by any of several simple electrostatic generators create conducting objects with at least controlled increments of charge upon them is by *induction* and *charge transfer* or *charge sharing*. We will discuss these in more detail next week after establishing the electrostatic properties of conductors.

Charge, as we shall see, is the fundamental quantity that permits objects to "couple" – affect one another – via the electromagnetic interaction. It therefore will serve us well to know some of the most important "True Facts" about charge. This initial listing is just to prime the pump, as it were – we will go over all of these ideas in much more detail, and repeatedly, later!

- Experimentally, objects can carry a ***net*** charge q when "electrified" various ways (for example by rubbing materials together).
- Almost all of the mass that makes up our everyday world consists of ***charged particles*** – the electron (-) and up or down quarks that in turn make up protons and neutrons in the atomic nucleus. Indeed, almost all of that mass is "baryonic" in those nuclei, as electrons constitute less than a thousandth of the total.
- Charge comes in two ***flavors***, **+** and **-**, but most matter is approximately charge-neutral most of the time. Consequently, as Benjamin Franklin correctly guessed, most charged objects end up that way by adding or taking away charge from this neutral base.
- **The SI unit of charge is called the *coulomb* (C).** As we shall see, a coulomb is a *lot* of charge, far more than one can usually place on or remove from a macroscopic object in the lab to do experiments on. Microcoulombs or even nanocoulombs are much more reasonable lab-scale electrostatic charges!
- "Like" charges exert a long range (action at a distance) repulsive force on one another. "Unlike" charges attract.

[34]Wikipedia: http://www.wikipedia.org/wiki/Charles-Augustin de Coulomb.

- The force varies with the inverse square of the distance between the charges and acts along a line connecting them. Coulomb's Law (covered next) describes this attraction or repulsion in extremely precise terms.

A quantity that is a constant througout all known interactions, neither created nor destroyed, is said (in physics) to be "conserved". In the first semester of this course, you learned of a number of quantities that were *conditionally* conserved – momentum or angular momentum, conserved when the net force or torque acting on a system is zero – or *unconditionally* conserved, such as net energy (or more properly, mass-energy). Our final True Fact is that:

- *Net* charge is an unconditionally conserved quantity in nature – we have *never observed an interaction that led to the creation or destruction of net charge*[35].

Later we will learn to write this conservation law mathematically in terms of the *flux of the current density*, but since we have not yet covered the mathematical tools to do this with, we will for now learn the experimental result that charge cannot be created nor destroyed; we can only move charge that already exists from one place to another[36].

1.1.1: Charge Quantization and Elementary Particles

Experimentally, we can readily see that charge can be isolated and moved around in very large to extremely small quantities. A natural question is then: Can we continue dividing charge indefinitely, and move an *infinitesimal* (in the formal sense of calculus) amount of charge? Is charge a *continuous* quantity, the way we classically imagine space and time to be? In Franklin's time it appeared so, and he spoke of at least one of the two kinds of charge as being a "fluid" that could be moved around in arbitrary amounts.

However, just as we learned in mechanics that solids and fluids are themselves *not* continuous, but rather microscopically particulate, composed of things like atoms and molecules, it turns out that atoms and molecules are in turn constructed out of quantized ***elementary*** particles, that many of these named particles are charged, and that the charge of each elementary particle is an integer multiple of an "elementary" quantum of charge. Indeed, we characterize elementary particles *by* their unique signature consisting of (among other things) their (rest) mass and their charge!

Even though this is a course in *classical* physics, we must never lose sight of the fact that somewhere down there underneath it all, a quantum universe is lurking, and sometimes it matters. In particular, it is very important for us to build an accurate *mental model* for things like the "matter" we wish to apply our theory to even as we treat most of its properties and interactions classically. In the spirit of this, let's try to understand in very simple terms – mostly

[35] At least, not so far. Good scientists always remain open minded as absence of evidence is not *certain* evidence of absence, it is at best *probable* or *practical*.

[36] Later in the study of physics you may learn of *quantum field* interactions that lead to e.g. *pair production* (or annihilation) – the simultaneous creation (destruction) of e.g. an electron-positron pair. Note well that while charges are indeed produced (destroyed) in this sort of interaction, the total charge of a produced (destroyed) pair is *zero*, justifying the careful use of the term "net" above in formulating the law. At the "everyday" energies of normal matter at normal temperatures, one pretty much can ignore this sort of thing.

Particle	Symbol	Charge	Mass-energy (m_0c^2)
Quarks			
Up quark	u	+2/3	~ 3 MeV
Up antiquark	$\bar{u}$	-2/3	~ 3 MeV
Down quark	d	-1/3	~ 6 MeV
Down antiquark	$\bar{d}$	+1/3	~ 6 MeV
Leptons			
Electron	e^-	-1	511 keV
Positron	e^+	+1	511 keV
Electron neutrino	ν_e	0	< 2 eV

Table 1: Charge and Mass of First Generation Fermions

pictures and ideas – the way everyday matter is put together, especially as it relates to the idea of charge.

There are two "kinds" of elementary particles observed in nature that form the massive building blocks of nearly everything we see, usually grouped into *families*. One family consists of the *quarks*[37] , which carry a charge that is quantized in units of $e/3$, where

$$1 \text{ e/esu/electrostatic unit} = 1.6 \times 10^{-19} \text{ C}. \tag{1.1}$$

This is a conversion factor you will need to know. Start to learn it right now!

The other family of particles is called the *leptons*[38] , which carry a charge that is quantized in units of e itself. Both of these massive particle families belong to a still larger group/kind of elementary particle called the *fermions*[39] .

All of these "normal matter" particles have so-called *antiparticles* – particles that are identical in all respects but that have the *opposite charge* of their normal matter counterparts. One of the great puzzles of modern physics is precisely why our Universe appears to be full of matter and nearly completely lacking in antimatter, but because that's the way it is we will mostly ignore antimatter in this class except as a convenient source of (oppositely) charged particles in certain problems.

Table 1 summarizes the names and charge properties of the "first generation" of the quarks and leptons. The particles exist in stable matter – the antiparticles exist in nature and can be observed in things like cosmic rays or the results of high energy particle collisions, but are much rarer and more difficult to produce or observe than the first generation particles. The table leaves out neutrinos, as they are uncharged and have a negligible mass and are hence of no particular interest to us in this course. There are further generations of both quarks and leptons but those are *also* very much beyond the limits of this course, as are the "heavy vector bosons" – massive elementary particles with integer spins, all part of something called the "Standard Model"[40] .

[37]Wikipedia: http://www.wikipedia.org/wiki/Quark.

[38]Wikipedia: http://www.wikipedia.org/wiki/Lepton.

[39]Wikipedia: http://www.wikipedia.org/wiki/Fermion. These are particles that have a "spin" angular momentum that is an *odd-half-integer multiple of* $\hbar/2$, not that it matters in this course.

[40]Wikipedia: http://www.wikipedia.org/wiki/Standard Model. Feel free to explore on your own, of course. You can start a good wiki-romp here, but of course the interesting physics now is going *beyond* the standard model.

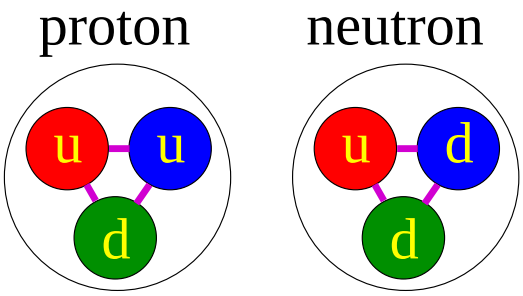


Figure 1.2: Simple model for protons and neutrons built out of three quarks. Note that the "diameter" of a proton or neutron is on the order of 10^{-15} meters, and atomic nuclei are made up of protons and neutrons that are basically "touching" and hence are of this same order in size.

Note that quarks and antiquarks come in charge units of $\pm 2e/3$ and $\pm e/3$, but we can never directly observe the thirds. In ordinary matter, these quarks are found in the *bound state* (bound together by nuclear forces we will not discuss here) into the *nucleons*: the *proton* (charge $+e$) and *neutron* (charge 0). In fact, a proton is made up of three quarks: uud, with charge $\frac{2}{3}e + \frac{2}{3}e + \frac{-1}{3}e = e$ – where the neutron is also made up of three quarks: udd, with charge $\frac{2}{3}e + \frac{-1}{3}e + \frac{-1}{3}e = 0$, as illustrated in figure 1.2 above. Experimentally, we *only see particles with a net charge quantized in units of* $\pm e$ *outside of a nucleon*, something called "confinement" in particle physics circles.

Protons have a (rest) mass around 938.3 MeV/c^2 (1.67×10^{-27} kg). This is tiny, but is still almost 2000 times larger than the mass of an electron at 0.511 MeV/c^2 (9.11×10^{-31} kg). Neutrons are just a hair more massive than a proton (939.6 MeV/c^2). Protons and neutrons are bound together by the strong interaction into an atomic nucleus on the order of 10^{-15} meters in diameter. This (positively charged) nucleus strongly attracts negatively charged electrons via the electrostatic force that is the first object of our study, which then arrange themselves around the nucleus to create a structured, electrically neutral object – the *atom*.

Since neutral atoms must contain an integer number of protons (charge +e) and an equal integer number of electrons (charge -e), we can name atoms according to the number of protons in their nucleus (the number of electrons is somewhat variable as we can comparatively easily add or take electrons away from most atoms). This is the basis of the *periodic table of the elements*, where every element is distinguished by its *atomic number*, symbol (usually) Z, which is the number of protons or electrons in the electrically neutral atom, arranged according to the chemical properties of the material, which *varies dramatically* with atomic number in families due to quantum mechanics (and, sigh, beyond the scope of this course).

Still, we need a simple mental model for atoms in order to understand some very important things in this course! I therefore encourage you to visualize atoms using one of the two pictures in figure 1.3, more the second model than the first. We will develop these models much more completely, and even semi-quantitatively, later in the course.

On the left, a heavy (carbon) nucleus – not to scale! – has its six electrons in "classical" elliptical orbits, as expected from Kepler's first law (or solution to the classical equations of motion) for inverse square force laws like Coulomb's Law and Newton's Law of Gravitation. On

Learning E&M is the first step on a long, very interesting, road!

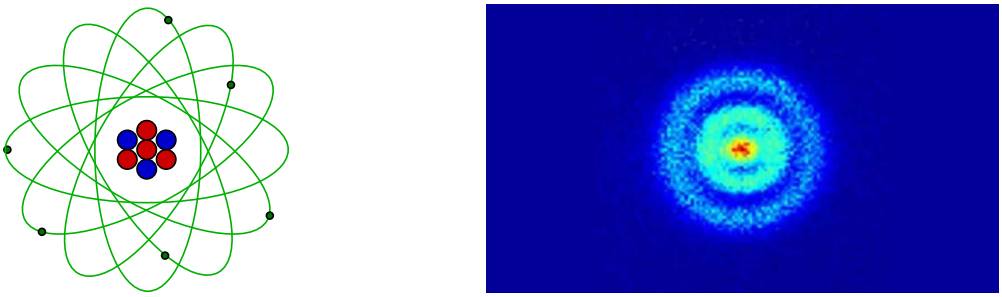


Figure 1.3: Simple mental models for “an atom”.

the right, however, an actual nanoscale “photograph” of a hydrogen atom reveals the quantum reality – an invisibly tiny, heavy nucleus surrounded by an electron smeared out in a “cloud” with colors that illustrate electron (probability) density.

The former figure is actually one of the models that led to the *death of classical physics* as it turns out that any sort of classical orbit is *inconsistent* with Maxwell’s Equations (the fundamental subject matter of this entire course) but it is still sometimes useful. It is best, however, to keep the photograph on the right in figure 1.3 – revealing a very massive and invisibly tiny nucleus surrounded more or less symmetrically surrounded by a much larger “cloud” of light, relatively mobile electrons with the same total charge – as your mental model of a neutral atom. This picture will turn out to be enormously useful to us as we seek to understand electronic properties of matter.

Finally, atoms in turn are “glued” together by electrostatic forces to form *molecules* (the object of the study of *chemistry*, so we will not dwell much on this in this text beyond noting the fact). Molecules, as it turns out, *also* tend to stick together for reasons we will explore a bit later on, and hence ***bulk ordinary matter is made up of molecules, that are in turn made up of atoms, that are in turn made up of electrons and nuclei, and the nuclei in turn are made up of protons and neutrons, which are (finally!) made up of quarks!*** Get it? As you can see from the small mass of the protons and neutrons that make up more than 99.9% of the weight of atoms , there are a *lot* of protons and neutrons – and atoms – in even micrograms of ordinary matter!

From our model we can see that nearly all the *mobile* charge in solid matter is made up of *electrons*, as the nucleus of any given atom is much more massive and is surrounded by a nearly “impenetrable” ball of negative electrical charge, locked into solids in a rigid structure in such a way that it isn’t terribly mobile. However, in *fluids* ionic charge can move around with *either* sign. In *semiconductors* the mobile charge can even be something called electron “holes” – de facto positive charge carriers consisting of regions of electron *deficit* that move against an otherwise stationary neutral electronic background, in a weird quantum sort of way.

Franklin, unfortunately, thought that the flavor of mobile charge in ordinary conductors was *positive*. In fact, as noted above, it is *negative* – associated with moving electrons. This is “Franklin’s mistake” – the bane of physics students for over two hundred years, where the *current* in a wire generally points in the *opposite direction* to the actual motion of the (negative) electrons in the wire. This will – rarely – matter in particular problems, so keep it in mind.

1.1.2: Coarse-Graining and Charge Density

With our picture of "an atom" in mind, we can proceed to figure out what happens when we consider "chunks" of matter in any of its common forms – solids, liquids or gases. First, we should note that atoms themselves – let alone the point-like elementary charges they are ultimately made up of – are quite tiny in terms of their mass and physical extent compared to the SI units describing bulk matter. This is so much so that physicists actually keep a few other sets of units in their pockets to use when doing atomic or nuclear physics! We won't do much of this now, but three important unit conversion numbers (two of them for length scales) to keep in mind are:

$$1 \text{ fermi (fm)} = 10^{-15} \text{ meters} \tag{1.2}$$

$$1 \text{ angstrom (Å)} = 10^{-10} \text{ meters} \tag{1.3}$$

$$N_A = 6 \times 10^{23} \text{ Avogadro's number} \tag{1.4}$$

A fermi is the typical length scale of the diameter of an atomic nucleus. An angstrom is the length scale of the diameter of an atom. Avogadro's number N_A is the number of atoms or molecules in one "mole" of matter – a quantity that has a mass on the order of tens of grams, centimeter sized chunks of liquids or solids. A nucleus is invisibly small indeed, relative to an atom, and an atom is invisibly small relative to "macroscopic" chunks of matter (which we will arbitrarily consider to be the smallest chunks of matter we can resolve and hence see with the naked eye through a microscope, around one micron in size), and even these smallest chunks are made up of a *lot* of atoms!

There is therefore a *very large number* of discrete charges in nearly any macroscopic piece of solid or liquid matter. We can easily estimate how much within a factor of two or three by assuming that anywhere from nearly 100% (in the case of hydrogen) to roughly 40% (in the case of Uranium) of the mass of matter consists of the *protons and neutrons* in the nuclei of the atoms that make it up. Protons and neutrons are themselves made up of *three* elementary charged particles (quarks, see above), and in neutral matter for every proton there is an electron. We can reasonably expect that close-packed solid hydrogen thus has the *fewest* charges per cubic micron, and we can estimate this number as 4 (three quarks and an electron) times $(10^4)^3$ (the volume of a cubic micron in angstroms, with roughly one hydrogen atom per cubic angstrom) or 4×10^{12} individual charged particles[41]!

The mass of such a chunk would be order of $1.67 \times 10^{-27} * \times 10^{12} \approx 1.67 \times 10^{-15}$ kg and it would contain around 10^{14} discrete charges. Our smallest visible chunk, the cubic micron of just about anything solid or liquid, will have at least 100 trillion discrete charges in it, or quite likely even more!

This makes precisely summing up fields produced by all of these charges in chunks of matter much bigger than atoms *all but impossible*, even with computers. It is also generally pointless to even try – with so many objects, surely an *average* would do for most purposes! We will therefore have frequent cause to "**coarse-grain**" our description of bulk matter – to

[41] Note that I say "roughly". Estimation is an important practice in physics! The estimates for the number of charges in this or that that I am presenting could easily be off by a factor of 10 by the time I lop off this or that smaller factor and treat it as '1', or solid hydrogen might well not be arranged with precisely one atom per cubic angstrom, but as you will see, this *will not matter* as the conclusion will still easily hold.

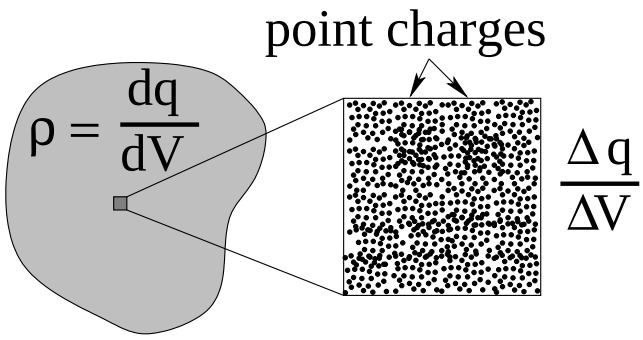


Figure 1.4: The "idea" behind coarse-graining: Any tiny block of ordinary matter – even one as small as a cubic micron in size – contains a *lot* of charges – so many that we can fairly use calculus instead of discrete summation to add it all up!

ignore the discrete particulate nature of charge and average out the *total* charge Δq in a *finite but still invisibly small* volume of matter ΔV, as illustrated in figure 1.4. By choosing ΔV **small enough that we can treat it like a volume differential but large enough that it contains a very large number of discrete charges** (of either or both signs), we can define a quasi-continuous charge *density*:

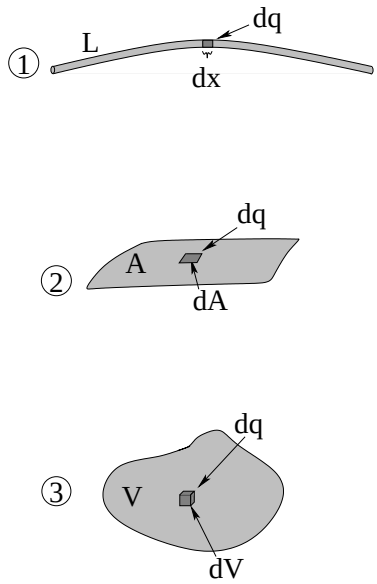
$$\rho = \lim_{\Delta V \to \text{“0”}} \frac{\Delta q}{\Delta V} \approx \frac{dq}{dV} \tag{1.5}$$

and use *calculus* to manage all of our sums!

This works because as individual charges of order e – electrons, or nuclei – move across the boundary of the "infinitesimal" micron-scaled volume ΔV, they make changes in the total charge inside the volume that only show up in some irrelevant decimal digit in the estimated total electrical charge in ΔV. After all, there are roughly 100 trillion, approximately equally balanced numbers of positive and negative charges in there! Even if we move a *million* charges across the line, that's only 0.00000000000001 of a coulomb of charge – a hundred thousandth of a *nano*coulomb of charge. Our calculus-based computation will surely work to well within any reasonable experimental accuracy.

Similarly, we can associate *surface* charge densities with "two dimensional" distributions of charge (for example, a charged piece of paper or a charged metal plate) and *linear* charge densities with thin "lines" of charged matter (for example, a wire or piece of fishing line). The calculus of all three of these distributions are illustrated in figure 1.5. In all of these forms, at the length scales associated with everyday objects, it is *indeed* better to think of charge as being the "fluid" that Franklin imagined it to be, and unimaginably difficult to consider solving problems using full sums over the trillions of trillions of discrete charges that make up the object.

Note that two of these are *further* approximations of truly three dimensional volume charge distributions. The surface charge distributions on conductors we will treat in this course aren't *truly* two-dimensional (with zero thickness); typically the unbalanced charges are confined to a few layers of atoms at the surface. However, this distance is on the order of *nanometers*, and it seems safe to ignore this thickness relative even to the side length of square micron "chunks" of the area, let alone the centimeter and meter length scales of actual charged objects. Similarly linear charges might in reality be confined to a similarly thin layer on the surface of a "thin"



$$\lambda = \frac{dq}{dx} \tag{1.6}$$

$$\sigma = \frac{dq}{dA} \tag{1.7}$$

$$\rho = \frac{dq}{dV} \tag{1.8}$$

Figure 1.5: Three charge density distributions we will use in this course – linear, surface, and volume.

wire or insulating string (perhaps one with a diameter on the order of 100 microns), but as long as the string is much thinner than it is long, we will make little error if we assume it is mathematically a one dimensional distribution.

I'll end this section with a "litany" of sorts that I advise students to use to help them solve actual problems involving continuous charge distributions. Nearly all such solutions involve using our knowledge of the forces, fields, potential energies or potentials (don't worry if you don't fully understand what these all are yet, you will soon enough) of *point* charges, summed up over all of the point(like) charges that make up the problem. To go from a discrete sum to an integral sum (the basic first step in integral calculus) you will need to remember that the charge of a coarse-grained differential "chunk" of material that has volume, area, or length as illustrated above can be obtained from:

> **The charge of the chunk is the charge per unit (volume, area, or length) of the chunk times the differential (volume, area, or length) of the chunk!**

That is:

$$dq = \rho dV \qquad dq = \sigma dA \qquad dq = \lambda d\ell$$

where the differentials are expressed in coordinates one can actually integrate over to sum up the desired result for a given continuous distribution of charge.

Say this to yourself every night before bed for a week or two. You'll be glad you did!

1.1.3: Insulators, Conductors, Semiconductors

The last property associated with the charges that make up matter that we wish to at least mention this early (although we'll examine it in more detail later) is that various materials can often be categorized, broadly speaking, into one of three types with quite distinct properties:

- **Insulators**. The charge in the atoms and molecules from which an insulating material is built tends to *not be mobile* – electrons tend to stick to their associated atoms and molecules tightly enough that *ordinary* electric fields cannot remove them (as we'll see, strong enough fields still can). Surplus charge placed on an insulator tends to remain where you put it. Vacuum is usually considered an insulator, as is air, although neither is a *perfect* insulator and even vacuum responds to and modifies electromagnetic fields[42] .

  Insulators still respond measurably to an applied field, however – the charges in the atoms or molecules distort as the molecules *polarize*, and the resulting microscopic dipoles *modify the applied field inside the material*. Since we live in air (a material) we do not generally see the *true* electric field produced by a charge but one that is very slightly reduced by the polarization of the air molecules through which the field travels. This is called *dielectric response* and we'll discuss it extensively later.

- **Conductors**. For many materials, notably metals but also ionic solutions and sufficiently hot gases (plasmas), at least one electron per atom or molecule is only *weakly* bound to its parent and can easily be pushed from one atom/molecule to the next by small electric fields. We say that these *conduction electrons* are *free* to move in response to applied field and that the material *conducts electricity*.

  Conductors also have some special properties when they respond to applied fields beyond this that we'll learn about later. Since electrons are bound to atoms by forces with a finite magnitude, *all matter* becomes a conductor in a *strong enough field!* Dielectric insulators that are placed in such a strong field experience something called *dielectric breakdown* and shift suddenly from an insulating to a conducting state. Lightning is a spectacular example of dielectric breakdown.

- **Semiconductors**. These are special "quantum" materials that can be shifted between being a conductor or an insulator depending on the potential difference at the interfaces between different "kinds" of semiconducting materials. This is an entirely quantum mechanical effect and is hence a bit beyond the classical bounds of this course, but it certainly doesn't hurt to know that semiconductors exist even in this course, as semiconductors are *extremely important* to our society. In particular, semiconductors are used in three critical ways: they are used to make diodes (one-way gates that allow electrical current to pass only in one direction, which we *will* discuss as electrical circuit elements when we talk of rectification in AM radios), as amplifiers (transistors) (used to make electronically played music and speech adjustably loud enough to listen to, for example), and as *switches* from which the digital information processing devices are built that dominate modern existence. This list is far from exhaustive – see Wikipedia: http://www.wikipedia.org/wiki/SemiconductorAn important contemporary use of semiconductors is as the basis for *solar cells*. One *very* crude way to think of at least some kinds of solar cell is as diodes where light (in the form of a photon) "kicks" an electron across the one-way semiconductor barrier. for a more complete discussion.

This concludes our discussion of charge per se for now. At this point you can see that charge is *indeed* ubiquitous! We (and everything around us) are made up of charged particles – even

[42]Wikipedia: http://www.wikipedia.org/wiki/Vaccum polarization. Beyond the scope of this course is quantum field theory, vacuum polarization, and pair production, but beyond the scope of not, in nature even an initially charge-free vacuum responds to strong electromagnetic fields.

the neutral neutrons in the nuclei that make up most of our mass are made up of charged particles! But just what force is it that binds electrons to nuclei, and in turn binds one atom to another to form a molecule? What force pushes atoms apart, so that we can set a coffee cup down on a table and not have it fall right through or get stuck there? It is time to learn about one of the most important force laws in the Universe, the one that is perhaps the *most* directly responsible for chemistry and biology: Coulomb's Law.

1.2: Coulomb's Law

If one charges various objects (for example, two conducting balls suspended from insulating strings so that they are near to one another but not touching) and measures the angular deflections of the strings when the balls are in force equilibrium, one can verify that:

- The force between the charges is proportional to each charge separately. The force is *bilinear* in the charge.
- The force acts along the line connecting the two charges.
- The force is repulsive if the charges have the same sign, attractive if they have different signs.
- The force is inversely proportional to the square of the distance between them.

These four experimental observations are summarized as ***Coulomb's Law***. They are a law of nature, on a par with Newton's Law of Gravitation (which it greatly resembles), although we will actually use an *equivalent* (and slightly more fundamental) version of this law, Gauss's Law for Electrostatics, as the version we will spend most of our time studying.

In general, while we like to understand laws like this verbally, they are more *useful* to us if we can formulate them *algebraically*, expressed in a suitable *coordinate frame*. Let's draw just such a frame in figure 1.6.

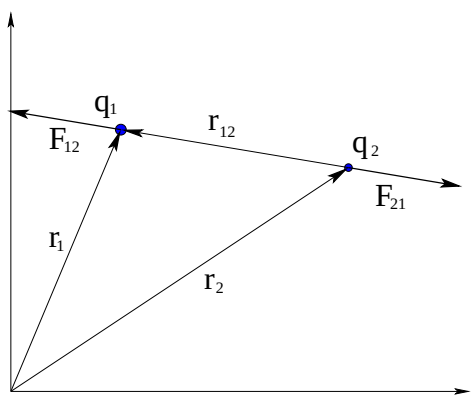


Figure 1.6: The geometry of Coulomb's Law.

We need the force acting on one of the charges, say q_1 to be bilinear in the charges (proportional to both of them as the labelling of each as 1 or 2 is arbitrary). It has to act along the

line connecting the charges, so we make a unit vector from 2 to 1 as $(\vec{r}_1 - \vec{r}_2)/|\vec{r}_1 - \vec{r}_2|$. It has to be *inversely* proportional to $|\vec{r}_1 - \vec{r}_2|^2$. Finally, we need a dimensioned constant to connect up the (SI) units on both sides of the equal sign.

If we put these things together, we get:

$$\vec{F}_{12} = k_e \frac{q_1 q_2}{|\vec{r}_1 - \vec{r}_2|^2} \times \frac{(\vec{r}_1 - \vec{r}_2)}{|\vec{r}_1 - \vec{r}_2|} = k_e q_1 q_2 \frac{(\vec{r}_1 - \vec{r}_2)}{|\vec{r}_1 - \vec{r}_2|^3} \tag{1.9}$$

as the force acting on charge q_1 at position $\vec{r}_1$ (in an arbitrary coordinate frame) due to charge q_2 at position $\vec{r}_2$. It acts on a line *from* charge 2 *to* charge 1 independent of frame (contains only relative vectors). It is proportional to both charges and satisfies Newton's *third* law. It is inversely proportional to the distance that separates them squared. It even has the benefit of being *repulsive* if both charges have the same sign and *attractive* if they have opposite signs, in either order! It is a perfect rendition of the verbal statements of the observations of Franklin and Coulomb, but now we can *compute* the force in a specific set of *coordinates* – if we have the constant of proportionality.

The constant of proportionality in SI units is given by:

$$k_e = \frac{1}{4\pi\epsilon_0} = 9.0 \times 10^9 \ \frac{\mathrm{N-m}^2}{\mathrm{C}^2} \tag{1.10}$$

(good to two significant digits, really it is $8.9875517923(14) \times 10^9$ at the current best precision but "9×10^9" is a lot easier to remember and compute with and at 0.14% accuracy is nearly always just as good in a course like this). This is a ***constant of nature***, the equivalent of "G" in the theory of gravitation, and effectively defines the "size" of the unit of charge in terms of the already known SI units of force and length.[43] It can be different numerically if we change to a different set of units, but in order for physics to be consistent it must still give exactly the same predictions for real charges in real space – in the new units.

Coulomb's Law may be simple, but it is very, very powerful – it describes the pervasive and ubiquitous force that holds the atoms and molecules of our experience (and hence *us*) together. However, it is also not in a terribly convenient form. We note that Coulomb's law (like Newton's Law of Gravitation) describes *action at a distance*. In our development above, it describes the force q_2 exerts on q_1 across the separation in space between them, but doesn't explain how q_1 "knows" where q_2 is. We rather expect that there must be something *local* to q_1, something that is there *at that point in space and time* that *encodes* this information – the magnitude of its charge and its relative location. We'd like there to be a *cause* for the observed force that is produced by q_2 at all points in space that acts *on* q_1 when we put it at some particular one. Lacking anything better to do we'll just invent the cause and call it the ***electrostatic field*** just as we similarly defined the *gravitational* field last semester.

Using fields is, as we will see, highly advantageous compared to always computing forces between *two* charges, and as we proceed, we will become more and more strongly convinced that the field itself is as "real" as anything else in physics. That is to say, it (like force itself) is a human invention, but one that is *consistent with the entire theoretical framework of physics and*

[43] Actually, a coulomb is *not* the "fundamental" unit in the SI system. The coulomb is defined to be one *ampere-second* where the *ampere* – the SI unit of electrical current – is standardized via measurements made of *magnetic* forces between fixed current carrying wires. We'll learn about this in a few weeks when we study magnetism.

in excellent quantitative agreement with observations. As in all of science, unprovable existential truth is entirely secondary to having a simple, consistent mathematical theory that *works* to accurately describe past quantitative observations and to correctly predict the outcome of future ones.

1.3: Electrostatic Field

Following this plan, then, let us propose an electrostatic field that is the supposed "local cause" of the electrostatic force between two charged objects. This means that *every* charged object in the Universe produces its own electrostatic field that emanates from the charge and presumable contributes (according to the superposition principle) to the total force any *other* charge experiences at any given point in space and time. Note well that this field is presumed to be present everywhere in space whether or not we measure it, whether or not there is a charge there to be acted on by it.

We leave until a future electrodynamics course – or at least, until later in this one – the deep discussion of whether or not this field is ***instantaneously*** linked to the position of the charge (so that moving a source charge moves its field with it everywhere, no matter how far away, synchronously) or if the field change *propagates* away from the charge at some finite speed. Empirically it is the latter (one of many ways we come to believe in the reality of our construct) but in the context of the stationary or slowly moving charges of electro***statics*** it won't really matter as the speed of propagation of the field – the speed of light, $c = 3 \times 10^8$ m/sec – is so rapid that it might as well be "instantaneous" in the laboratory scale electro*static* experiments and problems that dominate the first part of this course.

The fundamental definition of electrostatic field produced by a "source" charge q_s at position $\vec{r}_s$ is that it is the ***electrostatic force per unit charge*** acting on a small test charge q_0 placed at any given point in space $\vec{r}$ ***in the limit that the test charge vanishes***:

$$\vec{E} = \lim_{q_0 \to 0} \frac{F}{q_0} \tag{1.11}$$

or (dividing out the test charge from Coulomb's Law above):

$$\vec{E}(\vec{r}) = \lim_{q_0 \to 0} \frac{1}{q_0} k_e q_0 q_s \frac{(\vec{r} - \vec{r}_s)}{|\vec{r} - \vec{r}_s|^3} = k_e q_s \frac{(\vec{r} - \vec{r}_s)}{|\vec{r} - \vec{r}_s|^3} \tag{1.12}$$

A much more compact way to understand this results from putting the source charge at the origin of our coordinate system and (since now there is only one charge) labelling it q. In this case we get the following extremely simple way of representing the electrostatic field of a point charge q:

$$\vec{E}(\vec{r}) = \frac{k_e q}{r^2} \hat{r} \tag{1.13}$$

This result can be re-expressed in words. A point charge q produces a radially symmetric electrostatic field proportional to q that drops off like $1/r^2$ where r is the distance of the point of observation from q. It points radially *away* from positive charges and radially *in towards* negative charges (basically preserving the sign of q, in other words). If you remember this, then it is *easy* to figure out the changes needed when you have multiple charges and at least some of them are *not* at the origin.

A common question that students often ask is: "Why all of the hassle with letting point-like test charges go to zero after dividing if we're just going to divide it out anyway? Why not just start with the field of a point charge?" The reason is that – as we will see later – the presence of the test charge exerts a force in turn on the *source* distribution of charge that produced the field! If that charge distribution is not metaphorically "nailed down", if it can move *at all* in response to the test charge, it will rearrange and thereby *change* the field one is trying to measure. One is no longer measuring the field produced by (say) a charged conducting sphere, one is measuring the field produced by a charged conducting sphere in the presence of another charge that alters the charge distribution on the conducting sphere! By letting the test charge in the definition go to zero, one *formally* causes any disturbance caused by the measurement itself to go to zero, leaving you with the field that is (presumably) still there in the limit in the absence of any charges *but* those in the provided distribution.

However, those students are also correct in asking the question – we *do* start with the force and factor out and eliminate that test charge because the only way we have of *measuring* the field is via the force, and the measurement (like all measurements) is bound to alter the thing we are measuring so that we need to formally minimize this disturbance in the definition! Nevertheless, *operationally* it makes just as much sense to skip the force part of the definition entirely and go straight to the field. Since we are making up the entire idea of electrostatic field in the first place, we might as well just assert that an isolated point charge – a charged elementary particle from Table 1 above, for example – simply produces a radial electrostatic field as given in equation 1.12 above *by definition*, that the electrostatic force on any point charge is its charge times the electrostatic field at its location, and that it is *up to us* to account for any rearrangement or motion of the charges that make up the field when using this to solve problems.

As it happens, this latter plan – the one that is ***practically*** adopted by *all* physics textbooks – works quite well. The fundamental definition of the electrostatic field starts with the force between two charges and infers the field, but almost always, we'll actually *work* the other way around. In general we'll be given a distribution of charges (either discrete or a continuous charge distribution), from which we must determine the field. With the field known, we can then evaluate the force we expect these charges to exert on an *arbitrary* (e.g. test) charge placed placed in the field at an *arbitrary* point in space by means of the following rule:

$$\vec{\boldsymbol{F}} = q\vec{\boldsymbol{E}} \tag{1.14}$$

For two *point charges* and the definition of field given by 1.12 this result is *exact*. Indeed, it correctly describes the *alterations* in the force observed, when one accounts for the force-driven alterations in the positions of the source charges once the superposition principle (discussed next) is used to add up the total field! So we might as well use equation 1.12 as our definition for the electrostatic field of a point charge with equation 1.14 describing the force acting on any point charge *placed* in that field and move on from there!

So much for a single charge, but as we noted above, there are *lots* of charges in even *tiny* chunks of matter. We need a way of finding the total field produced by many charges, not just one. Furthermore, that way needs to work for charges counted "one at a time" (when there are only a few and they are enumerable) and it also needs to be useful in the limit of so many charges that a coarse-grained average yields an approximately continuous *charge distribution* in bulk matter.

Fortunately for all concerned, the electrostatic fields of many point charges simply *add right up* at any point in space! This too is a principle of nature (and is related to the linearity of the underlying equations that are the laws of nature). We call this key result the *Superposition Principle*.

1.4: The Superposition Principle

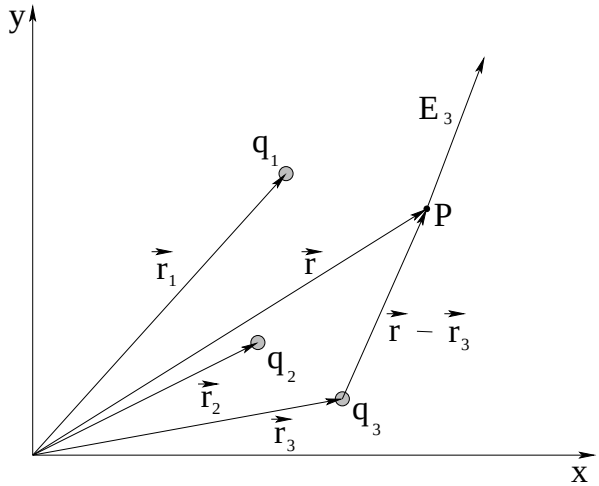


Figure 1.7: Geometry needed to evaluate the field of "many" (three) charges at an arbitrary point P at position $\vec{r}$. Only the field of the third charge $\vec{E}_3$ is shown explicitly. Note well the magnitude and direction of the vector $\vec{r} - \vec{r}_3$: ***head*** at $\vec{r}$, ***tail*** at $\vec{r}_3$. This is a vector ***from*** the position of the charge q_3 ***to*** the point of observation P at $\vec{r}$.

Given a collection of charges located at various points in space, the total electric field at a point is the sum of the electric fields of the individual charges:

$$\vec{E}(\vec{r}) = \sum_i \frac{kq_i(\vec{r} - \vec{r}_i)}{|\vec{r} - \vec{r}_i|^3} \tag{1.15}$$

This is illustrated in figure 1.7. Note that this is basically the *force* superposition rule we learned above, divided by the "test charge" q_0 in the standard definition. For fixed-position point charges, as we've seen, there can be no charge rearrangement of the sources so the limiting step $q_0 \Rightarrow 0$ is not strictly necessary, but it doesn't hurt to use it. As before, the vector field contributions from each charge carry the sign of the source charge – away from the source as in the figure for (assumed) positive charges, but if any charge is actually negative its field is directed *towards* the source charge.

Simple as it is, the superposition principle is *extremely important* in physics. It tells us that the electrostatic field results from a *linear* field theory and later in a study of physics you will learn that this means that the differential equations that describe the field produced by charge distributions are *linear* differential equations. Field theories don't *have* to be linear, but it turns out that the ones of the greatest importance in physics are[44].

[44]Mostly, anyway. In quantum mechanics things like vacuum polarization make even electrodynamics – the "poster child" of a linear theory – somewhat nonlinear at very short length scales very near a point charge.

Observe that the total $\vec{\boldsymbol{E}}$-field in equation 1.15 is a ***vector*** sum! This means that in most cases one will have to ***decompose the field produced by each charge into its vector components in the coordinate frame in question, then add the components separately, and finally reconstruct the total vector!*** It is easier to show you how that all works than tell you, so let's look at two simple examples of evaluating the total electric field produced y only *two* point charges. Both of these are very useful examples quite aside from illustrating the fairly simple math associated with summing up the field of point charges.

Example 1.4.1: Finding the Field of Two Point Charges – An 'Electric Dipole'

Two charges $\pm q$ located symmetrically on (say) the y-axis produce a field that is easy to evaluate at points on the x and y-axis. This arrangement of charges is called an *electric dipole* and is a very important concept that we will work with extensively this semester and beyond. So, suppose we have two point charges of magnitude $-q$ and $+q$, located on the y-axis at $y = -a$ and $y = +a$, respectively. We'd like to find the electric field first at an arbitrary point on the y axis, and again at an arbitrary point on the x axis.

At that point you *should* be able to find an expression for the electric field at an *arbitrary* Cartesian point (x, y) by generalizing the steps we take in these problems. Note that we are not really *ignoring* z even though the field is three dimensional and we are omitting it in the example, because we expect the $\vec{\boldsymbol{E}}$-field for this example to be *azimuthally symmetric*, that is, not to change as we rotate the solution around the y-axis, turning the solution in the x-y plane into the solution in other planes containing the y-axis.

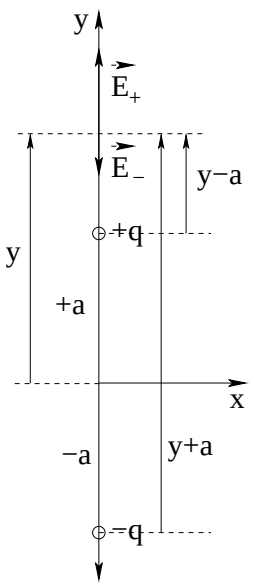


Figure 1.8: The geometry needed to solve for the field on the y-axis.

Let's start by drawing a good figure like the one on the left of the two charges on the y-axis, as well as the *arbitrary* point y on the axis where we wish to evaluate the field. Recall that the field of a point charge is:

$$\vec{\boldsymbol{E}} = \frac{k_e q}{r^2}\hat{\boldsymbol{r}}$$

The field due to the positive charge $+q$ at $y = +a$ points directly away from it in the *positive* y direction at a point $y > a$ as drawn and therefore its y-component is equal to:

$$\vec{\boldsymbol{E}}_+(0, y) = \frac{k_e q}{|y - a|^2}\hat{\boldsymbol{y}} \tag{1.16}$$

The field of the negative charge $-q$ at $y = -a$ points *towards* it (in the *negative* y-direction and hence its y-component is equal to:

$$\vec{\boldsymbol{E}}_-(0, y) = -\frac{k_e q}{|y + a|^2}\hat{\boldsymbol{y}} \tag{1.17}$$

As we might have expected just from the figure, the problem is basically ***one dimensional*** – there is only one field component to add up with no vector decomposition per se really

needed. The total field on the y axis is just:

$$\vec{\boldsymbol{E}}_{\text{tot}}(0,y) = \boxed{k_e q \left(\frac{1}{|y-a|^2} - \frac{1}{|y+a|^2} \right) \hat{\boldsymbol{y}}} \tag{1.18}$$

That was pretty easy!

The field on the x-axis is a *tiny* bit more difficult. As before, we start with a good figure defining the coordinate system and the point on the x-axis where we want to evaluate the field:

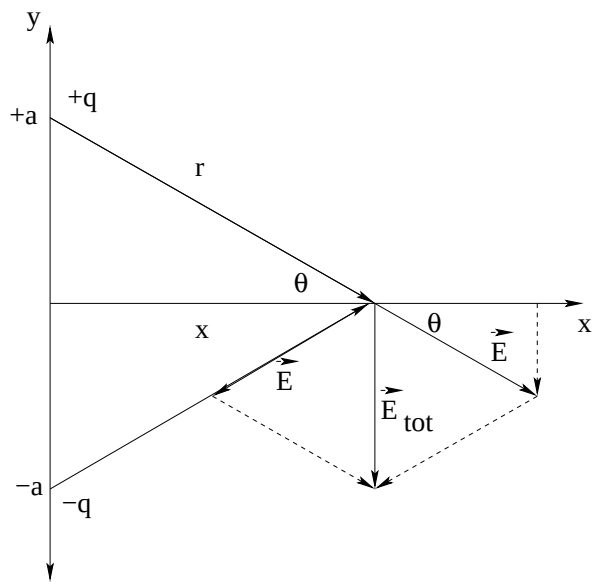


Figure 1.9: The coordinate frame and geometry needed to evaluate the $\vec{\boldsymbol{E}}$-field at an arbitrary point on the x-axis. ***Note Well*** that we defined an angle θ in the figure to aid us in decomposing the vector components of the field even though there is no 'θ' in the coordinate system!

Here the field produced by *each* charge has *two* (both x and y) components. We now have to solve the problem in three steps:

- Find the ***magnitude*** of the field of each charge, using e.g. the pythagorean theorem.
- Find the ***vector components*** of the field of each charge. This is tricky! Remember θ is *not given* so we will have to find things like $\sin\theta$ and $\cos\theta$ in terms of the givens and the coordinates of the point in question!
- Finally, we have to add up the components and reconstruct the full vector field. As always, there are multiple ways we might represent the final vector and we're not picky as long as your answer uniquely and correctly specifies that vector!

To find the vector field, we must first find the magnitude of the field. Observe that the distance from either charge to the point of observation drawn above is $r = (x^2 + a^2)^{1/2}$. Then the magnitude of the electric field vector of either charge is just:

$$|\vec{\boldsymbol{E}}| = \frac{k_e q}{r^2} = \frac{k_e q}{(x^2 + a^2)} \tag{1.19}$$

This magnitude is represented by the length of the field arrows in figure 1.9. We orient the arrows away from the upper (+) charge and toward the lower (-) charge.

Now look at the right triangle formed by x, a and r with θ drawn in one corner. By definition (think about it):

$$\cos\theta = \frac{x}{r} = \frac{x}{(x^2+a^2)^{1/2}} \tag{1.20}$$

$$\sin\theta = \frac{a}{r} = \frac{a}{(x^2+a^2)^{1/2}} \tag{1.21}$$

(where we are writing down the *positive,* quadrant 1 values, and will handle the signs needed in our final algebraic expressions from the picture). Using these, we can find the components:

$$\begin{aligned} E_x &= |\vec{\boldsymbol{E}}|\cos\theta = \frac{k_e q}{(x^2+a^2)} \cdot \frac{x}{(x^2+a^2)^{1/2}} \\ &= \frac{k_e q x}{(x^2+a^2)^{3/2}} \end{aligned} \tag{1.22}$$

and

$$\begin{aligned} E_y &= -|\vec{\boldsymbol{E}}|\sin\theta \frac{k_e q}{(x^2+a^2)} \cdot \frac{a}{(x^2+a^2)^{1/2}} \\ &= -\frac{k_e q a}{(x^2+a^2)^{3/2}} \end{aligned} \tag{1.23}$$

This is for a single charge ($+q$). The other charge has components that are the same *magnitude* but its E_x obviously *cancels* E_x from the first charge while its E_y obviously *adds* to it (doubling it). The total field is thus:

$$\vec{\boldsymbol{E}}_{\rm tot}(x,0) = \boxed{-2\frac{k_e q a}{(x^2+a^2)^{3/2}}\hat{\boldsymbol{y}}} \tag{1.24}$$

Our next topic will be the further investigation of the electric dipole. In it, we will define the *electric dipole moment* and see that the dipole moment of this arrangement of charges is:

$$\vec{\boldsymbol{p}} = 2qa\hat{\boldsymbol{y}}. \tag{1.25}$$

Thus the $\vec{\boldsymbol{E}}$-field can be expressed in terms of the magnitude of the dipole moment $p = |\vec{\boldsymbol{p}}| = 2qa$ as:

$$\vec{\boldsymbol{E}}_{\rm tot}(x,0) = -\frac{k_e p}{(x^2+a^2)^{3/2}}\hat{\boldsymbol{y}}. \tag{1.26}$$

The field on both the x and y axes seems to drop off like the (approximate) distance from the origin *cubed* in the limits where $x \gg a$ or $y \gg a$ (you can easily show this using the binomial expansion). This goes to zero *faster* than the one over r-*squared* field of an "electric monopole" (single point charge) but is still certainly not *zero* anywhere in space any more than the field of a single charge is!

At this point you should be able to evaluate the electric field vector of a y-directed dipole at an arbitrary point (x, y) in the x-y plane, not just on the x and y axes themselves! The geometry and trigonometry are just a little bit more difficult, but are still pretty straightforward if you use the pythagorean theorem and carefully draw the triangles needed to find the x and y components of the field of each charge. Now however, there will be no fortuitous cancellation of field components – the $\vec{\boldsymbol{E}}$-field will generally not point along the y axis unless you are at a

point on the y axis itself or on the x-z plane through the origin. Sadly, cartesian coordinates are not the ideal coordinate system in which to study electric dipoles, and it will turn out to be much easier to define the *electric potential* (a scalar field) of the dipole rather than the vector electrostatic field, once we know what that is, so we'll come back to dipoles again later when we've reviewed a couple more coordinate systems and learned about potential.

Before we leave, however, we do need to spend a bit of time on just what an "electric dipole" is, and why they are important to us. We'll start, then, by formally definining an electric dipole and learning how it interacts with the electric field and at least *look* at the shape of the field it produces at more points than just those on the x or y axis.

1.5: Electric Dipoles

As we just noted, the arrangement of two equal but opposite charges above is called an *electric dipole*[45] . Dipole fields play an enormously important role in physics! That is because dipolar arrangements of charge and the forces and torques that act on them and the fields that they produce are *common* in nature and play a critical role in things like, well, life! *Our* life! Let's see why.

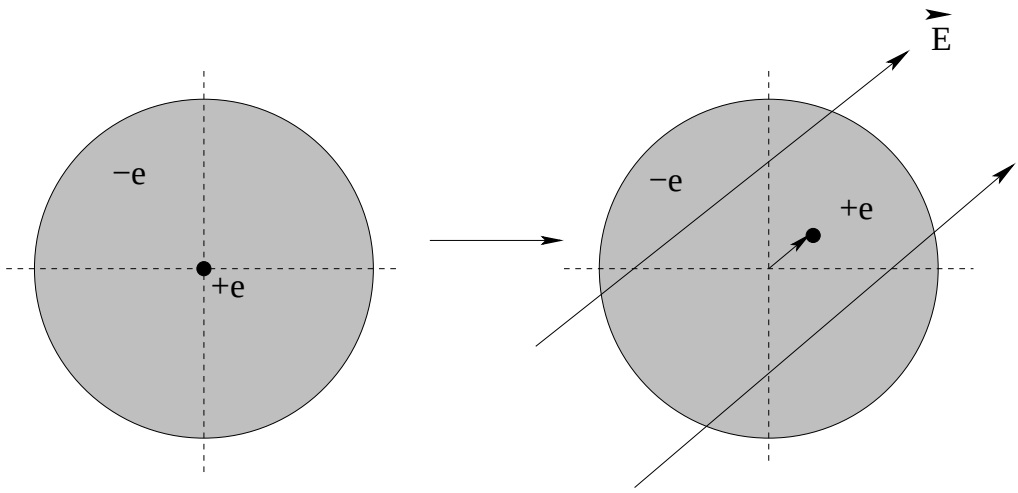


Figure 1.10: An atom in an electric field *polarizes* as its nucleus is displaced relative to its electron cloud when it is pushed one way while the cloud is pulled the other.

A simple model for an atom has a nucleus symmetrically surrounded by a spherical ball of charge in such a way that the result is electrically neutral and obviously produces no electric field outside the atom when isolated. If such an atom is placed in an electric field, however, the nucleus is pulled one way and the electron cloud is pushed the other way! While the atom remains electrically neutral (up to a point) the vector fields produced by the positive and negative charges are symmetric about different centers and *no longer precisely cancel* – they add up to make a *dipole* field!

In a few weeks we will consider the field produced by polarized atoms *on average* inside a solid as this field *modifies* the field that polarizes the atoms and we will learn some wonderful

[45]Wikipedia: http://www.wikipedia.org/wiki/dipole.

things. The model for "an atom" we develop will be very much like that illustrated in figure 1.10 above. In particular we will develop this picture into something called the *Lorentz Oscillator Model* which idealizes an atom as a uniform ball of negative charge symmetrically surrounding a small (pointlike) positively charged nucleus such that the total charge is zero. This particular atomic model predicts *harmonic oscillation* of an atomic dipole moment as well as a *linear response* model for the polarization of an atom. It actually works remarkably well all the way up to *graduate* electrodynamics classes to help students understand the general principles of dielectric polarization, electric conduction, radiation, and more!

For the moment, however, it suffices for us to recognize that since we *are* a big pile of atoms and those atoms spontaneously polarize in electrical fields (which are also ubiquitous), the forces and torques acting on dipoles, and the fields produced by dipoles, are both of great interest to us as we seek to understand ourselves and everyday "stuff" about the world around us such as why charged balloons stick to walls, why the sky is blue and the sunset is red, why matter hangs *together* even though it is generally electrically neutral. We will eventually learn that electric dipoles and dipolar forces and fields are literally everywhere, and *are very important indeed* in our efforts to build a rational worldview that explains the world of our everyday experience in simple, intuitive terms.

Let's start by modelling the resulting charge distribution of a polarized atom (or any other dipolar system) as a "basic neutral electric dipole" constructed directly out of two pointlike charges of opposite sign separated by a vector distance $\vec{l}$ ***from the negative to the positive charge***: Note that we are not, in this figure, assuming that any particular E-field (like the one

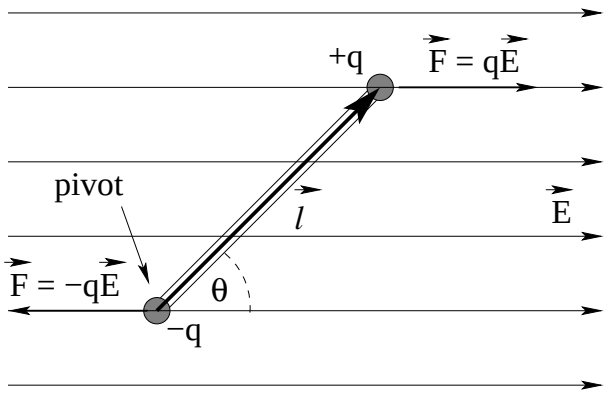


Figure 1.11: The basic dipole consists of two equal and opposite charges $\pm q$ separated by a vector displacement $\vec{l}$, in which case the *dipole moment* of the arrangement is defined to be $\vec{p} = q\vec{l}$.

drawn) or other force is *creating* the dipole; rather we are assuming that it is fixed, with the charges rigidly separated by e.g. a massless rod in between. Because we are interested in determining what the force and torque acting on the dipole are when it is located *in* a field, in figure 1.11 we have gone ahead and placed our basic dipole in a uniform field pointing to the right.

The dipole moment of the two charges is ***defined*** to be:

$$\vec{p} = q\vec{l} \tag{1.27}$$

where q is the magnitude of the charge and $\vec{l}$ is the vector that points ***from*** the negative

charge ***to*** the positive charge. As our work evaluating the electric field of a dipole in the previous section might suggest, this definition is far from arbitrary – it turns out that this is precisely the quantity that behaves somewhat "like a charge" in the equations that describe its field. To emphasize the similarity, note well that we can refer to isolated charges as electric *monopoles*, and will eventually learn to speak of monopolar and dipolar (and quadrupolar!) *moments* of arrangements or distributions of charge. In future electrodynamics courses, you will spend a considerable amount of time learning to expand electromagnetic fields in terms of the multipolar moments of source distributions, using math that is considerably more complex than the simple stuff we will use here, but *the idea will be exactly the same*, which is why it makes sense to understand the terminology and point of it all *here*, where it is still very simple!

1.5.1: Force and Torque Acting on a Dipole

We'll start by considering the *force* and the *torque* exerted by a *uniform electric field* on a dipole as illustrated in figure 1.11 above. When an electric dipole is placed in a *uniform* electrical field, the forces on the two poles are ***equal in magnitude and opposite in direction***, that is, they form what we learned to call a *force couple* in the mechanics course preceding this one.

In that course, we learned (and it is easy to see explicitly) that the net force exerted by a force couple (and hence the net force acting on a dipole in a uniform field) is *zero*. Algebraically:

$$\vec{\boldsymbol{F}} = -q\vec{\boldsymbol{E}} + q\vec{\boldsymbol{E}} = 0 \tag{1.28}$$

If the dipole is not *aligned* or *antialigned* with the uniform field, however, the force couple produced by the field clearly exerts a *pure torque* on the dipole. In particular, this torque is *independent of our choice of pivot*[46].

If we pick (say) the negative charge as the pivot, we can evaluate the torque most easily, as it is due to the force exerted on the positive charge only, at position vl relative to the pivot. The torque is therefore:

$$\begin{aligned}\vec{\boldsymbol{\tau}} &= \vec{\boldsymbol{r}} \times \vec{\boldsymbol{F}} \\ &= \vec{\boldsymbol{l}} \times q\vec{\boldsymbol{E}} \\ &= q\vec{\boldsymbol{l}} \times \vec{\boldsymbol{E}}\end{aligned}$$

or:

$$\tau = \vec{\boldsymbol{p}} \times \vec{\boldsymbol{E}} \qquad \text{with} \qquad \vec{\boldsymbol{p}} = q\vec{\boldsymbol{\ell}} \tag{1.29}$$

Note well that that charge is a *scalar* quantity so we can pull it back through the cross-product to associated it with $\vec{\boldsymbol{\ell}}$ instead of $\vec{\boldsymbol{E}}$. Also note that the "1D" directed magnitude of the torque is:

$$\tau = -pE\sin\theta \tag{1.30}$$

[46] Do not hesitate to look back at *Introductory Physics I* if necessary to review this and other aspects of torque.

where the torque points in the ***opposite direction*** to θ (e.g. into the page of figure 1.11 where θ as drawn is out of the page)[47]. This is a very important result; learn this picture and mini-derivation well so you can easily remember and apply it[48].

The two expressions above (equations 1.28 and 1.29) are only generally *exact* if $\vec{\boldsymbol{E}}$ is *uniform*. At the very least, we can see from their derivation that the field has to be the same at the two ends of the dipole so the forces form a force couple and cancel. In nature, however, many dipoles of interest have length scales that range from the sizes of nuclei (fermi, 10^{-15} meters) through the size of molecules (angstroms, 10^{-10} meters). Water or ammonia molecules, for example, are *polar* – they have permanent dipole moments.

What happens to the force and torque acting on molecules like this that are in *not quite* uniform fields, fields that vary with position. This is not an easy question, as the field can point in one direction at one point in space with some value, point in another direction at another point in space with a *different* value, and the dipole can be rotated around in this varying field!

As we'll see, as long as the dipoles in question are small *relative* to the scale over which $\vec{\boldsymbol{E}}$ varies, $\vec{\boldsymbol{\tau}} = \vec{\boldsymbol{p}} \times \vec{\boldsymbol{E}}_{\rm avg}$ will work quite well for the torque (note the use of the *average* field at the dipole location), but what about the force? Will it still be *zero*, or "small", or maybe not even all that small? To answer this question for "point-like" dipoles, let's turn to a discussion of work and energy.

As you can see from the diagram above, it requires *work* to twist the dipole around in the electric field. The work *we* do to twist it (positive or negative) or equivalently the negative of the work done by the *field* is the energy stored in the dipole as potential energy, as the electrostatic force is clearly *conservative* (it has the same form as conservative Newtonian gravitation, recall – we'll discuss this in more detail in a couple of chapters). Since forces and torques should point in the direction that *reduces* the potential energy, we expect the potential energy function to be minimum when the dipole moment aligns with the applied field.

Consider, then the amount of work done by only the component of the force perpendicular to the arc of motion as we twist the dipole above, pivoted at the negative charge, from position $\theta = \pi/2$ at right angles to the field (where we ***define the potential energy to be zero***) to the arbitrary angle θ drawn. A bit of consideration and a good picture (see homework) should convince you that:

$$\begin{aligned} U &= -\int F_t\, ds \quad (\text{or } = -\int \tau\, d\theta) \\ &= -\int_{\pi/2}^{\theta} (-qE\sin\theta)\, \ell d\theta = (q\ell)E\int_{\pi/2}^{\theta} \sin\theta\, d\theta \\ &= -pE\cos\theta + pE\cancel{\cos\frac{\pi}{2}} = -pE\cos(\theta) \end{aligned}$$

or:

$$U = -\vec{\boldsymbol{p}} \cdot \vec{\boldsymbol{E}} \tag{1.31}$$

Note that $U(\theta)$ is minimum (maximally negative) when the dipole is aligned with the field with

[47]There's bound to be a small-angle harmonic oscillator in there somewhere...

[48]Since this is the first time this semester that you have seen a *cross product*, if you have started to forget it is *also* a really good time to go back and review *that*, as well! You need to be very comfortable with its its pictorial representation, its algebra and geometry, and of course the good old right hand rule!

$\theta = 0$, maximum (maximally positive) when antialigned at $\theta = \pi$. Alignment (when both force and torque are zero and energy is minimum) is a point of *stable equilibrium* for this system!

From this we can see that from the relation between torque and potential energy in one dimension:

$$\tau = -\frac{dU}{d\theta} = -p\cos\theta\frac{dE}{d\theta} + pE\frac{d\cos\theta}{d\theta} \approx -q\ell E\sin\theta \tag{1.32}$$

will work pretty well as long as:

$$\frac{dE}{d\theta} \approx 0$$

which we expect to usually be the case over the scale of a e.g. a molecule or other point-like dipole. This justifies being able to use:

$$\vec{\boldsymbol{\tau}} = \vec{\boldsymbol{p}} \times \vec{\boldsymbol{E}}$$

as being *exact* for true "point" dipoles and an excellent approximation using $\vec{\boldsymbol{E}}_{\rm avg}$ in "slowly" varying fields where the field doesn't twist much to new angles across the length of a not-quite-point dipole so $\vec{\boldsymbol{E}}_{\rm avg}$ has a sensible direction at its location.

Sadly, from our general knowledge of intro-level mechanics, we do *not* expect that the *force* on a dipole, point-like or not, will vanish in a non-uniform field! Recall that:

$$\vec{\boldsymbol{F}} = -\vec{\boldsymbol{\nabla}}U \tag{1.33}$$

was the inverse of our definition of potential energy in mechanics! Each component of a conservative force was related to a (partial) derivative of the scalar potential energy. From this we expect that:

$$\boxed{\vec{\boldsymbol{F}} = \vec{\boldsymbol{\nabla}}(\vec{\boldsymbol{p}} \cdot \vec{\boldsymbol{E}})} \tag{1.34}$$

This formula – obviously zero if $\vec{\boldsymbol{E}}$ is uniform/constant and has no derivatives – *can* be difficult to compute – $\vec{\boldsymbol{p}}$ has no derivatives but is in a dot product with an electric field $\vec{\boldsymbol{E}}$ that does. All of this results in an "interesting" tensor form for the explicit force acting on a point dipole in a completely general field that varies with position, one well beyond the scope of this course to work out in any detail.

Still, we would like to *understand* it at some level beyond just memorizing a formula we might not know how to evaluate (at least, until after you've taken a course in multivariate calculus). Fortunately it is easy enough to work through for the *simple* case when $\vec{\boldsymbol{p}}$ and $\vec{\boldsymbol{E}}$ *point in the same direction* (say, x). A couple of homework problems will walk you through this as well, but we'll do this particular example to get you started.

Suppose we have a dipole with charge(s) $\pm q$ separated by length Δx lined up with the x-axis (so $p_x = q\Delta x$). Let $\vec{\boldsymbol{E}} = E_x\hat{\boldsymbol{x}}$ be the field at (say) the location of the negative charge and $\vec{\boldsymbol{E}}' = (E_x + \Delta E_x)\hat{\boldsymbol{x}}$ at the location of the positive charge. We can easily write an exact expression for the total force on the dipole in this case:

$$\begin{aligned}
F_x &= -qE_x + q(E_x + \Delta E_x) \\
&= \cancel{-qE_x} + \cancel{qE_x} + q\Delta E_x \\
&= (q\Delta x) \times \frac{\Delta E_x}{\Delta x} \\
&= p_x\frac{dE_x}{dx} = \frac{d}{dx}p_xE_x
\end{aligned}$$

where we took the limit $\Delta x \to 0$ implicitly in the last step, effectively making p_x into a *point* dipole. Note that this is exactly equal to:

$$F_x = -\frac{d}{dx}(-p_x E_x) = -\frac{dU}{dx} \tag{1.35}$$

in this special case, since $U = -\vec{\boldsymbol{p}} \cdot \vec{\boldsymbol{E}} = -p_x E_x$ only.

In the *general* case, we could tediously repeat this for each spatial direction and each general term in the dot product forming U, and we'd end up with the expression above (which technically has 9 distinct terms in it). The more general case, in very rough terms, results from letting $\vec{\boldsymbol{p}} = q\Delta\vec{\boldsymbol{l}}$ (a very short point-like dipole) while $\Delta\vec{\boldsymbol{E}} \approx \Delta\vec{\boldsymbol{l}} \cdot \vec{\boldsymbol{\nabla}}\vec{\boldsymbol{E}}$ is basically the first term of a 3D Taylor series expansion of $\vec{\boldsymbol{E}}$, where the gradient has to be applied to each component of the field separately.

If this still confuses you, don't worry. You'll have the opportunity to set yourself straight while exploring the idea on your own in a couple of *one-dimensional* homework problems with *explicitly given* fields where everything lines up and you can use the ordinary Taylor series or binomial expansion as demonstrated in the example to see how it goes.

1.5.2: Electric Field of a Dipole

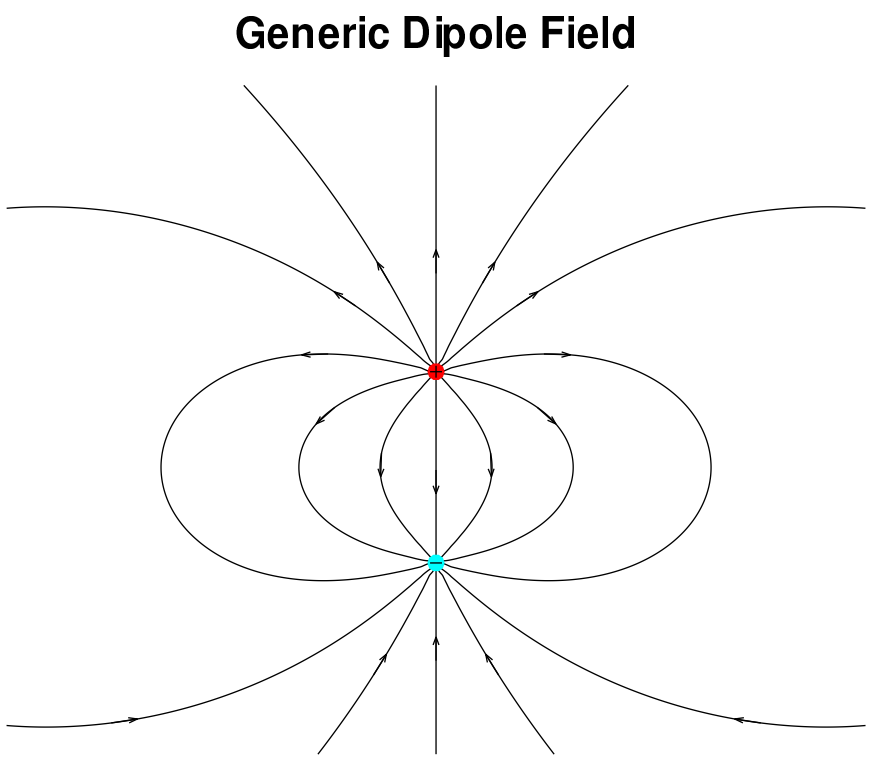


Figure 1.12: The characteristic "butterfly" electric field of an electric dipole oriented in the positive y-direction (up, in the figure) located at the origin. The field has *azimuthal* symmetry around the axis of the dipole, and the units are arbitrary.

We've already started to get a feel for the field produced by an electric dipole, in at least two possible (rather symmetric) directions. Not to worry! It is *already familiar* to any student who has done the simple school experiment of sprinkling iron filings onto a sheet of paper sitting on a small bar magnet (which as we will eventually learn is a *magnetic* dipole) as the "butterfly pattern" that emerges as the filings line up with the magnetic field. It is even possible to do the same thing with an electric dipole under a sheet of paper or plexiglass, but one has to use something like husks of rice in place of the iron filings. Both of these *experimental* results suggest that the "lines of force" associated with our 'imaginary' electric field (and later magnetic field) as an explanation or cause of electrostatic or magnetostatic forces is a bit *more* than just a metaphor, but is supported by observations of nature!

In many cases, the physical length of the dipole ($2a$ in this case worked out above and extended below) will be *small* compared to x, the distance of the point of observation to the dipole. In this limit, the field (or later, potential) produced is that of an *ideal* dipole, or a *point* dipole. The general butterfly shape of the *electric* dipolar field of a point dipole is illustrated in figure 1.12 by drawing the *lines of force* for the electrostatic field (discussed in the next chapter) that are everywhere tangent to the electric field and proceed smoothly ***from*** positive ***to*** negative charges.

We can actually *find* the dipolar field on, say, the x axis to very high accuracy in the limit that $x \gg a$. Here's a ***rubric*** for doing so, one that you should likely write down and keep handy as you do the homework for the next few weeks until you master it:

- Factoring out the larger of the two quantities (in this case x) in the denominator.
- Moving the remaining part of the denominator (the part that is now in the form $(1+z)^n$ for $|z| < 1$ and n arbitrary) to the numerator by changing the sign of n (giving it a negative exponent on top).
- Performing a *binomial expansion* on $(1+z)^{-n}$ in the numerator and keeping terms to any desired or required degree of precision.

This is easier than it originally sounds, and you'll be giving lots of practice. Note that *usually* (but ***not always***), keeping just the ***first surviving non-zero term*** is enough as that is often the only one important at long distances away from the dipole or other charge distribution.

Let's use this rubric in an example.

Example 1.5.1: Find the field of a y-directed electric dipole at an arbitrary point on the x-axis *in the limit where* $x \gg a$*!*

We start with the result we derived above for $\vec{\boldsymbol{E}}(x,0)$ on the x axis.

$$\vec{\boldsymbol{E}}_{\text{tot}}(x,0) = -\frac{k_e|\vec{\boldsymbol{p}}|}{(x^2+a^2)^{3/2}}\hat{\boldsymbol{y}}$$

Then we proceed according to the rubric:

$$\begin{aligned} E_y(x) &= -\frac{k_e|\vec{\boldsymbol{p}}|}{x^3\left(1+\left(\frac{a}{x}\right)^2\right)^{3/2}} \\ &= -\frac{k_e|\vec{\boldsymbol{p}}|}{x^3}\left(1+\left(\frac{a}{x}\right)^2\right)^{-3/2} \\ &\approx -\frac{k_e|\vec{\boldsymbol{p}}|}{x^3}\left(1-\frac{3}{2}\left(\frac{a}{x}\right)^2+\ldots\right) \end{aligned}$$

where each step is taken ***directly*** off of the rubric to end up with:

$$\boxed{E_y(x) \approx -\frac{k_e|\vec{\boldsymbol{p}}|}{x^3}\hat{\boldsymbol{y}} + \mathcal{O}\left(\frac{1}{x^5}\right)} \tag{1.36}$$

The last term in this result is read in mathematese (the language of mathematics) as: "plus neglected terms of *order* $1/x^5$" or higher).

It turns out that the field of a point dipole *generally* scales like $1/r^3$ where r is the distance from the dipole to the point of observation. It thus vanishes more rapidly than the electric monopolar moment (the field of a single bare charge, which goes like $1/r^2$) with distance, but that *does not mean the field is negligible* because the electric force is *very powerful*, far stronger than gravity, and the strongest force of nature outside of the nucleus of an atom.

Indeed, for most problems in physics that *don't* involve planet-sized masses, the electromagnetic forces – whatever form or magnitude they might have – are by far the *largest* forces acting within a system. To decide whether or not *any* algebraic expression for the field can be neglected requires specific numbers; for that reason many problems will have you find the *leading order term(s)* in a binomial or taylor series expansion of the field or potential.

Please go review both the binomial and taylor series expansions, as they will be very useful to us as we solve problems and work examples in this course. The binomial expansion in particular is a wonderful way to do "in your head" estimates of quantities that would otherwise require a calculator to evaluate.

Homework for Week 1

Before you Begin...

There are "no numbers" in most of the homework problems in this text. **This is deliberate** – algebra is a ***reasoning*** tool and physics is all about empirically founded reason! Arithmetical evaluation of formulas given numerical data on their contents, on the other hand is a process of more or less mechanical substitution and evaluation – often irreverently referred to as "plug and chug" – that can be and often is performed by entities that understand nothing at all about the origins or meaning of the formula into which numbers are being substituted or the result of the computation[49].

That is not meant to suggest that arithmetical practice is useless in physics problem, only to explain why this text de-emphasizes it. Arithmetic's primary virtue in physics *practice* problems is to permit students to get a concrete feel for reasonable/typical sizes or scales of real-work results *once a student understands those results!* A secondary virtue is that well, yeah, physics *is* supposedly a quantitatively precise theory of how everything works and one needs numbers in order to compare that theory with reality via measured experimental numbers – the basis for the lab part of a typical physics course. On both grounds, a physicist should never be completely *incompetent* at arithmetic even when done by hand, and the ability to perform quick and accurate *numerical estimates* of results has long been prized.

In each of the following chapters, most of the provided homework problems are intended for all students of physics and should *all* be completed by those students at the end of each week/chapter. They are sometimes followed by a few clearly marked "advanced" problems that are intended to be assigned primarily to physics majors, math majors taking physics, or engineering students, who are expected to know and be able to skillfully use a bit more general mathematics (especially calculus) than e.g. life science students, but note well that there is *plenty of math including calculus in the general problems* even for non-major life science students. It is impossible to learn and understand physics without at least *some* competence in calculus[50]. Newton invented calculus *just so he could formulate physics* and this course *teaches and reinforces* the use of algebra, geometry, trigonometry and calculus both to permit all of classical physics to be *consistently developed* from Newton's Laws and a handful of empirical (e.g. force) laws and to solve problems that exemplify and illuminate each new set of concepts and results as they are developed.

Please do not skimp on or skip the homework, if you are using this text to learn physics! Students who work homework problems to ***mastery*** – the state where you can do each assigned problem, perfectly, ***without using any external resource including your notes or the textbook itself*** – will almost certainly excel in the course and earn high marks as a result. Students that only work hard enough to *barely get through the homework* with the book in one hand and their lecture notes in the other or worse, don't *even* honestly complete the homework via personal struggle and effort but copy the work of others or hand it in incomplete, well, what does your *reason* tell you is a likely outcome gradewise when confronted with problems you still haven't mastered and don't really understand on a test?

[49] Such as computers, or students armed with calculators who were taught physics as a pile of formulas to be memorized instead of understood.

[50] It is the opinion of the author that so-called "algebraic physics" is taught as an empty exercise in the memorization of formulas whose origins are concealed from the student, shrouded in the mists of *calculus*...

Problem 1.

Physics Concepts

In order to solve the following physics problems for homework, you will need to have the following physics and math concepts first at hand, then in your long term memory, ready to bring to bear whenever they are needed. Every week (or day, in a summer course) there will be new ones.

To get them there efficiently, you will need to carefully organize what you learn as you go along. This organized summary will be a *standard, graded part of every homework assignment!*

Your homework will be graded in two *equal* parts. Ten points will be given for a complete crossreferenced summary of the physics concepts used in each of the assigned problems. One problem will be selected for grading in detail – usually one that well-exemplifies the material covered that week – for ten more points.

Points will be taken off for egregiously missing concepts or omitted problems in the concept summary. Don't just name the concepts; if there is an equation and/or diagram associated with the concept, put that down too. Indicate (by number) all of the homework problems where a concept was used.

This concept summary will eventually help you prioritize your study and become your own personal study guide to review for exams! To help you understand what I have in mind, I'm building you a list of the concepts for *this* week, and indicating the problems that (will) need them:

- Coulomb's Law:
$$\vec{\boldsymbol{F}}_{ij} = \frac{k_e q_i q_j (\vec{\boldsymbol{r}}_i - \vec{\boldsymbol{r}}_j)}{|\vec{\boldsymbol{r}}_i - \vec{\boldsymbol{r}}_j|^3}$$
(with $k_e = 9 \times 10^9$ N-m^2/C^2). Needed in problem(s) 4, 5, 6, 7, 8, 9, 10, 11. A core concept!

- Electric Field:
$$\vec{\boldsymbol{E}} = \lim_{q_0 \to 0} \frac{\vec{\boldsymbol{F}}_0}{q_0} = \frac{k_e q (\vec{\boldsymbol{r}}_0 - \vec{\boldsymbol{r}})}{|\vec{\boldsymbol{r}}_0 - \vec{\boldsymbol{r}}|^3}$$
or of a point charge, located at the origin:
$$\vec{\boldsymbol{E}} = \frac{k_e q}{r^2} \hat{\boldsymbol{r}}$$
Needed in nearly all of the problems.

- This definition ensures that we can find the force on a charge as follows:
$$\vec{\boldsymbol{F}} = q \vec{\boldsymbol{E}}$$
which is the version of Coulomb's Law that we will most often use in the problems – find field first, then find force if necessary. Used in nearly all of the problems in this context.

- The Superposition Principle for the Electric Field:
$$\vec{\boldsymbol{E}}(\vec{\boldsymbol{r}}) = \sum_i \frac{k_e q_i (\vec{\boldsymbol{r}} - \vec{\boldsymbol{r}}_i)}{|\vec{\boldsymbol{r}} - \vec{\boldsymbol{r}}_i|^3}$$

or, for a continuous distribution of charge:

$$\vec{\boldsymbol{E}}(\vec{\boldsymbol{r}}) = k_e \int \frac{\rho(\vec{\boldsymbol{r}}_0)(\vec{\boldsymbol{r}} - \vec{\boldsymbol{r}}_0)d^3r_0}{|\vec{\boldsymbol{r}} - \vec{\boldsymbol{r}}_0|^3}$$

One can also integrate over sheets or lines of charge, using their *charge densities*:

$$\begin{aligned} \rho &= \frac{dq}{dV} \\ \sigma &= \frac{dq}{dA} \\ \lambda &= \frac{dq}{dx} \end{aligned}$$

Needed in problems 2, 3.

- We should keep in mind that **charge is conserved**. The net charge of objects cannot change; charge can only move around, not be created or destroyed. A basic concept.
- The electric dipole moment of a pair of equal and opposite point charges of magnitude q separated by a vector $\vec{\boldsymbol{l}}$ is:

$$\vec{\boldsymbol{p}} = q\vec{\boldsymbol{l}}$$

We sometimes need the *idea* of quadrupole moments and monopole moments in this chapter. Needed in problems 2, 3, 5, 6, 9.

- The force on a dipole in a uniform electric field is:

$$\vec{\boldsymbol{F}} = 0$$

(more generally it is $\vec{\boldsymbol{F}} = -\vec{\boldsymbol{\nabla}}(-\vec{\boldsymbol{p}} \cdot \vec{\boldsymbol{E}})$. The torque on a dipole in a uniform field is:

$$\vec{\boldsymbol{\tau}} = \vec{\boldsymbol{p}} \times \vec{\boldsymbol{E}}$$

Needed in problems 2, 3, 5, 6, 9.

- Yes, we use Newton's Second Law:

$$\vec{\boldsymbol{F}} = m\vec{\boldsymbol{a}}$$

(problems 3, 4, 8 and 11); Newton's Second Law for torque:

$$\tau = I\alpha$$

(problem 9); our knowledge of the Simple Harmonic Oscillator equation and its solutions:

$$\frac{d^2x}{dt^2} + \omega^2 x = 0$$

(problems 9 and 11); and gravity near the Earth's surface:

$$\vec{\boldsymbol{F}}_g = -mg\hat{\boldsymbol{y}}$$

(down, in problems 7 and 8); and the ideas associated with stable versus unstable equilibrium in problem 3.

Our knowledge of Newton's Laws, rotation and oscillation and gravity near the earth's surface from the Mechanics part of this course is essential in this part as well!

- Two pieces of *math* that we will use repeatedly in this part of the course are the **Taylor Series Expansion of a function** in terms of its derivatives:

$$f(a+\Delta a) = f(a) + \frac{df(a)}{dx}\Delta a + \frac{1}{2!}\frac{d^2 f(a)}{dx^2}\Delta a^2 + \frac{1}{3!}\frac{d^3 f(a)}{dx^3}\Delta a^3 + ...$$

  which converges for small Δa (used in problems 3, 5, 6, 11) and the Taylor series of a particular functional form, the **Binomial Expansion**:

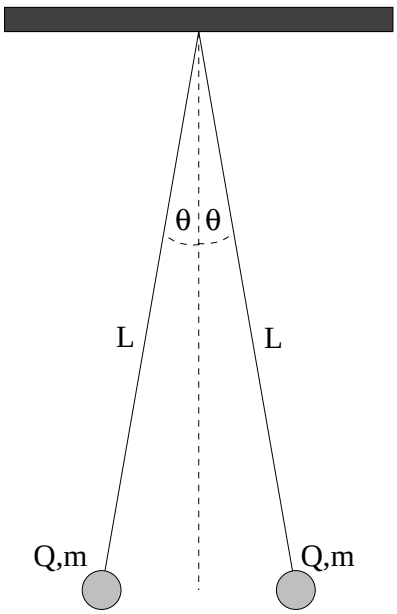
$$(1+z)^n = 1 + nz + \frac{n(n-1)}{2!}z^2 + \frac{n(n-1)(n-2)}{3!}z^3 + ...$$

  which only converges unconditionally if $|z| < 1$ (used in problems 2, 3, 5, 6, 11).

Note well the similarity between this concepts summary *needed for the homework* and the concepts summary that started the chapter. This is no accident; the chapter summary is there at the start for a reason! However, there may be additions or deletions – don't just copy the summary, and **be sure to cross-reference the problems**. The latter step is what will really help you when you are studying for a quiz or exam. What are the most important ideas, the ones you *must* know for the exam? Your concept review will (eventually) let you see at a glance...

Also, I included more concepts than are strictly needed by the problems – *don't hesitate* to add important concepts to your list (including concepts from Introductory Physics 1 in this series) even if none of the problems seem to need them! Some concepts are *ideas* and underlie problems even when they aren't actually/obviously used in an algebraic way in the solution! Remember, anything that you needed to know to solve the problems should (in the end) be in this list along with a list of the problems where it is needed.

Problem 2.



Two small spheres of mass m are suspended from a common point by threads of length L. When each sphere carries a charge Q, each thread makes an angle θ with the vertical as shown.

a) Show that the charge Q is given by:

$$Q = 2L\sin\theta\sqrt{\frac{mg\tan\theta}{k_e}}$$

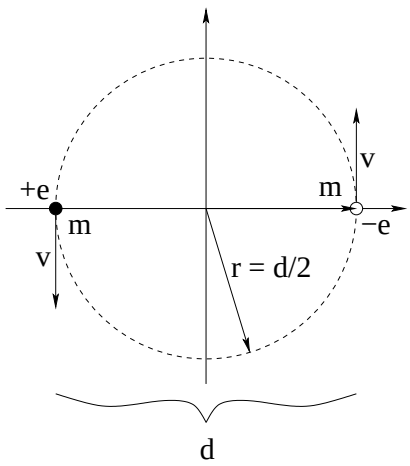
where k_e is the electrostatic constant.

b) Find Q if $m = 10$ grams, $L = 1$ m, and $\theta = 0.05$ radians. You may make the small angle approximation and can use $g \approx 10$ m/sec^2 to keep the arithmetic simple!

c) What would happen – really – if both charges Q equalled ***1 Coulomb*** instead of the tiny charge you obtained in your answer to b)?

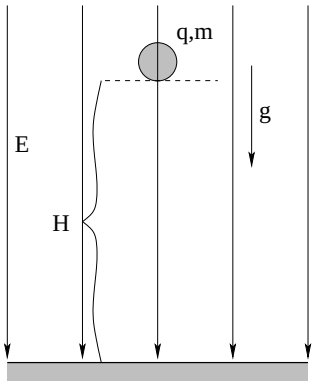
Part c) of this problem is there so you can confront just what a reasonable "size" is for isolated electric charges in the laboratory. It is much, much smaller than a Coulomb!

Problem 3.



An electron (charge $-e$, mass m) and a positron (charge $+e$, mass m) revolve around their common center of mass under the influence of their attractive coulomb force. This bound state is sometimes called Wikipedia: http://www.wikipedia.org/wiki/positronium and can actually be created for very brief periods of time in the laboratory (it is very unstable quantum mechanically as the positron and electron rapidly anihillate one another). ***Find the speed of each particle*** v in terms of e, m, k and their separation (the *diameter* of the orbit) $d = 2r$, assuming that they are both classical particles experiencing the classical electrostatic force we are studying (Coulomb's Law).

Problem 4.



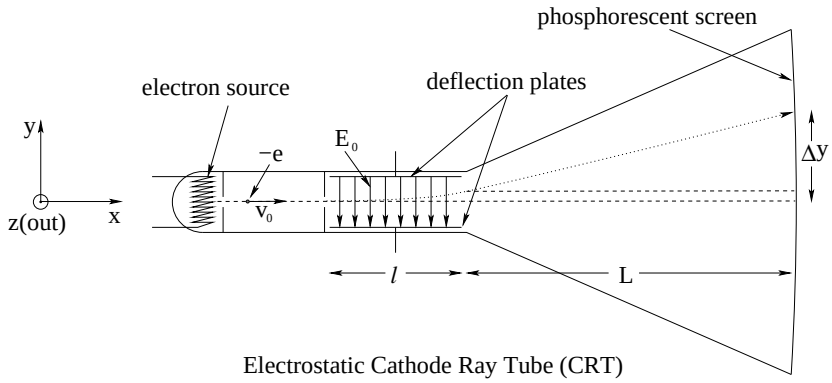
A ball of known charge q and mass m, initially at rest, falls freely from a height H in a uniform electric field of magnitude E that is directed vertically downward. It is observed that the ball hits the ground at a vertical speed $v = 2\sqrt{gH}$.

a) What is the *sign* of the charge q?

b) Find m in terms of E, q, g, and H (as needed).

c) Suppose the field was reversed to point up. With what speed would it hit the ground now (if it hits it at all)?

Problem 5.



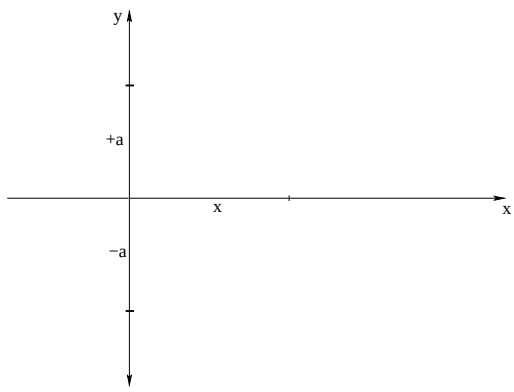
Electrostatic Cathode Ray Tube (CRT)

From the first commercial production in the mid-1930's until the year 2000, "television" and other electronic video displays were predominantly ***cathode ray tubes*** (CRTs). They were subsequently superceded by the various high resolution flat panel displays and CRT-based TVs ceased production in the US and Canada by 2010 (a good thing, since the screens contained lead, a toxic heavy metal, to prevent X-ray damage to the skin and eyes of viewers). They are, however, good examples of *physics-based engineering!*

In a (somewhat oversimpified) "electrostatic" CRT design, an electron of mass m and charge $-e$ (boiled off of a negatively charged heated "cathode") emerges from a collimating hole after a "fall" across an accerating field to move directly to the right with speed v_0 along the axis of a cathode ray tube. Assume that there is a *uniform electric field* $\vec{\boldsymbol{E}} = -E_0\hat{\boldsymbol{y}}$ in the region between the vertical ***deflection plates*** (of length l) and that everywhere else, $\vec{\boldsymbol{E}} = 0$. A nearly flat phosphorescent screen (that glows where the electron beam strikes) is a distance L from the end of the plates.

Ignoring the effect of the gravitional force on the electron as it is irrelevant for electrons travelling at such high speed, **find** $\boldsymbol{\Delta y}$, the deflection from the center point where an undeflected electron beam would hit the screen. **Hint:** You might want to break the ***trajectory problem*** up into two parts, across *l* and then across L.

Problem 6.



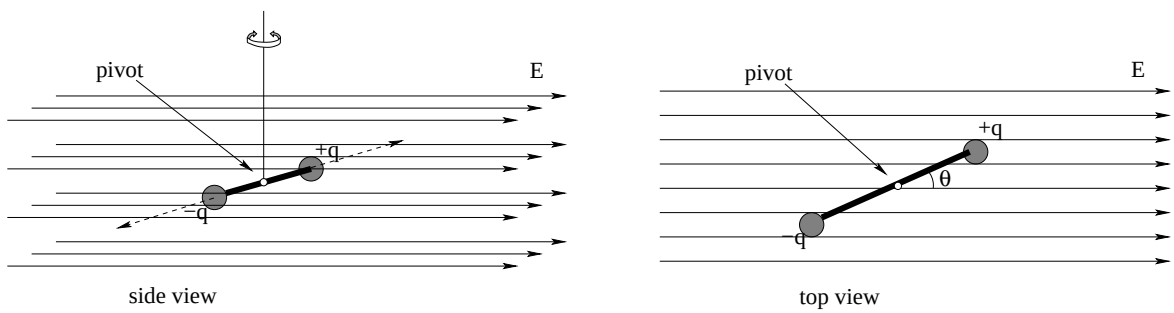
Find the electric field at an arbitrary point on the x axis in the x-y plane in cartesian coordinates for the three provided charge distributions below. Then use the ***binomial expansion*** to find its asymptotic form (the first non-zero term in the expansion) when both the case $x \ll a$ (near the origin) and $x \gg a$ (far from the origin).

a) Two equal positive charges $+q$ located at $y = -a$ and $y = +a$. Note that the (***monopolar***) far field resembles that of a *single* charge located at the origin of magnitude $2q$.

b) A positive charge $+q$ located at $y = +a$ and a negative charge $-q$ located at $y = -a$. Note that the far field dies off like $1/r^3$ instead of $1/r^2$, characteristic of an electric ***dipole***.

c) **Advanced/optional:** Two equal positive charges $+q$ located at $y = -a$ and $y = +a$, and a ***third*** charge of $-2q$ at the origin. Note that in this arrangement, the net charge is zero (so we expect no monopolar field far away). The two visible dipoles *also* cancel, so we expect no *dipolar* field far away. The leading order term in the field diminishes like $1/r^4$, characteristic of ***quadrupolar*** charge distribution.

Problem 7.

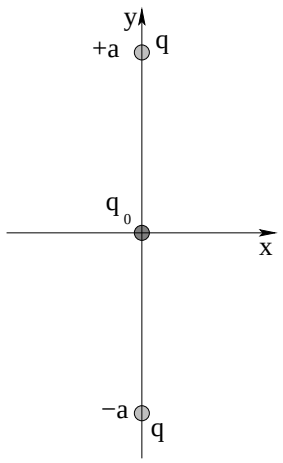


Suppose you have a "dumbbell" consisting of two identical (pointlike) masses m attached to the ends of a thin (massless) rod of length ℓ that is suspended by a string and pivoted at its center so that it can rotate freely in the horizontal plane. The masses carry a charge of $+q$ and $-q$, and the system is located in an uniform horizontal electric field $\vec{E}$ parallel to the plane of rotation.

Show that for ***small*** values of of the angle θ between the direction of the dipole and the electric field, the system displays ***simple harmonic motion***, and obtain an expression for the ***period*** of that motion. You may want to review simple harmonic motion and torsional oscillators from chapter/week 9 of:

http://www.phy.duke.edu/∼rgb/Class/intro-physics-1.php

Problem 8.



Two equal positive charges are on the y axis, one at $y = +a$ and the other at $y = -a$ as illustrated above. From symmetry, the electric field at the origin is zero. A positive charge $+q_0$ placed at the origin $(0, 0)$ will therefore have zero total force acting on it and will therefore be *in equilibrium*. But is the equilibrium *stable?* Let's find out.

a) Find an expression for the total ***force*** acting on q_0 at the position $(0, y)$ on the y-axis. Expand this force when $y \ll a$ and keep only the leading order term. Based on what you find, is the equilbrium at the origin *stable* for small displacements in the y-direction?

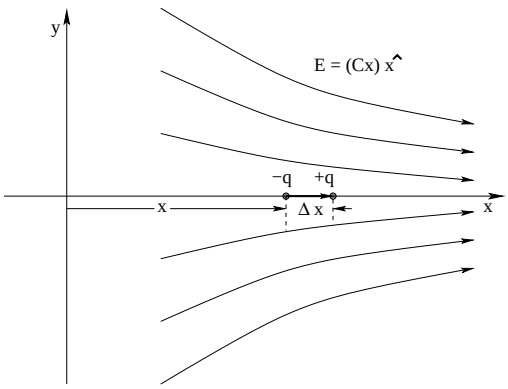
b) Find an expression for the total ***force*** acting on q_0 at the position $(x, 0)$ on the x-axis. Expand this force when $x \ll a$ and keep only the leading order term. Based on what you find, is the equilbrium at the origin *stable* for small displacements in the x-direction?

c) How would these results change if q_0 were a *negative* charge?

d) Find the ***magnitude and sign*** of a specific charge q_0 that, when placed at the origin $(0, 0)$ makes the net force on *all three of the charges at once* vanish so they are all *three* in equilibrium. In this case, what will happen if *any* of the charges are displaced slightly from equilibrium in *any* different direction (that is, is the equilibrium stable or unstable)?

This particular kind of "stable in one direction, unstable, in another" equilbrium is called a *saddle point*. We'll see *why* in a couple of chapters.

Problem 9.



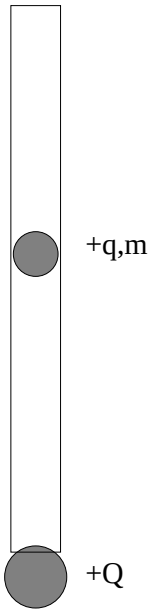
An electric dipole consists of charge $-q$ at position x and a second charge $+q$ at position $x + \Delta x$ on the x-axis as shown. This dipole is in a ***nonuniform electric field*** $\vec{E} = (Cx)\,\hat{x}$, where C is a (given) constant. Sum the total electrostatic force acting on the two charges that make up the dipole, and write the result in terms of p_x, the x-component of the dipole moment. Then, show that in the limit that $\Delta x \to 0$ your answer to b) can be written:

$$F_x = p_x \frac{dE_x}{dx} = -\frac{dU}{dx}$$

evaluated at x.

newpage

Advanced Problem 10.



A small (point) mass m, which carries a charge q, is constrained to move vertically inside a narrow, frictionless cylinder. At the bottom of the cylinder is a point mass of charge Q having the same sign as q.

a) Show that the mass m will be in equilibrium at a height:

$$y_{\text{eq}} = \sqrt{\frac{kqQ}{mg}}$$

.

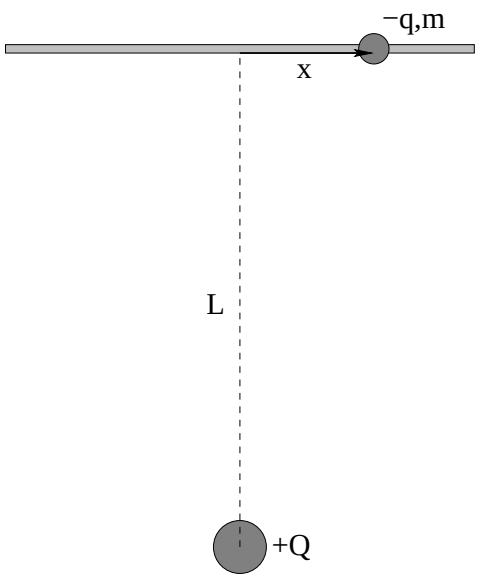
b) Show that if the mass m is displaced by a small amount Δy from its equilibrium position and released, it will exhibit simple harmonic motion with angular frequency:

$$\omega = (2g/y_{\text{eq}})^{1/2}$$

.

You will need to use e.g. the binomial or taylor expansions to solve this problem.

Advanced Problem 11.



A small bead of mass m and carrying a negative charge $-q$ is constrained to move along a long, thin, frictionless rod. A distance L from the center of this rod is a positive charge Q. Show that if the bead is displaced a distance x from the center (where $x \ll L$) and released, it will exhibit simple harmonic motion. Obtain an expression for the ***period of this motion*** in terms of the parameters L, Q, q, and m. Neglect gravity (if present and vertical, it would be cancelled by an additional normal force exerted on the bead by the wire).

You will need to use e.g. the ***binomial expansion*** to solve this problem.

Week 2: Continuous Charge and Gauss's Law

- **Continuous Charge**

  Charge distributions can often be continuous. We therefore define the following *charge densities*:

$$\begin{aligned} \rho &= \frac{dq}{dV} \\ \sigma &= \frac{dq}{dA} \\ \lambda &= \frac{dq}{dL} \end{aligned}$$

  for the charge per unit volume, per unit area, and per unit length respectively.

- **Superposition Principle**

  To find the electrostatic field produced by a continuous charge density distribution, we use the superposition principle in *integral* form:

$$\vec{\boldsymbol{E}}(\vec{\boldsymbol{r}}) = k \int \frac{\rho(\vec{\boldsymbol{r}}_0) \cdot (\vec{\boldsymbol{r}} - \vec{\boldsymbol{r}}_0) d^3 r_0}{|\vec{\boldsymbol{r}} - \vec{\boldsymbol{r}}_0|^3}$$

  where $dV_0 = d^3 r_0$ is the "volume element" – the volume of an infinitesimal chunk of the charge in the charge distribution located at $\vec{r}_0$.

  Because one has to integrate over the differential *vectors*, this integral is remarkably difficult to perform. We'll revisit it in a much simpler form when we get to electrostatic *potential*, a scalar quantity that one can usually integrate more easily without this complication.

  There are two more ways of writing this for the other two kinds of charge distribution:

$$\vec{\boldsymbol{E}}(\vec{\boldsymbol{r}}) = k \int \frac{\sigma(\vec{\boldsymbol{r}}_0) \cdot (\vec{\boldsymbol{r}} - \vec{\boldsymbol{r}}_0) d^2 r_0}{|\vec{\boldsymbol{r}} - \vec{\boldsymbol{r}}_0|^3}$$

$$\vec{\boldsymbol{E}}(\vec{\boldsymbol{r}}) = k \int \frac{\lambda(\vec{\boldsymbol{r}}_0) \cdot (\vec{\boldsymbol{r}} - \vec{\boldsymbol{r}}_0) dr_0}{|\vec{\boldsymbol{r}} - \vec{\boldsymbol{r}}_0|^3}$$

where in all cases the integral is over the entire charge distribution in question. Note that $dA_0 = d^2r_0$ and $dL_0 = dr_0$ are the "area element" and "length element" one uses in an infinitesimal chunk of the distribution in the last two expressions.

- **Gauss's Law for the Electric Field**

  Gauss's Law is written:

$$\oint_{S/V} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}}\, dA = 4\pi k \int_V \rho\, dV = \frac{Q_{\text{in S}}}{\epsilon_0}$$

  or in words, the flux of the electric field through a closed surface S equals the total charge inside S divided by ϵ_0, the permittivity of the electric field.

  Gauss's law can be used to easily evaluate the electric field for charge density distributions that have the symmetry of a coordinate system, but its real importance is that it is one of *Maxwell's Equations*, the fundamental laws of nature that govern charge and the electromagnetic field.

- **Gauss's Law and Properties of Conductors**

  One can easily use Gauss's Law to prove the following properties of conductors *in electrostatic equilibrium*. Note well that these properties *only* apply in equilibrium when no charge is actually moving.

  1 The electric field ***vanishes inside*** a conductor in ***electrostatic equilibrium*** (ESE). (It really vanishes across the first few layers of atoms, not at a mathematical surface, but we will consider changes on the scale of a few angstroms as being "instantly" and treat it as a perfect surface).

  2 There is ***no net charge in the volume of matter inside*** a conductor in ESE. This follows from **1** from Gauss's Law applied "backwards".

  3 All unbalanced charge placed or distributed on a conductor in ESE ***must reside on the surface***. This follows from **2** – if it isn't on the inside it must be on the surface.

  4 There can be ***no field component parallel to the surface of a conductor*** in ESE. In equation form, we write this:

$$\boxed{\vec{\boldsymbol{E}}_{||} = 0}$$

  The argument is identical to that used to deduce rule **1**.

- **5** Since the field at the surface of a conductor in ESE can at best be $\vec{\boldsymbol{E}}_{\perp}$ only outside and *zero* inside, if we consider an infinitesimally thin Gaussian pillbox with inner surface in the conductor and outer surface just outside, we can easily show that:

$$\boxed{\vec{\boldsymbol{E}}_{\perp} = 4\pi k_e \sigma = \frac{\sigma}{\epsilon_0}}$$

  The field at the surface is directly proportional to the surface charge density!

2.1: The Field of Continuous Charge Distributions

2.1.1: Coarse-Graining and Charge Density Revisited

In natural matter, charges are very, very small compared to the length scales we can directly perceive. An atom is order of 1 Å (10^{-10} meters) in size where a nucleus is order of 1 fermi (10^{-15} meters) in size. An electron is a pointlike particle with no physical extent at all. In a tiny piece of solid matter – one only 10^{-6} meters cubed, say – there are around $(10^4)^3 = 10^{12}$ *atoms*, and each atom is made up of 2 to 200 electric *charges* in its electron cloud and nucleus, and this is still only a chunk *one micron* in size!

Clearly, if we want to evaluate the electric field produced by a macroscopic piece of matter, we're going to have to do something other than *just sum* over the $\vec{E}_i$ fields produced by all of these charges. Instead we *average* over the amount of charge inside all of the tiny micron-scale blocks that might make up a large object. For each block there is a certain *net charge* ΔQ, in the block of size (volume) ΔV. We can use this to define the *average charge density* of the object:

$$\rho = \frac{\Delta Q}{\Delta V} \tag{2.1}$$

Now we can sum over a lot *fewer* objects. There aren't as many blocks a micron in size as there were charges, but there are *still* way, way too many blocks in an object even the size of a centimeter – 10^{12} of them, in fact – too many for us to actually sum up by hand, or even too many to sum over with a computer if we want the answer quickly. Fortunately, for typical real-world macroscopic charge distributions ρ varies only a *little* from block to block. Also, on a centimeter-plus scale, those micron sized blocks are *infinitesimal*, small enough to treat as if they are *differential* in size. We can then consider using *calculus* to do our sums. Here's how it works.

2.1.2: Using Calulus to Fine $\vec{E}(\vec{r})$ from a Charge Distribution

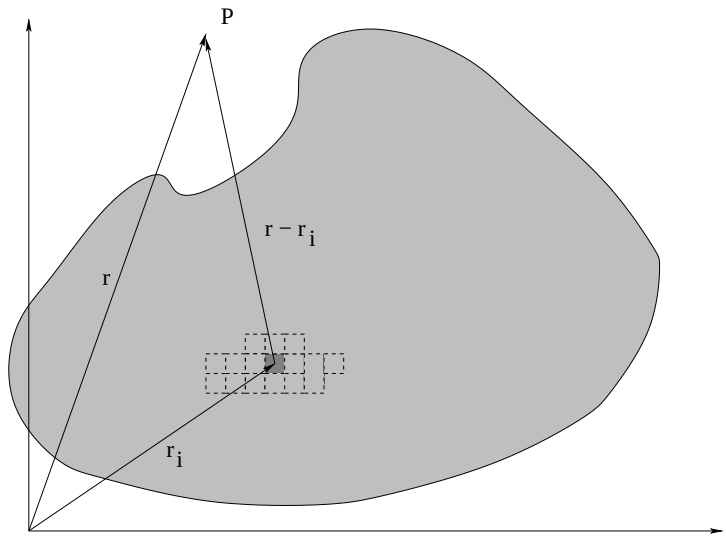


Figure 2.1: Coarse grained average leading to an integral.

Consider figure 2.1, illustrating an "arbitrary" charge density distribution. If we want to evalute the field at position $\vec{r}$ due to all of this charge, one approach might be to (mentally, at least) chop the entire amoebic blob shaped into little chunks of volume, ΔV in size (highly exaggerated in the picture so you can see them compared to their ideal scale). We've tallied up the charge in each block ΔQ, and labeled (in our minds) each block with an index i at position $\vec{r}_i$. We can then compute the field using the straight-up ***superposition principle*** at the point P (position $\vec{r}$) as:

$$\vec{E}_{\rm tot}(\vec{r}) = \sum_i \frac{k\Delta Q_i}{|\vec{r}-\vec{r}_i|^3}(\vec{r}-\vec{r}_i) \tag{2.2}$$

For many problems in physics, especially ones where the charge distribution in each block isn't any sort of convenient, integrable function, this is *exactly how physicists would evaluate the field*, only they'd use a computer to do the sums[51].

Even if we *can* use a computer, though, there can be a *lot* of blocks as noted – quite possibly too many chunks in the blob for us to sum over in a reasonable amount of time even with a computer! Often, as well, the charge density distribution in question *is* simple in form and hence likely easy enough to integrate over with ordinary calculus! In these cases, we pretend that the (coarse-grained) charge is *continuously distributed* according the prescription given above and in the first chapter:

$$\rho = \lim_{\Delta V\to 0}\frac{\Delta Q}{\Delta V} = \frac{dQ}{dV} \tag{2.3}$$

and turn the summation into an *integral* (remember both Σ_i and $\int_V$ stand for S(umming) over a collection; they are both summation symbols, the latter the one we use for *continuous* things):

$$\vec{E}_{\rm tot}(\vec{r}) = \sum_i \frac{k\Delta Q_i}{|\vec{r}-\vec{r}_i|^3}(\vec{r}-\vec{r}_i) = \int_V \frac{k\rho(\vec{r}')dV'}{|\vec{r}-\vec{r}'|^3}\vec{r}-\vec{r}') \tag{2.4}$$

where we've used $dQ = \rho dV$ (in the primed coordinates we use to replace the $\vec{r}_i$'s). This is just the *field of every little differential sized chunk that makes up the entire object, summed over all the chunks!*

This is a lot to remember, so we'll create a little mnemonic to help you. Just as we found the electric field last week by using the field of a single point charge in its simplest form and then putting it into suitable coordinates, we'll find it this week the exact same way, but the point charge in question will be dq and not q. That is:

$$\vec{E} = \frac{kq}{r^2}\hat{r} \quad \Longleftrightarrow \quad d\vec{E} = \frac{k\,dq}{r^2}\hat{r} \tag{2.5}$$

To use the latter, we just have to find dq for the particular kind of distribution, and be able to do the final integrals.

We used charge per unit volume in this discussion, but we will find that charge often distributes itself on surfaces, and we'll often need to find the field produced by lines as well. Recalling the litany given above:

> **The charge of the chunk is the charge per unit (volume, area, or length) of the chunk times the differential (volume, area, or length) of the chunk!**

[51] Well, they'd more likely evaluate the scalar *potential* of the distribution but that isn't covered until the next chapter so we'll stick to the electrostatic field for now.

we therefore expand out the charge densities we might need to handle integration over these kinds of charge distributions as:

$$\rho = \frac{dq}{dV} \quad \Longleftrightarrow \quad dq = \rho\, dV \tag{2.6}$$

$$\sigma = \frac{dq}{dA} \quad \Longleftrightarrow \quad dq = \rho\, dA \tag{2.7}$$

$$\lambda = \frac{dq}{d\ell} \quad \Longleftrightarrow \quad dq = \rho\, d\ell \tag{2.8}$$

There are then *three remaining steps* associated with setting up and solving an actual problem involving integration over a distribution:

a) Draw a picture ("like" figure 2.1), add a suitable coordinate system, identify the "right" differential chunk – one you can integrate over in those suitable coordinates – and draw in the vectors needed to express $d\vec{\boldsymbol{E}}$ as given above using the dq appropriate to the charge distribution.

b) Put down an expression for $d\vec{\boldsymbol{E}}$ (or rather, usually, $|d\vec{\boldsymbol{E}}|$) in terms of the coordinates, and find its *vector* components in terms of those same coordinates, using symmetry to eliminate unnecessary work.

c) Do the integral(s), find the field $\vec{\boldsymbol{E}}$ at the desired point.

The first two are pretty simple, and are worth most of the credit. The last will be easy enough if you've done the homework and are working hard to relearn all the calculus you need to do the integrals required in this course, which are *carefully chosen* to be ***not too difficult*** to do – believe it or not – using fairly basic calculus[52]! If your grader has a generous heart, at the beginning of this course you won't be *too* heavily penalized if you do the first two steps correctly, but this *is* the *second* semester in a calculus-based physics course and by now you really should have mastered all the math but the "new" calculus specific to E&M – line, surface, and volume integrals, especially those in cylindrical and spherical coordinates – that were only touched on in the Mechanics course.

Let's try some examples.

Example 2.1.1: Circular Loop of Charge

In figure 2.2 above we see a circular ring of charge of radius R and uniform charge per unit length:

$$\lambda = \frac{Q}{L} = \frac{Q}{2\pi R} \tag{2.9}$$

Our job is to find the electric field at an arbitrary point on the z-axis, a point with sufficient symmetry to make the evaluation fairly straightforward[53].

[52] http://www.phy.duke.edu/ rgb/Class/one-sheet-math-review.php The "One Sheet Math Review" pages here may help! These encapsulate just about all of the math you will need for this course – algebra, calculus, vectors – with everything you need to know in each topic on a single sheet. Print them out and put them in with your notes and study them from time to time as needed on problems!

[53] We *could* use the same general approach to find the field at an arbitrary point in space, but the *calculus* and *geometry* required to get an actual *answer* would become very difficult in cartesian coordinates – so difficult that

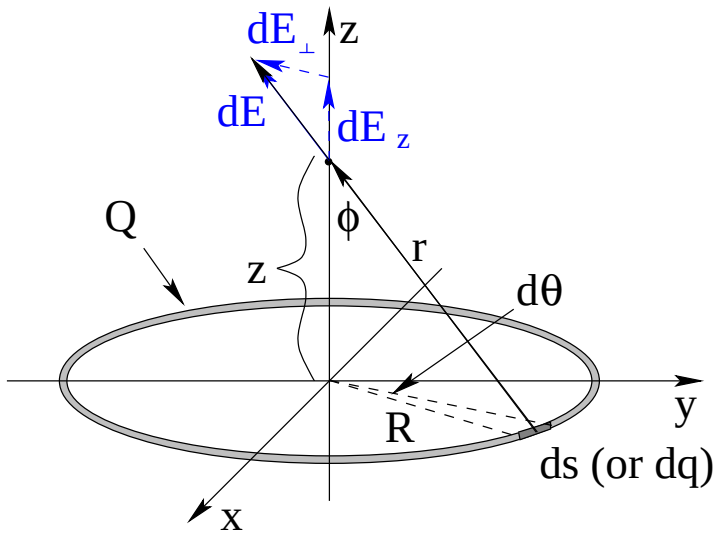


Figure 2.2: A charged ring with charge per unit length λ.

We begin by finding a small chunk of charge on the ring expressed in some coordinate we can integrate over. In this case the best possible coordinate system to use is (fairly obviously) *cylindrical* coordinates, so that we can locate a small chunk on the ring at an angle ϕ swung around in the counterclockwise direction from the positive x-axis. The angular width of the chunk is then $d\phi$, and the length of the arc subtended is $ds = R\ d\phi$.

From the previous section we recall that we need to find the charge of this little chunk of arc, repeating the litany: "the charge in the chunk is the charge per unit length, times the length of the chunk". That is:

$$dq = \lambda\ ds = \lambda R\ d\phi = Q\frac{d\phi}{2\pi} \tag{2.10}$$

where the last form is clearly the *fraction* of the total charge that lies inside the tiny subtended arc. The magnitude of the field produced by this little chunk of charge at the point z on the axis is:

$$|d\vec{\boldsymbol{E}}| = \frac{k_e dq}{r^2} = \frac{k_e \lambda a d\phi}{z^2 + R^2} \tag{2.11}$$

where we have used the pythagorean theorem to evaluate $r = \sqrt{z^2 + R^2}$ as drawn in the figure.

This vector has three components. All we need to worry about is the z-component from the *symmetry of the ring*. The field at a point on the axis cannot change as we rotate the coordinate system around the z-axis because the ring of charge looks the same as we do. Therefore it cannot have x or y components as these would *change* as we rotated the coordinate system. However, for the sake of completeness (and to give you something to figure out on the picture) I'll put down the x and y components as well:

$$dE_x = -|d\vec{\boldsymbol{E}}| \sin\theta \cos\phi \tag{2.12}$$

$$dE_y = -|d\vec{\boldsymbol{E}}| \sin\theta \sin\phi \tag{2.13}$$

$$dE_z = |d\vec{\boldsymbol{E}}| \cos\phi \tag{2.14}$$

in real life one would be very likely to concede finding an analytic solution as too difficult and resort to the use of a computer instead, or at least try it in a coordinate system more adapted to the symmetry of the ring.

In these equations, we must evaluate $\sin\phi$ and $\cos\phi$ using the right triangle Rzr:

$$\sin\phi = \frac{a}{r} = \frac{R}{(z^2+R^2)^{1/2}} \tag{2.15}$$

$$\cos\phi = \frac{z}{r} = \frac{z}{(z^2+R^2)^{1/2}} \tag{2.16}$$

so that:

$$E_z = \int_0^{2\pi} \frac{k_e \lambda z \; a d\theta}{(z^2+R^2)^{3/2}} = \frac{k_e \; \lambda(2\pi R) \; z}{(z^2+R^2)^{3/2}} = \frac{k_e Q \; z}{(z^2+R^2)^{3/2}} \tag{2.17}$$

Physicists are as lazy as they can be *when* it is possible to be lazy without compromising correctness! We will often invoke *symmetry* – as I did above to conclude that $E_x = E_y = 0$ – to avoid doing a tedious computation whose outcome we can already see and concentrate our effort on the one integral we actually cannot avoid (at least not without more practice than you've had so far in avoiding integrals:-).

You are encouraged to follow this practice yourselves, whenever you can "see" where symmetry can help out! However, you may well be confused as to exactly what the argument is that leads us to this hands-free conclusion, so just this once let's work through it in words.

The problem itself has *cylindrical symmetry*. In particular, if we mentally rotate the ring around the z-axis (or rotate the coordinate frame itself around its z-axis without moving the ring), nothing in the problem changes! If we had a nonzero, say, x component in the efield, then as we rotate the ring or coordinate, the direction of this component would *have* to change along with the ring, but since the problem itself doesn't change the solution cannot change either and the only way this is possible is if $E_x = 0$ (and ditto for E_y). E_z, on the other hand, would *not* change with this rotation even if it isn't zero, so it is allowed!

There are several other arguments that work just as well. For every chunk ds on the ring, there is an identical chunk that is its mirror image across the z axis, and the vector components of these two chunks *in* the plane of the ring cancel just like the E_x field of two point charges at $x = \pm R$ cancel (while the vertical components add).

Now, this all sounds (I hope) just great to you, and illustrates verbally some of the reasoning that flashes through a physicists mind before digging in to compute E_z only, but you *might* well still be suspicious. Words are all well and good, but math is the real language of physics – does this intuitive conclusion actually work? So again, just this once, let's do the calculus explicitly:

$$E_x = -\int_0^{2\pi} \frac{k_e \lambda a^2 \cos\theta d\theta}{(z^2+a^2)^{3/2}} = -\frac{k_e \lambda a^2}{(z^2+a^2)^{3/2}} \cdot \sin\theta \Big|_0^{2\pi} = 0 \tag{2.18}$$

(and ditto, of course, for E_y)! The reason $E_x = E_y = 0$ *mathematically* is because:

$$\int_0^{2\pi} [\sin\theta, \cos\theta] \, d\theta = 0$$

but why bother setting this up and doing it when we can just see that it must be so? Invoking symmetry *when appropriate* is thus a perfectly legitimate step in solving many of the physics problems you will enounter in this textbook and course. It will take a bit of practice to see just when that is, and you won't lose points if you get answers the hard way, but it will take *time*, time you might wish to spend some other way during an exam, so do give it a try.

Example 2.1.2: Long Straight Line of Charge

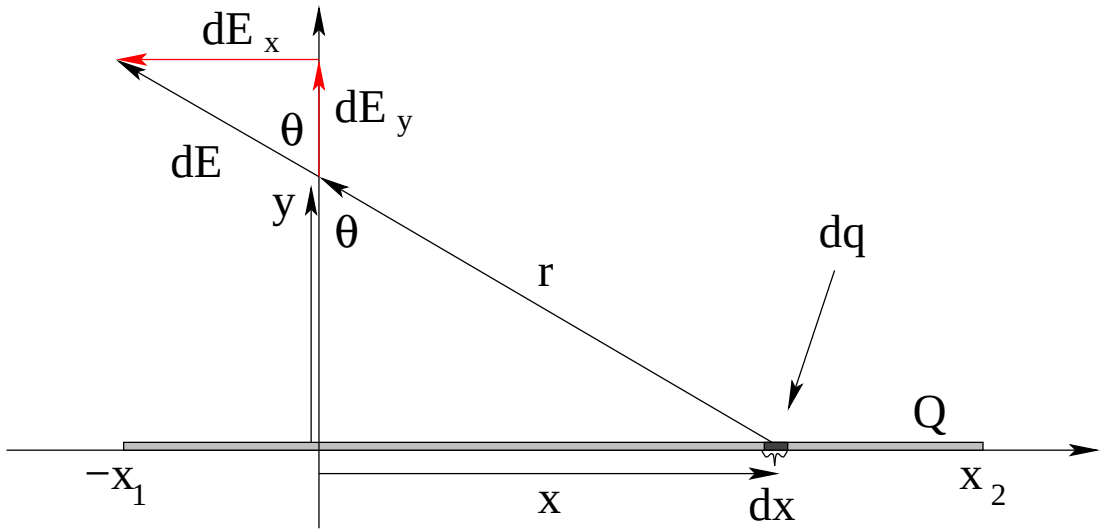


Figure 2.3: A straight line of charge with uniform charge per unit length λ.

In figure 2.3 we see a long straight line of charge. As before, we have to choose a coordinate system in terms of which to do the integral to add up the field components produced by all the little chunks of charge that make up the line.

At first glance, it seems as though cartesian components are a natural choice for the problem, so we start by using them. We want to find the field at an arbitrary point P in space, so we pick one, make the x-axis lie along the line of charge, and draw a y-axis through the x-axis such that P is the (shortest) perpendicular distance y from the x-axis. In this coordinate frame, the left hand end of the rod is x_1, right right hand is x_2. Either x_1 or $x2$ may be positive or negative depending on where P is relative to the line, and by making P and the x-y coordinate frame all lie in one plane, we have made z irrelevant (that is, expect $E_z = 0$ from good old "symmetry").

Next, we pick a chunk of charge of length dx, a distance x out from the origin directly under the point P. The charge of our chunk is *again* given by our 'magic' incantation: "The charge of the chunk is the charge per unit length of the chunk times the length of the chunk", or:

$$dq = \lambda\, dx \tag{2.19}$$

Finally, the magnitude of the field is given by:

$$|d\vec{\boldsymbol{E}}| = dE = \frac{k_e dq}{r^2} = \frac{k_e \lambda\, dx}{(x^2 + y^2)} \tag{2.20}$$

We need in this case to evaluate *both* the (red) dE_x *and* dE_y, components in figure 2.3 as E_x and E_y will in general both be nonzero (unless P happens to be in the middle of the line, in which case we expect $E_x = 0$ – from symmetry). From the triangles in the figure it is pretty obvious that:

$$dE_x = -dE \sin\theta \tag{2.21}$$

$$dE_y = dE \cos\theta \tag{2.22}$$

where we will assume that the θ we have drawn is *positive* when swung out to the right in the positive x direction, and negative when it swings out in the direction of negative x.

Noting (from the xyr right triangle) that $\cos\theta = y/r$ we get:

$$dE_y = \frac{k_e\lambda\, dx}{r^2}\cos\theta = \frac{k_e\lambda\, dx}{(x^2+y^2)}\cos\theta = k_e\lambda y\frac{dx}{(x^2+y^2)^{3/2}} \tag{2.23}$$

(for example). This, unfortunately, doesn't look terribly easy to integrate!

In fact, this is one of the most difficult integrals we have to do in this course, not because it is *particularly* difficult but because it is one of the few times we have to integrate something other than $x^n dx$, a simple trig function, or an exponential function with fairly obvious u-substitutions. The problem is that we have *too many mutually dependent variabls* – as we vary x, both r and θ vary as well and vice versa!

It turns out that this problem is easier to do if we convert it into a *trigonometric* form using nothing but y (which is fixed) and θ as our *one* independent variable. Here's how it works. Start with:

$$x = y\tan\theta \tag{2.24}$$

so

$$dx = \frac{y\, d\theta}{\cos^2\theta} \tag{2.25}$$

and

$$y = r\cos\theta \tag{2.26}$$

If we substitute equation 2.25 into the expressions for dE_x and dE_y above we get:

$$dE_y = \frac{k_e\lambda\, dx}{r^2}\cos\theta = \frac{k_e\lambda\, y d\theta}{r^2\cos^2\theta}\cos\theta = \frac{k_e\lambda\, \cancel{y}}{y^{\cancel{2}}}\cos\theta d\theta = \frac{k_e\lambda}{y}\cos\theta d\theta \tag{2.27}$$

which looks *easy* to integrate!

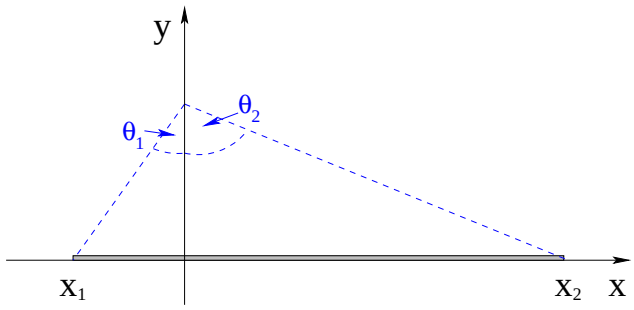


Figure 2.4: Definition of the limits of integration θ_1 and θ_2 in terms of our coordinate frame.

The limits of integration are the angles to the dotted lines that point at the ends of the line, which we will call θ_1 on the left, θ_2 on the right as indicated in figure 2.4. Using these limits:

$$E_y = \frac{k_e\lambda}{y}\int_{\theta_1}^{\theta_2}\cos\theta d\theta = \frac{k_e\lambda}{y}(\sin\theta_2 - \sin\theta_1) \tag{2.28}$$

where we should carefully note that the *specific* θ_1 in the figure above is a *negative* angle as drawn above (just as x_1 happens to be negative) and would go into this formula as a negative number in radians.

If we evaluate E_x *everything is the same* except that there is an overall minus sign and we integrate over $\sin\theta\ d\theta$ instead, to get:

$$E_x = -\frac{k_e\lambda}{y}\int_{\theta_1}^{\theta_2}\sin\theta d\theta = \frac{k_e\lambda}{y}(\cos\theta_2 - \cos\theta_1) \tag{2.29}$$

An interesting consequence of this result is that we can easily evaluate the field a distance y away from an *infinite* line of charge (that still has a uniform charge per unit length λ. In that case, $\theta_1 = -\pi/2$ and $\theta_2 = \pi/2$. We get:

$$E_x(\infty) = 0 \tag{2.30}$$

$$E_y(\infty) = \frac{2k_e\lambda}{y} \tag{2.31}$$

where we should recall that *every* point P has an x-coordinate in the middle of an infinite line of charge so that $E_x = 0$ *from symmetry* in this case! Isn't symmetry useful?

Remember this result for an infinite line of charge for later, where we will obtain it again using Gauss's Law and hence use it to *check* that Gauss's Law *works* as expected.

Example 2.1.3: Circular Disk of Charge

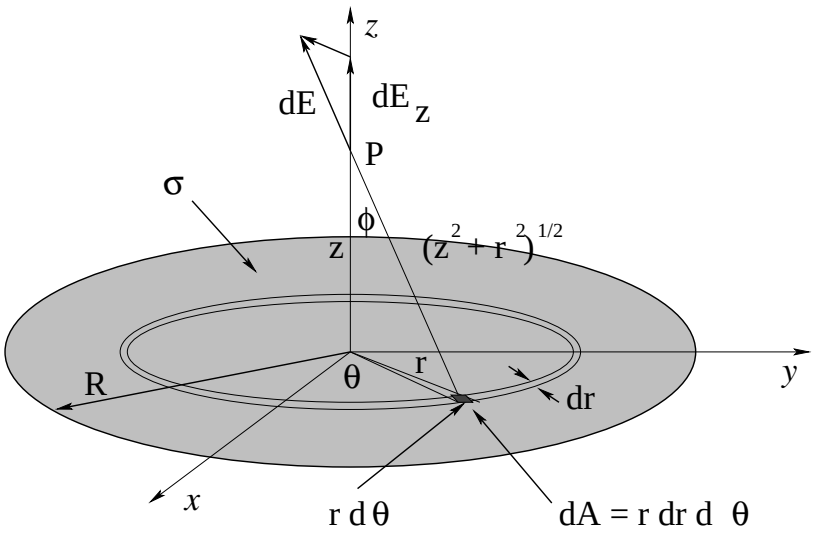


Figure 2.5: A charged disk with charge per unit area σ.

In figure 2.5 above we see a disk of charge with a uniform charge density:

$$\sigma = \frac{Q}{\pi R^2} \tag{2.32}$$

As before with a ring, we can only easily evaluate the field on the z-axis where we now *from symmetry* that $E_x = E_y = 0$. Also as before, we will proceed by finding the field of a tiny chunk of charge in suitable coordinates and sum it up using integration(s).

In order for us to be able to sum over all of the chunks of charge that make up the disk, we have to use coordinates in which integrating over the disk's *area* is *easy*. It will not be

easy at all in cartesian coordinates (try it, if you enjoy suffering)! Instead, we use *plane polar coordinates* (r, θ) for the disk itself, but keep the cartesian coordinate z to describe the point P.

We have just invented a 3D coordinate frame called ***cylindrical coordinates*** (r, θ, z). They are in all respects equivalent to the cartesian coordinates (x, y, z), and one can freely go from a description in one frame to the other if it turns out the solution is easier there. Cylindrical coordinates are often quite useful when a problem has *cylindrical symmetry* – does not change when rotating around the z (polar) axis – or when a domain we must integrate over has clean boundaries in this coordinate frame.

We actually implicitly used them to do the ring of charge example above, but in that case $r = R$ and our integral was basically both one-dimensional and trivial, so it wasn't worth pointing out. Later in the chapter (after working out Gauss's Law) we will spend some time discussing the *three* coordinate frames that are the most useful in electrodynamics problems: cartesian, cylindrical, and ***spherical polar*** coordinates.

At the moment we can get by without the full discussion if we note that the easiest way to integrate over the disk of charge in plane polar coordinates locates a point at (r, θ) inside the disk. There we swing out a small chunk of arc length $r\ d\theta$ as before for the ring, and then (mentally) *push the tiny arc out in the r-direction* by a distance dr to sweep out a tiny "rectangular" differential chunk of area dA. This area is then:

$$dA = r\, d\theta\, dr. \tag{2.33}$$

As an exercise, let's use this differential area element to find the area of the disk of radius R itself. To do this, all we have to do is integrate dA between appropriate limits that *cover* the disk *exactly once*. The advantage of using plane polar (or cylindrical) coordinates is that in these coordinates, the two-dimensional integral *separates* into two independent *one dimensional* integrals which are easy to do using our standard set of integrals (from, say, the one-sheet reviews). Here's the algebra:

$$A = \int dA = \int_0^R \int_0^{2\pi} r dr\, d\theta = \left(\int_0^R r dr\right)\left(\int_0^{2\pi} d\theta\right) = \frac{R^2}{2}(2\pi) = \pi R^2. \tag{2.34}$$

I *hope* you already know that the area of a disk is πR^2, but you may have wondered *how* we know it! Now you can see – we've explicitly evaluated the area of a disk using calculus!

This is an *important* exercise, as it shows that the integral can be grouped so that it *separates*. That is, the r integration and θ integration are *independent*. We will only do integrals over more than one coordinate in this course when they separate (in a suitable coordinate frame!), so that a student can easily work enough non-trivial problems to master physics at this level if they have mastered (a rather small subset of) *one-dimensional integration methods*. These separable problems are trivially multivariate, so to speak, and do not require that a student have taken a course in multivariate calculus to fully understand.

At any rate, we are now ready to proceed to solve our actual problem. We can easily find dq, the charge of a tiny chunk of the disk at the specific coordinates in the plane r, θ from our mantra: "The charge of the chunk is the charge per unit area times the area of the chunk", or:

$$dq = \sigma dA = \sigma\, r dr\, d\theta = \frac{Q}{\pi R^2}\, r dr\, d\theta \tag{2.35}$$

As before, we find

$$|d\vec{\boldsymbol{E}}| = dE = \frac{k_e dq}{(r^2+z^2)} = \frac{k_e \sigma\, r dr\, d\theta}{(r^2+z^2)} \tag{2.36}$$

and

$$dE_z = dE\cos\phi = \frac{k_e \sigma z\, r dr\, d\theta}{(r^2+z^2)^{3/2}} \tag{2.37}$$

(where now $\cos\phi = z/(r^2+z^2)^{\frac{1}{2}}$ from the $0rz$ right triangle in figure 2.5 above).

Finally:

$$E_z = \int_{\text{disk}} dE_z = k_e \sigma z \int_0^R \int_0^{2\pi} \frac{r dr\, d\theta}{(r^2+z^2)^{3/2}} = k_e \sigma z \left(\int_0^R \frac{r dr}{(r^2+z^2)^{3/2}}\right)\left(\int_0^{2\pi} d\theta\right) \tag{2.38}$$

Note that this integral *exactly covers the disk!* It runs from $r=0$ to $r=R$, and for *each* r it runs from $\theta = 0$ to $\theta = 2\pi$, catching the entire ring with radius r (and differential thickness dr). These integrals are *independent* since r is independent of θ and the limits of integration are fixed and do not vary with *either* coordinate.

The θ integral is trivial and yields 2π. What's left is:

$$\begin{aligned} E_z &= 2\pi k_e \sigma z \int_0^R \frac{r dr}{(r^2+z^2)^{3/2}} \\ &= \pi k_e \sigma z \int_0^R (r^2+z^2)^{-3/2}(2r dr) \\ &= -2\pi k_e \sigma z (r^2+z^2)^{-1/2}\Big|_0^R \\ &= 2\pi k_e \sigma \left(1 - \frac{z}{(R^2+z^2)^{1/2}}\right) \\ &= 2\pi k_e \sigma\,(1-\cos\Phi) \end{aligned} \tag{2.39}$$

where (as was pointed out to me by one of my many clever students) $\cos\Phi = z/\sqrt{R^2+z^2}$ where the angle Φ points from P to the edge of the disk.

There are two useful limits for us to explore for this problem. One is the limit that $R \to \infty$ (which we can also interpret as $\Phi \to \pi/2$). In this limit, the disk of charge is *infinite* in extent – it is an infinite plane of uniform charge. Since $\cos\pi/2 = 0$ the field in this limit is obviously:

$$E_z(\infty) = 2\pi k_e \sigma \tag{2.40}$$

and doesn't depend on the distance from the plane. Again, *every* point is in the middle of an infinite plane of charge, so the field of an infinite plane (or any large sheet of charge where P is close enough to the sheet so that the angles from it to the edges of the sheet are close to $\pi/2$) is uniform and has this magnitude, away from the (presumed positive) sheet of charge.

The other is when $z \gg R$. This limit is a bit tricky. We have to use the *binomial expansion*

to evaluate the field to leading order. We get (showing *every step* to guide your later practice):

$$\begin{aligned}
E_z &= 2\pi k_e \sigma \left(1 - \frac{z}{(R^2+z^2)^{1/2}}\right) \\
&= 2\pi k_e \sigma \left(1 - \frac{z}{z\left(1+\frac{R^2}{z^2}\right)^{1/2}}\right) \\
&= 2\pi k_e \sigma \left(1 - \left(1+\frac{R^2}{z^2}\right)^{-1/2}\right) \\
&\approx 2\pi k_e \sigma \left(1 - \left(1 - \frac{1}{2}\frac{R^2}{z^2} + ...\right)\right) \\
&\approx \pi k_e \sigma \left(\frac{R^2}{z^2}\right) \\
&\approx \frac{k_e(\pi R^2 \sigma)}{z^2} \\
&\approx \frac{k_e Q}{z^2}
\end{aligned} \tag{2.41}$$

or ***the field far away from the disk is the field of a point charge of the same magnitude as the disk***.

As we saw in examples done in the previous chapter, when we are far away from a charge distribution the *details* of that distribution are averaged away and we are left with a field whose leading order behavior is determined by what is called its ***multipolar moment*** – if the distribution has a net charge it is monopolar; if it has no net charge but has a $+/-$ asymmetry it is dipolar; if it has no net charge but two *balanced* dipolar charges it is quadrupolar; and so on. This means that we can often *guess* or very simply calculate what the field of a charge distribution will look like (to leading order) far away from the distribution; all we need to know (or calculate) are the total charge and/or the total separated charge and distance and direction of separation.

In future electrodynamics courses, you will learn how to express the multipolar moments of the electric and magnetic fields as specific integrals of the charge and current distributions multiplied with some very special functions that make at least *formulating* the electromagnetic field produced by those distributions simple enough – if you can do the integrals. Fortunately, with modern computers we can basically *always* do the integrals numerically, even if it would be better to give yourself a root canal with a rusty Black and Decker drill than try to do them analytically...

At this point we are *almost* finished with examples of how to use direct integration over a charge distribution to find the vector electric field. At this point you should be able to tackle all of the problems on the homework and/or the in-class problem sets (or, for that matter, in other textbooks at this introductory level). We will do *one more* (optional) highly advanced example below, *after* we cover *Gauss's Law*, where it will serve both to *directly* prove the 'shell theorem' covered with Newton's Law of Gravitation in the first semester of this course *and* to help validate Gauss's Law in application to a *nontrivial* spherically symmetric charge distribution.

2.2: Gauss's Law for the Electrostatic Field

Gauss's Law for the electrostatic field is, as we shall see, one of *Maxwell's Equations.*[54] Maxwell's equations are, in turn, the equations of motion for the unified *dynamic* electromagnetic field, considered to be 'laws of nature', and in my option at least, are one of the most *beautiful* things (mathematically and conceptually speaking) in all of physics. It is therefore of critical importance that you work hard developing a *conceptual understanding* of this law that permits you to *visualize* the relationship between the mathematics of its expression and the geometry of the field in addition to "just" learning to solve problems with it.

For that reason we will begin this chapter with a derivation of this law from the field equation of the point charge (which in turn is basically Coulomb's Law in disguise) and the superposition principle. Derivations, of course, work both ways and physicists today generally consider Gauss's Law the fundamental law of nature and the field of a point charge and Coulomb's law are rather consequences to be derived from it instead of the other way around. You will not be responsible for being able to "do" the derivation yourself in a problem or on an exam, but it is strongly advised that you work through it a couple of times anyway and get to where you intuitively understand the relationship between flux integrals and conservation, as we'll use this idea in a critical way later when we add the Maxwell Displacement Current to Ampere's Law in order to be able to show that light is an electromagnetic wave!

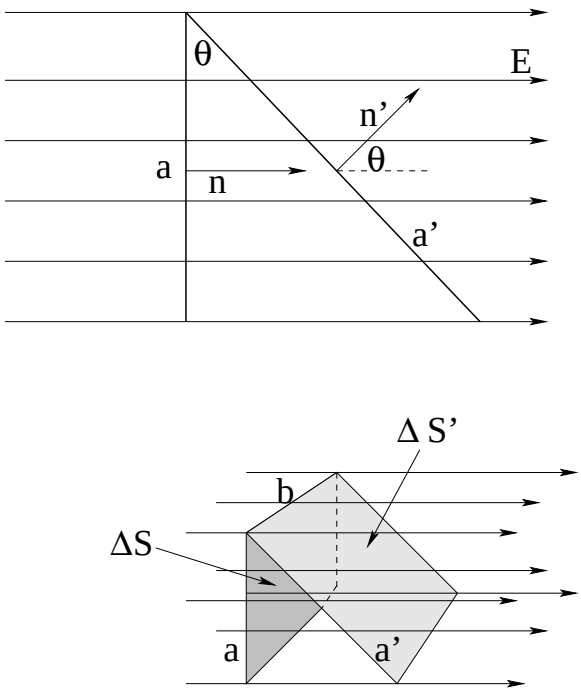


Figure 2.6: Geometry of the flux integral over a small surface area

We begin our derivation of Gauss's Law by considering the *flux* of the electrostatic vector field through a small rectangular patch of surface ΔS. To compute this, we first must understand what the flux of an arbitrary vector field $\vec{F}$ through a surface S *is*. Mathematically, the

[54]Wikipedia: http://www.wikipedia.org/wiki/Maxwell's Equations.

flux of a vector field through some surface is defined to be:

$$\phi_f = \int_{\Delta S} \vec{\boldsymbol{F}} \cdot \hat{n}\, dS \tag{2.42}$$

Note that the word flux means *flow*, and this integral measures the *flow* of the field *through* the surface. It's mathematical purpose is to detect the *conservation of flow* in the vector field. Basically it takes the magnitude of the field $\vec{\boldsymbol{F}}$ at all points on the surface, computes the component of $\vec{\boldsymbol{F}}$ that goes *through* the surface at right angles (instead of tangent to the surface, which doesn't really go "through"), multiplies it times a tiny differential chunk of the area, and then adds up all the differential chunks thus computed.

Let's look at this in more detail, specializing to the case of the electric field. Consider figure 2.6, where we show electric field lines flowing through a small $\Delta S = ab$ at right angles to the field lines (so that a unit vector $\hat{\boldsymbol{n}}$ normal to the surface is *parallel* to the electric field). ΔS is small enough that the continuous field is approximately uniform across it (we will eventually make it differentially small, of course, so this is no problem).

Since the field is uniform and at right angles to the field, the flux through just this little chunk is easy to evaluate. It is just:

$$\Delta\phi_e = |\vec{\boldsymbol{E}}|\Delta S = |\vec{\boldsymbol{E}}|ab \tag{2.43}$$

That was easy enough! Let's make things a little more complicated.

Suppose that we consider a rectangular surface $\Delta S' = a'b$ that is *tipped* with respect to the first surface at an angle θ, that shares the length b of the first surface, and that has a length a' that is long enough that it precisely subtends the same "stream" of the vector field $\vec{\boldsymbol{E}}$ as shown. Basically, all the field lines that pass through the first surface pass through the second surface, and again we are assuming that the field is continuous and we can make the picture as small as we like (differentially small in the limit) so that a conserved $\vec{\boldsymbol{E}}$ doesn't change its *magnitude or direction* in between the two surfaces.

Note that $a = a'\cos(\theta)$, so that:

$$\Delta S' = a'b = \frac{ab}{\cos(\theta)} \tag{2.44}$$

If we just multiply $|\vec{\boldsymbol{E}}|$ by $\Delta S'$, we see that we'll get $\Delta\phi_e' = \Delta\phi_e/\cos(\theta)$, right? And we'd like to get the same thing, as we'd like the flux integral to *measure* the continuity and conservation of the electric field across the tiny region between the two surfaces. So we multiply by $\cos(\theta)$ on top to compensate and get:

$$\begin{aligned}
\Delta\phi_e' &= |\vec{\boldsymbol{E}}|\cos(\theta)a'b \\
&= |\vec{\boldsymbol{E}}|\cos(\theta)\frac{ab}{\cos(\theta)} \\
&= |\vec{\boldsymbol{E}}|ab \\
&= \Delta\phi_e
\end{aligned} \tag{2.45}$$

We can interpret this as meaning (in words) "If $\vec{\boldsymbol{E}}$ is a continuous, constant vector field in the region between ΔS and $\Delta S'$, then $\Delta\phi_e' = \Delta\phi_e$ and the flux through the two surfaces is conserved."

Note that $|\vec{\boldsymbol{E}}| = \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}}$ and $|\vec{\boldsymbol{E}}| \cos(\theta) = \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}}'$, so that we can write:

$$\lim_{\Delta S \to 0} \Delta\phi_e = \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} \Delta S = \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}}' \Delta S' \tag{2.46}$$

which *does not vary* for any possible tipping of a surface element ΔS originally perpendicular to the field lines as long as the tipped surface $\Delta S'$ stretches to cover the exact same field lines as illustrated above.

This is all very specific to the case where the field is *uniform* and points in a single direction, but in fact it is easy to see that the result is more general than that.

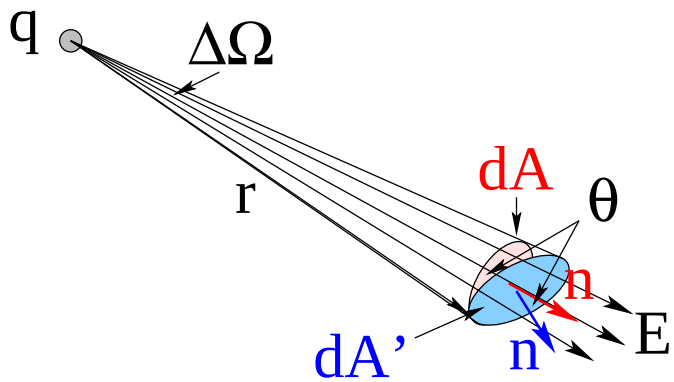


Figure 2.7: A spherical surface area dA subtended by the cone with solid angle $d\Omega$ with a charge q at the apex, a distance r away, and its outward directed unit vector normal $\hat{\boldsymbol{n}} = \hat{\boldsymbol{r}}$ (in red), and an area at the same distance r *tipped* so it is *still* subtended by the same solid angle, but now has an angle θ between its unit normal $\hat{\boldsymbol{n}}'$ (in blue) and $\hat{\boldsymbol{r}}$ (the direction of the outflowing electric field).

Suppose that we consider a single a point charge, producing the usual, symmetric electric field:

$$\vec{\boldsymbol{E}} = \frac{k_e q}{r^2} \hat{\boldsymbol{r}} \tag{2.47}$$

Some of that field streams out in the narrow *solid cone* with apex on the charge q pictured in figure 2.7 – this is basically a cone with a "two-dimensional angle" – called a ***solid angle*** – in the apex. At the radius r, this cone chops off a chunk of a surrounding (closed) spherical surface with area:

$$dA = r^2 d\Omega \tag{2.48}$$

As one can see in the section discussing spherical polar coordinates, one can actually formulate $d\Omega$ in coordinates and integrate it over the entire solid angle around the charge such that (for fixed r):

$$A = \oint dA = r^2 \oint d\Omega = r^2 \int_0^{2\pi} d\phi \int_0^{\pi} \sin\theta d\theta = 4\pi r^2 \tag{2.49}$$

Thus the solid angle of 4π exactly "covers the sphere" a single time the area of a sphere of radius r is $4\pi r^2$.

If one tips up this surface so that it is still at the distance r from the charge (within differential scales) and still is subtended by the *same* conical solid angle so that the *same* field lines flow through both, its area *still* increases to $dA' = \dfrac{r^2 d\Omega}{\cos\theta}$ (because all of the lengths on this tipped

surface along the tipping angle increase precisely as they do in figure 2.6 above, while the ones at right angles to this are unchanged). On a differential scale, then, this is *still* precisely compensated by taking a dot product with $\hat{\boldsymbol{r}}$, the direction of the electric field.

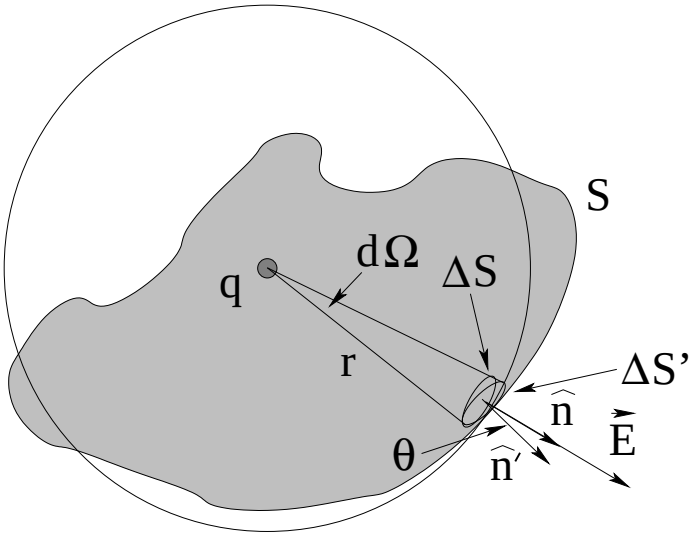


Figure 2.8: Point charge inside a closed surface S' *and* inside the closed *spherical* surface S of radius r that osculates S' at the patch dA. We have just shown that the electric field flux $d\phi_e$ through dA is equal to the flux through the tipped dA'.

Now let's think about what this means if we *deform* a closed spherical surface S surrounding the charge q into an *arbitrary* closed surface S' that still completely contains the charge as illustrated in figure 2.8. From the arguments given above, the flux of the electric field from the point charge through the *tipped* differential patch dA' that osculates (kisses) dA at one end but is tipped up through an angle θ so it is actually a part of the blob shaped *arbitrary* closed surface S', the flux through the two patches is the same:

$$d\phi_e' = \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}}' dA' = |\vec{\boldsymbol{E}}| \cos\theta \frac{r^2 d\Omega}{\cos\theta} = |\vec{\boldsymbol{E}}|\, r^2 d\Omega = \Delta\phi_e \tag{2.50}$$

In the differential limit, then, we can compute the flux through a small chunk of the arbitrary surface S' as:

$$\begin{aligned} d\phi_e &= \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA' \\ &= |\vec{\boldsymbol{E}}|\, r^2 d\Omega \\ &= \frac{k_e q}{r^2} r^2 d\Omega \\ &= k_e q\, d\Omega \end{aligned} \tag{2.51}$$

which is *independent* of the shape of S' and involves only the differential solid angle swept out from the charge as one does the integral. The point is that the r^2 in the differential area precisely cancels the r^2 in the electric field, while the $\cos\theta$ in the dot product precisely compensates for the increased area of the tipped differential patch that still is subtended by the (same) solid angle $d\Omega$

This result no longer depends on anything but the solid angle! to compute the *total* flux through the closed surface S *or* S', then, we just have to integrate the complete solid

angle surrounding the point charge! We already did this above – the result in spherical polar coordinates was:

$$\int d\Omega = \int_0^{\pi} \int_0^{2\pi} \sin(\theta) d\theta \; d\phi = 4\pi \quad \text{“steradians”} \tag{2.52}$$

(where steradians is the name of the dimensionless solid angle “coordinate” just as radians is the name of he usual planar angle “coordinate”). Thus we let S itself be *any* closed surface (since it doesn't matter if it is spherical and concentric or neither to get:

$$\phi_e = \oint_S \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} \; dS = 4\pi k_e \; q \tag{2.53}$$

independent of the shape of the closed surface S that we integrate over that encloses the charge q!

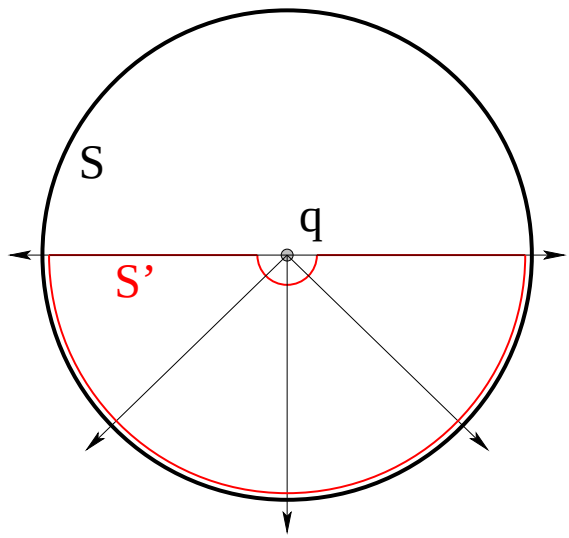


Figure 2.9: In this figure, closed surface S is a sphere concentric with the charge q, while the red closed surface S' consists of two hemispheric surfaces and a flat plane bisecting S and does *not* contain the charge q.

But what about closed surfaces with *no* point charge inside? It is easy enough to see that if the charge q is *outside* the closed surface S', the net flux through S' is *zero*. Consider figure 2.9. The lower half of the sphere S subtends 2π steradians (half the total solid angle of the sphere), so exactly *half* of the flux from charge q, $2\pi k_e q$, emerges from it. I deliberately drew S' so that all of the electric flux that *enters* the closed (red) surface S' on the inner red hemispheric surface also *exits* through the outer red hemispheric surface. There is no contribution to the flux through S' from the bisecting plane part of S' as the field is parallel to the surface. Hence:

$$\phi_{S'} = \phi_{\text{inner}} + \phi_{\text{outer}} + \phi_{\text{plane}} = -2\pi k_e q + 2\pi k_e q + 0 = 0 \tag{2.54}$$

It might look like this result is a special case, but really it is quite general. Anytime the field enters on one side of a closed surface but exits on the other, the contribution of the flux through that part of the surface cancels because the two sides subtend the same solid angle with one side going in and the other out (hence having opposite signs in the flux integral) for a net contribution of zero. Closed surfaces don't even have to form a simply connected domain (that's mathspeak for “they can have tunnels through them”, the topology of e.g. a donut-shaped surface or torus) – they just have to have a surface that splits space into two pieces

such that the only way to go from the inside to the outside is *through a surface*, not *around* it the way one can go around a piece of paper to get from one side to the other without going through.

To get our final result, Gauss's Law itself, all that remains is using the *superposition principle*. If we enclose more than one charge in S:

$$\oint_S \vec{\boldsymbol{E}}_{\rm tot} \cdot \hat{\boldsymbol{n}}\, dA = \oint_S (\vec{\boldsymbol{E}}_1 + \vec{\boldsymbol{E}}_2 + ...) \cdot \hat{\boldsymbol{n}}\, dA = 4\pi k_e (q_1 + q_2 + ... = 4\pi k_e Q_{\rm tot}) \tag{2.55}$$

because integration is a *linear* operation – the integral of a sum is the sum of the integrals. Clearly this result doesn't depend on the charges being "point charges", as we can follow the usual ritual and coarse grain the sum to convert it to an integral so it applies to continuous charge distributions as well as usual. When we do so, we (finally!) arrive at:

★ Gauss's Law for the Electric Field ★

$$\boxed{\oint_{S/V} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}}\, dA = 4\pi k_e \int_V \rho\, dV = \frac{Q_{\rm in\ S}}{\epsilon_0}} \tag{2.56}$$

in integral form. In this expression, we have (re)introduced a new quantity called ***the permittivity of free space***, ϵ_0, seen but not remarked upon in equation 1.10. It is trivially related to the Coulomb constant we have been using exclusively to find the electric field or force up to now:

$$k_e = \frac{1}{4\pi\epsilon_0} \qquad \Leftrightarrow \qquad \epsilon_0 = \frac{1}{4\pi k_e} = 8.85 \times 10^{-12} \quad \frac{\rm Coulomb^2}{\rm Newton\text{-}meters^2} \tag{2.57}$$

The reason we have ignored it up to now is that it is more difficult to remember than k_e, and we *don't need it yet* to help us understand, for example, electric fields in matter and polarizability. However, I will start using it occasionally in algebraic expressions of Gauss's Law just so you can get used to it and learn where it goes gradually without having to work very hard at it, so it is there when we need it. At that time we will also learn more useful ways of expressing its SI units!

In words, Gauss's Law for Electrostatics is:

> The flux of the electric field through a closed surface S equals the total charge inside S divided by ϵ_0 (the permittivity in a vacuum of the electric field).

This is our first of four ***Maxwell's Equations*** that we will cover, in considerable detail, this semester. Those four equations, plus perhaps a few definitions of things like force in terms of field, (almost) completely determine ***electrodynamics*** – the study of the unified electromagnetic field! Because we will use it often all semester long, I will usually abbreviate it instead of write it out as "GLE".

Note well that I used integration to expressed the total charge of a continuous distribution in the boxed expression, but of course Gauss's Law is equally well the discrete sum immediately preceding it that I got from directly considering superposition and the linearity of integrals. The main reason to write GLE in *just* this way is that if we (again) apply the multivariate calculus result known as the *divergence theorem*, we can convert the integral form into a *partial differential equation* that is *also* GLE:

$$\vec{\boldsymbol{\nabla}} \cdot \vec{\boldsymbol{E}} = \frac{\rho}{\epsilon_0} \tag{2.58}$$

So, what's GLE good for? Lots! But for the moment, we'll *start* but using GLE to easily evaluate the electric field for charge density distributions that have the symmetry of a coordinate system that we'd otherwise have to evaluate using painful direct integration. We will also use it to help us reason about things like the distribution of charge on a conductor in electrostatic equilibrium. And don't forget, we consider *it* to be the actual Law of Nature for the electrostatic field, so things like the field of a point charge and Coulomb's Law and so on are actually *consequences* of Gauss's Law (or consistently equivalent to Gauss's Law) rather than the other way around. So basically, everything else we do with the electrostatic field this semester will be a "use" of Gauss's Law.

2.3: Using Gauss's Law to Evaluate the Electric Field

One of the first and most important applications of Gauss's law for our current purposes will be to easily evaluate the electric field for certain symmetric charge distributions that we'd otherwise have to integrate over, painfully. There are precisely *three* symmetries we can manage in this way:

- point (spherical symmetry)
- infinite line (cylindrical symmetry)
- infinite plane (planar symmetry)

That's it! No more. For charge distributions that are spherically symmetric, cylindrically symmetric, or planarly symmetric, we can do the flux integral in Gauss's law *once and for all* for the symmetry. As we'll see, all that remains for us to be able to easily obtain the field from algebra is for us to evaluate the total charge inside a Gaussian surface for any given symmetric distribution. Here's the recipe:

a) Draw a closed *Gaussian Surface* that has the symmetry of the charge distribution. The various pieces that make up the closed surface should *either* be *perpendicular to the field* (which should also be constant on those pieces) or *parallel to the field* (which may then vary but which produces no flux through the surface).

b) Evaluate the flux through this surface. The flux integral will have exactly the same form for every problem with each given symmetry, so we will do this once and for all for each surface type and be done with it!

c) Compute the *total charge inside this surface*. This is the only part of the solution that is "work", or that might be different from problem to problem. Sometimes it will be easy, adding it up on fingers and toes. Sometimes it will be fairly easy, multiplying a constant charge per unit volume times a volume to obtain the charge, say. At worst it will be a problem in integration if the associated density of charge is a function of position.

d) Set the (once and for all) flux integral equal to the (computed per problem) charge inside the surface and solve for $|\vec{E}|$. That's all there is to it!

Now, you don't want to be *memorizing* these steps, you want to be *learning* them, so please use *exactly these steps* and *show all of your work doing them* in *every homework problem* that requires using them. If you use them five or six times in a row, in slightly different contexts, it will get quite easy! At the very least, even if you get a problem where you can't "do" (say) an integral to find the charge inside a given surface, you'll get most of the credit for laying out the precisely correct method except for an integral you can't quite do.

Note Well: You *cannot* use Gauss's Law to e.g. evaluate the field of a ring of charge, or a disk over charge, or a line segment of charge or any other continuous distribution that does not have the symmetry of sphere, infinite cylinder, or infinite plane. Sorry, that's just the way it is. It isn't that it isn't true for these distributions, it is that we cannot compute the flux integral. Let's do some examples, at least one for each symmetry.

Example 2.3.1: Spherical: A spherical shell of charge

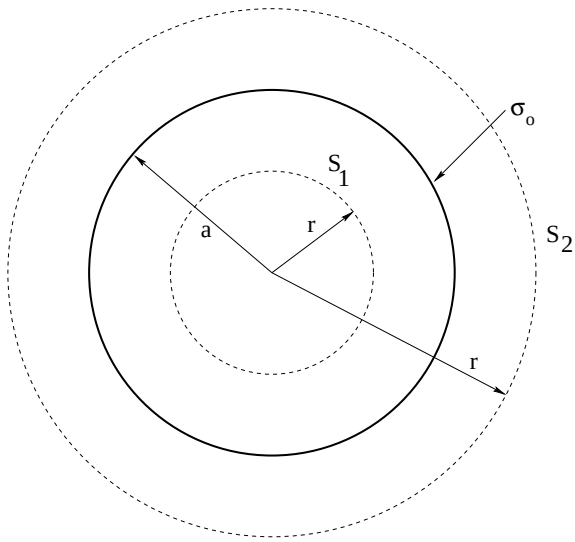


Figure 2.10: A spherical shell of radius a, carrying a uniform charge per unit area σ_0. Two spherical concentric *Gaussian surfaces* S_1 (with radius $r < a$ and S_2 (with radius $r > a$) are shown.

> Suppose you are given a spherical shell of charge with a uniform charge per unit area σ_0 and radius a. Find the field everywhere in space.

As you can see in figure 2.10, there are two distinct regions where we must find the field: *inside* the shell and *outside* the shell. Draw a *spherical* Gaussian surface S_1 inside the sphere (for $r < a$). From the symmetry of the distribution we know that the field $\vec{\boldsymbol{E}}$ must point in the direction of $\vec{\boldsymbol{r}}$ and (hence) be perpendicular and constant in magnitude at all points on the Gaussian surface S_1. Hence:

$$\phi_e = \oint_{S_1} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{r}}\, dA = E_r \oint_{S_1} dA = E_r(4\pi r^2) \tag{2.59}$$

where it is presumed that everybody knows how to integrate to evalute the area of a sphere *and* knows the result.

The total charge Q_S inside this sphere is *zero* by inspection – the fingers and toes thing. That was easy! Now we write Gauss's law:

$$\phi_e = \oint_{S_1} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{r}}\, dA = E_r(4\pi r^2) = \frac{Q_{S_1}}{\epsilon_0} = 0 \tag{2.60}$$

and solve for E_r:

$$\begin{aligned} E_r(4\pi r^2) &= 0 \\ &= \frac{0}{4\pi r^2} \\ E_r &= 0 \qquad \text{for } r < a \end{aligned} \tag{2.61}$$

We've just shown that *in general* the electric field of a spherical shell of charge (like the gravitational field of a spherical shell of mass last semester) *vanishes* inside, but using Gauss's law the derivation was *trivial!*

Outside the shell we draw a *second* spherical Gaussian surface S_2 at $r > a$. Again, the field must be constant and normal to all points on this surface from symmetry. The flux integral is *algebraically identical*:

$$\phi_e = \oint_{S_2} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{r}}\, dA = E_r \oint_{S_2} dA = E_r(4\pi r^2) \tag{2.62}$$

and in fact it will *always* have this algebraic form for a spherical problem, to the point where we will get bored writing this line out umpty times doing homework. Don't let that stop you! Do it every time, as when you know something well enough to be slightly bored writing it out, that's just about perfect, isn't it?

Again we can count up the charge inside S_2 on the thumbs of one hand. It is the total charge on the shell! Which is, in fact (noting that dA for a spherical shell of radius a is $a^2 \sin(\theta)\, d\theta\, d\phi$):

$$\begin{aligned} Q_S &= \int_S \sigma_0\, dA = \int_0^{2\pi} d\phi \int_0^{\pi} \sin\theta d\theta\, a^2\sigma_0 = 2\pi a^2 \sigma_0 \int_{-1}^{1} d(\cos\theta) \\ &= 4\pi a^2 \sigma_0 \end{aligned} \tag{2.63}$$

which we *could* have done using our heads instead of calculus, but there is a clever trick in this example (using $\sin\theta d\theta = -d(\cos\theta)$ to change variables and limits on the θ integral) which we used above when explicitly integrating above and which we'll have occasion to use again in other problems.

Finally, we write out Gauss's law and solve for E_r:

$$\phi_e = E_r(4\pi r^2) = \frac{Q_S}{\epsilon_0} \tag{2.64}$$

or

$$E_r = \frac{Q_s}{4\pi\epsilon_0\, r^2} = \frac{k_e Q_s}{r^2} \tag{2.65}$$

where once again Gauss's law gets us extremely simply something we probably should remember from last semester, which is that *the field of a spherically symmetric charge distribution outside that distribution* is the same as that of a *point charge with the same net charge located the origin*.

This is ***exactly what we got the hard way earlier in this chapter!*** The hard way being an explicit (and quite difficult) integral over the actual charge distribution. The fact that we get the same answer should give us some confidence that Gauss's Law is true and correct. It also convinces us that when we can use it it is *much easier* than explicit integration!

In lecture your instructor will probably do a few more difficult problems – perhaps a solid sphere of charge, or multiple spherical shells, or even a solid sphere with a charge distribution like $\rho(r) = Ar$ where A is a constant! You should be able to do *any* problem with a spherical distribution of charge that you can integrate or sum inside any given Gaussian sphere using this method.

Also note that once one has done a *single* spherical shell, one can easily do as many concentric shells as you might have on your fingers and toes using the *superposition principle*. Simply add the field produced by each shell at the point in question (which might be inside or outside the given shell) to that produced by all the other shells! There's a homework problem to help you learn that – do it!

Example 2.3.2: Advanced: Spherical Shell of Charge

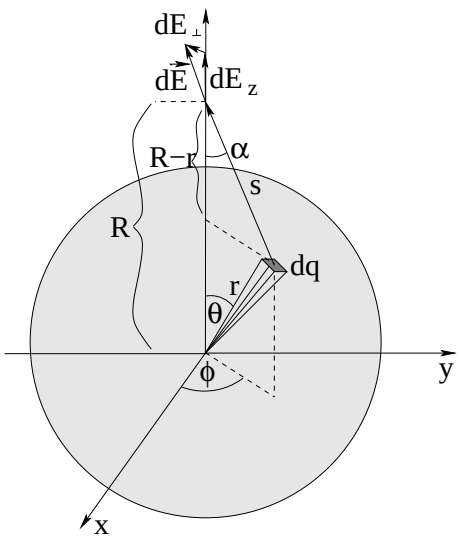


Figure 2.11: Geometry for finding the field of a uniform spherical ***shell*** of constant charge density σ by direct integration, both inside and outside. Note that θ is the angle swept *down* from the positive z axis (the equivalent of "latitude', although measured down from the north pole and not up from the equator) and ϕ is the angle to the x-y *projections* of the point, measured counterclockwise from the positive x-axis, the equivalent of 'longitude'). We call ϕ the *azimuthal angle*.

We will now proceed to set up and find the electric field inside and outside a uniform spherical shell of charge by direct integration. This is just difficult enough that this section is marked "Advanced". However, even normal humans – that is, humans who don't plan to major in physics or mathematics – who probably won't spend a lot of their lifetime integrating nontrivial functions and solving partial differential equations in spherical coordinate systems might want to look the solution over just to see how it works and so that they can use it as a

check for Gauss's Law, which we will cover next.

We begin by choosing a *spherical polar coordinate system*, where a point is represented by the triplet (r, θ, ϕ). Physicists usually use θ and ϕ as represented on the figure above, although in recent years some mathematics texts (and even a few physics texts) swap them so that θ is the usual polar angle in the x-y plane. Sadly, I am an 'old guy' and learned it so thoroughly the other way that I just don't want to change, so we'll stick with the variable representation as given above.

Because the charge distribution (and hence the field) has *spherical symmetry* we lose nothing by choosing the point P where we want to evaluate the field on the z-axis and giving it a z-coordinate R (which is also the distance of the point from the origin). Furthermore, although it is not strictly necessary, we can ignore $dE_\perp$ in the figure above because the problem has *azimuthal symmetry* and hence cannot have a total field component *in* the x-y plane. I'm assuming that you have some familiarity with spherical polar coordinates[55] and things like the area element on the surface of a sphere:

$$dA = r^2 \ \sin(\theta) \ d\theta \ d\phi = -r^2 \ d(\cos\theta) \ d\phi \tag{2.66}$$

but if you are not, it is a great time to review them.

For example, from this point on I'm simplifying ***all spherical integrals over*** θ by using the clever identity:

$$\sin(\theta) \ d\theta = -d(\cos(\theta)) \tag{2.67}$$

to *change variables* from θ to $\cos(\theta)$ so that:

$$\int_0^\pi f(\cos(\theta)) \ \sin(\theta) \ d\theta = \int_{-1}^1 f(\cos(\theta)) \ d\cos(\theta) = \int_{-1}^1 f(x) dx \tag{2.68}$$

This trick doesn't always work, but in physics a lot of time it does and when it does it is really useful!

Consider, then, the small differential chunk of area dA of charge in figure 2.11. We know from our usual rule that the charge in the chunk is the charge per unit volume times the volume of the chunk, or:

$$dq = \sigma dA = \sigma r^2 \ d(\cos\theta) \ d\phi \tag{2.69}$$

We know that the field of ***just this chunk*** at the point P is has a magnitude:

$$dE = \frac{k_e \ dq}{s^2} = k_e \sigma \left(\frac{r^2 \ d(\cos\theta) \ d\phi}{s^2} \right) \tag{2.70}$$

Finally, we only care (for the moment, anyway) about dE_z so we might as well write it down too:

$$dE_z = dE \ \cos(\alpha) = k_e \sigma \left(\frac{r^2 \ d(\cos\theta) \ d\phi}{s^2} \right) \cos(\alpha) \tag{2.71}$$

which we can rewrite using the geometry in figure 2.12 as:

$$dE_z = k_e \sigma \left(\frac{r^2 \ d(\cos\theta) \ d\phi}{s^2} \right) \left(\frac{R - r\cos(\theta)}{s} \right) \tag{2.72}$$

[55] Wikipedia: http://www.wikipedia.org/wiki/Spherical Coordinate Systems. Note well that I'm using the *physics* convention, that is, the second of the two pictures on the right.

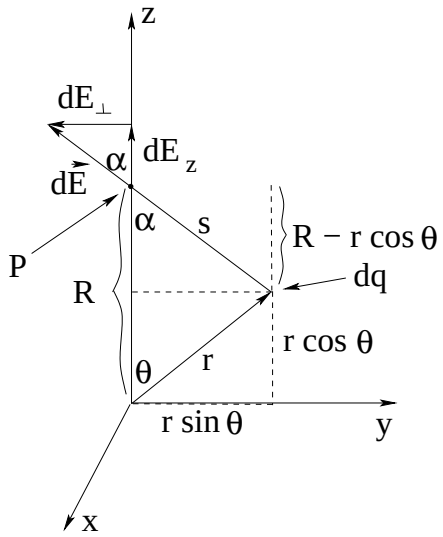


Figure 2.12: Geometry for the vector decomposition of $d\vec{\boldsymbol{E}}$ into dE_z.

Piece of cake, right? Well, not quite. Sadly, s and $\cos(\alpha)$ depend on P, r and θ via e.g. the law of cosines[56] for s and the geometry of the triangle with sides s, $R - r\cos(\theta)$ and $r\sin(\theta)$ for the other. On the other hand, the result still has azimuthal symmetry, which is good! This means we can immediately do the (trivial) ϕ integral and rearrange the result so we can tackle it:

$$\begin{aligned}
E_z &= 2\pi r^2 \sigma k_e \int_{-1}^{1} \frac{(R - r\cos(\theta))\, d(\cos\theta)}{s^3} \\
&= 2\pi r^2 \sigma k_e \int_{-1}^{1} \frac{(R - r\cos(\theta))\, d(\cos\theta)}{(R^2 + r^2 - 2rR\cos(\theta))^{3/2}} \\
&= 2\pi r^2 \sigma k_e \left\{ \int_{-1}^{1} \frac{R\, d(\cos\theta)}{(R^2 + r^2 - 2rR\cos(\theta))^{3/2}} \right. \\
&\qquad \left. - \int_{-1}^{1} \frac{r\cos(\theta)\, d(\cos\theta)}{(R^2 + r^2 - 2rR\cos(\theta))^{3/2}} \right\}
\end{aligned} \tag{2.73}$$

This integral looks difficult, and perhaps it is, but it isn't *that* difficult. The worst thing about it is that we have to integrate the second piece of it by parts. Let's start with the first (fairly

[56]Wikipedia: http://www.wikipedia.org/wiki/Law of Cosines.

easy) piece:

$$\begin{aligned}
\int_{-1}^{1} \frac{R\, d(\cos\theta)}{(R^2+r^2-2rR\cos(\theta))^{3/2}} &= -\frac{1}{2r}\int_{-1}^{1} (R^2+r^2-2rR\cos(\theta))^{-3/2}\,(-2rR\, d(\cos\theta)) \\
&= \frac{1}{r}\frac{1}{(R^2+r^2-2rR\cos(\theta))^{1/2}}\Bigg|_{-1}^{1} \\
&= \frac{1}{r}\left\{\frac{1}{(R^2+r^2-2rR)^{1/2}} - \frac{1}{(R^2+r^2+2rR)^{1/2}}\right\} \\
&= \frac{1}{r}\left\{\frac{1}{(R-r)} - \frac{1}{(R+r)}\right\} \\
&= \frac{1}{r}\left\{\frac{2r}{(R^2-r^2)}\right\} \\
&= \frac{2}{(R^2-r^2)}
\end{aligned} \tag{2.74}$$

That's not so horrible. All I had to do is multiply by $\frac{-1}{2r} \times 2r = 1$ to get it set up for u-substitution as $\int u^{-3/2} du$ (easy), and the rest is all algebra.

The second integral is also easy enough, at least if you you remember how to ***integrate by parts***:

$$\int u dv = uv - \int v du \tag{2.75}$$

Our chore, then, is to identify a u and a dv in the integral:

$$\int_{-1}^{1} \frac{r\cos(\theta)\, d(\cos\theta)}{(R^2+r^2-2rR\cos(\theta))^{3/2}} = \left(\frac{1}{-2R}\right)\int_{-1}^{1} \frac{-2rR\cos(\theta)\, d(\cos\theta)}{(R^2+r^2-2rR\cos(\theta))^{3/2}} \tag{2.76}$$

(where I've gone ahead and multiplied and divided by $-2R$, thinking ahead).

Let's let:

$$u = \cos(\theta) \tag{2.77}$$

and

$$\zeta = R^2 + r^2 - 2rR\cos(\theta) \tag{2.78}$$

so that:

$$dv = \frac{-2rR\, d(\cos\theta)}{(R^2+r^2-2rR\cos(\theta))^{3/2}} = \zeta^{-3/2}\, d\zeta \tag{2.79}$$

We integrate this to get:

$$v = \int dv = -2\zeta^{-1/2} = \frac{-2}{(R^2+r^2-2rR\cos(\theta))^{1/2}} \tag{2.80}$$

Note that this is *just the first integral* before we plugged in the limits!

So let's dig into the algebra. This bit isn't exactly trivial – be patient and try to understand

each step.

$$\begin{aligned}
\left(\frac{1}{-2R}\right)\int_{-1}^{1}\frac{-2rR\,d(\cos\theta)\quad\cos(\theta)}{(R^2+r^2-2rR\cos(\theta))^{3/2}} &= \left(\frac{1}{-2R}\right)\left\{\frac{-2\cos(\theta)}{(R^2+r^2-2rR\cos(\theta))^{1/2}}\bigg|_{-1}^{1}\right. \\
&\quad \left.-\int_{-1}^{1}\frac{-2\,d\cos(\theta)}{(R^2+r^2-2rR\cos(\theta))^{1/2}}\right\} \\
&= \left(\frac{1}{-2R}\right)\left\{\left(\frac{-2}{R-r}-\frac{2}{R+r}\right)\right. \\
&\quad \left.-\left(\frac{1}{rR}\right)\int_{-1}^{1}\frac{-2Rr\;d\cos(\theta)}{(R^2+r^2-2rR\cos(\theta))^{1/2}}\right\} \\
&= \left(\frac{1}{-2R}\right)\left\{\left(\frac{-4R}{R^2-r^2}\right)\right. \\
&\quad \left.-\left(\frac{2}{rR}\right)(R^2+r^2-2rR\cos(\theta))^{1/2})\bigg|_{-1}^{1}\right\} \\
&= \left(\frac{1}{-2R}\right)\left\{\frac{-4R}{R^2-r^2}\right. \\
&\quad \left.-\left(\frac{2}{rR}\right)[(R-r)-(R+r)]\right\} \\
&= \left(\frac{1}{-2R}\right)\left\{\frac{-4R}{R^2-r^2}+\frac{4}{R}\right\} \\
&= \left\{\frac{2}{R^2-r^2}-\frac{2}{R^2}\right\} \qquad (2.81)
\end{aligned}$$

Putting it all together we get:

$$\begin{aligned}
E_z &= 2\pi r^2\sigma k_e\left\{\frac{2}{R^2-r^2}-\frac{2}{R^2-r^2}+\frac{2}{R^2}\right\} \\
&= 2\pi r^2\sigma k_e\frac{2}{R^2}=\frac{k_e\left(4\pi r^2\sigma\right)}{R^2}=\frac{k_eQ}{R^2} \qquad (2.82)
\end{aligned}$$

Ouch! That was a lot of work! And technically, we're not even done – we should really pick a point where $R < r$ (inside the sphere) to prove that the electric field *vanishes* inside. At an interior point, one has to break the $\cos(\theta)$ integral up into two pieces with *different signs* because the charge from the part of the sphere above R creates a field that points *down*, where the charge from the part of the sphere below R points *up*. The integral limits change to:

$$E_z = 2\pi r^2\sigma k_e\left\{\int_{-1}^{R/r}\ldots-\int_{R/r}^{1}\right\} \qquad (2.83)$$

(but otherwise all geometry remains the same).

Ahhhh, too much work. We'll quit here – we've done enough to verify that *even in this extreme example* Gauss's Law gives the correct answer outside of the spherical shell and this gives us an excellent reason to think that it works inside as well!

Example 2.3.3: Electric Field of a Solid Sphere of Charge

Find the electric field at all points in space of a solid insulating sphere with uniform charge density ρ and radius R

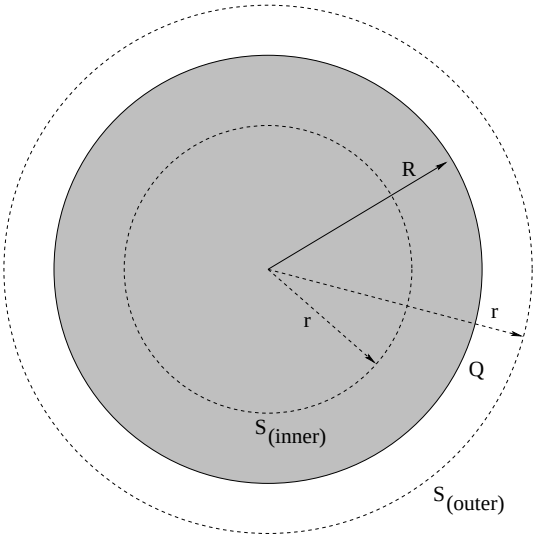


Figure 2.13: A solid sphere of uniform charge density ρ and radius R.

Just for grins, let's do a teensy bit of your homework together. Note well that you *don't get to just copy this onto your paper!* In order to *learn* this and get it right three weeks from now on an exam, you have to be able to do it *without* looking, or copying. So by all means, go through the example, study it, figure it out, then close this book or put aside your digital interface, get out paper, and *do it on your own without looking* – as many times as necessary to make the steps, and reasoning, *easy* to you. Go over it in multiple passes, work on it in your groups, review it in your notes (your teacher/professor probably did this example in class), discuss it in recitation. *Learn* it.

We begin by writing Gauss's Law for the outer surface in the figure **??**:

$$\begin{aligned}
\oint_{S_{\text{outer}}} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA &= 4\pi k_e \int_{V/S} \rho dV \\
E_r 4\pi r^2 &= 4\pi k_e \left\{ \int_0^R \int_0^{2\pi} \int_0^{\pi} \rho r^2 \sin(\theta)\, d\theta\, d\phi\, dr \right. \\
&\qquad \left. + \int_R^r 0\, dV \right\} \\
&= 4\pi k_e (2\pi\rho) \int_0^R r^2 dr \int_{-1}^{1} d(\cos(\theta)) \\
&= 4\pi k_e (\frac{4\pi R^3}{3}\rho) \\
&= 4\pi k_e Q_{\text{total}}
\end{aligned} \tag{2.84}$$

We divide both sides by $4\pi r^2$ and get:

$$E_r = \frac{k_e Q}{r^2} \qquad r > R \tag{2.85}$$

or (as by now you should come to expect) the spherical distribution of charge creates a field *outside* of the sphere that is identical to that of a point charge of the same total value at the origin.

Note that we did a bunch of stuff that we didn't really "have" to do – in an actual solution you'd be tempted to skip those steps or do them by inspection, which is fine, but that risks

confusing at least some of you who *don't* just see what we are skipping and why it is OK to do so. So note well – to find the total charge inside $S_{\rm outer}$, we integrated over the charge distribution from 0 to r *including the region where it was zero* – getting, of course, a zero value for that value. Zero regions drop out, and we'd usually just integrate over the *support* of ρ (the volume where it is nonzero) without thinking about it. Note also that this integral explicitly illustrates doing multiple integrals of a symmetric function – we just do the integrals over each coordinate independently (which is then really easy).

Finally, note the *clever trick* for integrating θ in spherical coordinates. $\sin(\theta)d\theta = -d(\cos(\theta))$, so we change variables from $\theta \to \cos(\theta)$ (and change and swap order of the limits to get rid of the minus sign). It is *very often* much easier to integrate with $\cos(\theta)$ as the variable instead of θ in spherical coordinates – in this case one can just look at it and see that one gets "2" from the integral in your head, for example.

Now we redo the whole thing for the interior integral:

$$\begin{aligned}
\oint_{S_{\rm inner}} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA &= 4\pi k_e \int_{V/S} \rho dV \\
E_r 4\pi r^2 &= 4\pi k_e \int_0^r \int_0^{2\pi} \int_0^{\pi} \rho r'^2 \sin(\theta)\, d\theta\, d\phi\, dr' \\
&= 4\pi k_e (2\pi\rho) \int_0^r r'^2 dr' \int_{-1}^{1} d(\cos(\theta)) \\
&= 4\pi k_e (\frac{4\pi r^3}{3}\rho) \qquad (2.86)
\end{aligned}$$

We divide both sides by $4\pi r^2$ and get:

$$E_r = k_e \left(\frac{4\pi\rho r}{3}\right) \qquad r < R \qquad (2.87)$$

This is a common, and important, example – so let's plot it to make it easier to remember: Things to note and remember: The field increases *linearly* inside the sphere and is *zero* at the origin, not infinite! Outside, the field drops off like $1/r^2$ – as you do more and more of these, you'll come to expect this to the point where you don't think twice about it. Any charge distribution with compact support and a net charge (spherical or not) produces a field that is dominantly *monopolar* and drops off like $1/r^2$ far away from the distribution.

This is very cool! The fact that the field is bounded at the origin means that the *singularity* that appears implicitly in the electrostatic field of a *point* charge need not trouble us if the charge isn't really a *point* charge but is rather a small ball of charge. However, if charge is bound up in a small finite size ball it produces *other* problems – such as the need for a force to hold it all together, as electrostatic charge of a single sign *repels itself*. In the case of a proton, there *is* such a binding force – the strong nuclear force. In the case of electrons, quarks, elementary particles, there *is* (as far as we can tell experimentally or predict theoretically) no such force, and hence those particles "should" be, and experimentally appear to be, truly *pointlike*. Which leads to a whole new set of problems (oops, that nasty infinity is back and has to be dealt with), the invention of renormalizable quantum field theories that soften or throw away the infinity – and in the process, makes physics an *enormously interesting discipline!* Much as we *do* understand at this point, the problem of understanding our Universe, especially

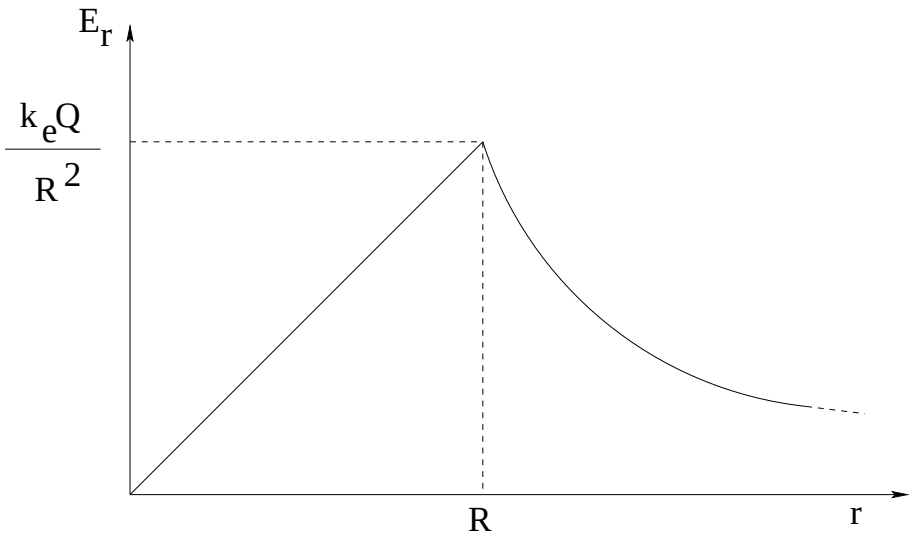


Figure 2.14: Electric field produced by a uniform sphere of charge both inside and outside, as a function of r.

at the smallest length and time scales, is far from solved[57].

The uniform ball of charge is the basis for a model of the *neutral atom* – a positive nucleus surrounded by a uniform ball of negative charge – that helps us understand *polarization* in a few weeks. This model is still used (dressed up with damping and a time dependent driving field) in *physics graduate school* where the model is called the *Lorentz Oscillator Model* for the atom and where the result of analyzing the model is understanding of *dispersion* – basically time dependent dielectric response and the absorption of electromagnetic energy by matter! It sounds complicated, but it isn't, not really. It is *almost* within your reach at the end of taking this introductory course (where we will cover the static part of the result perfectly adequately) – all that separates you is a bit more work with the damped driven harmonic oscillator to help you manage even the dynamics. The reward for the effort is that afterwards, you understand *microscopically* why, e.g. rainbows happen, why the sky is blue, how light from the sun warms the earth, and much more. So keep it in mind for later.

Example 2.3.4: Cylindrical: A cylindrical shell of charge

> Suppose you are given an infinite cylindrical shell of charge with a uniform charge per unit area σ_0 and radius a. Find the field everywhere in space.

We solve this problem *exactly* like we did the sphere. In fact, I block-copied the solution from above to write this and changed only a few minimal things.

There are two distinct regions, inside the cylinder and outside the cylinder. Draw a *cylindrical* Gaussian surface S_1 *of length* L inside the cylinder (for $r < a$). We don't know that the

[57] Students who are interested in reading something accessible for the lay person on the subject are encouraged to pick up a copy of *The Black Hole War: My Battle with Stephen Hawking to Make the World Safe for Quantum Mechanics* by Leonard Susskind. Great fun, and it will help make many of the concepts discussed here clearer in context.

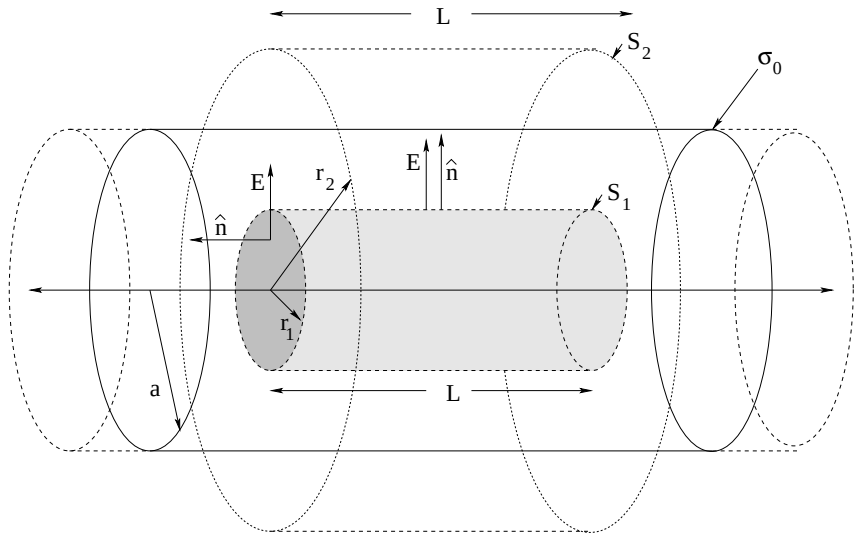


Figure 2.15: A cylindrical shell of radius a, carrying a uniform charge per unit area σ_0. Two cylindrical concentric *Gaussian surfaces* S_1 (with radius $r < a$ and S_2 (with radius $r > a$) are shown.

field is on this surface yet, but we do know that on the *cylinder* part it must lie along $\vec{\boldsymbol{r}}$ and be constant in magnitude and perpendicular to the surface at all points on our Gaussian surface from the symmetry of the distribution. On the end caps the field may well vary with r, but it is *parallel* to those surfaces and therefore there is no net flux through the caps. Hence:

$$\begin{aligned} \phi_e &= \oint_{S_1} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{r}}\, dA \\ &= \phi_{\text{caps}} + E_r \int_{\text{Cyl}} dA \\ &= 0 + E_r (2\pi r) L \end{aligned} \tag{2.88}$$

where it is presumed that everybody knows how to integrate to evalute the area of a cylindrical surface of radius r and length L *and* knows the result[58]. Note that I indicate explicitly that the *flux* through the end caps is zero even though the field there may not be.

The total charge Q_{S_1} inside this cylinder is *zero* by inspection – the fingers and toes thing. That was easy! Now we write Gauss's law:

$$\phi_e = \oint_{S_1} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{r}}\, dA = E_r (2\pi r L) = \frac{Q_{S_1}}{\epsilon_0} = 0 \tag{2.89}$$

and solve for E_r:

$$\begin{aligned} E_r (2\pi r L) &= 0 \\ &= \frac{0}{2\pi r L} \\ E_r &= 0 \qquad \text{for } r < a \end{aligned} \tag{2.90}$$

We've just shown that *in general* the electric field of a cylindrical shell of charge *vanishes* inside.

[58]Think of the label of a soup can. Use mental scissors to snip, snip, snip it off. Unroll it in your mind. It is $2\pi r$ long and L wide.

Outside the shell we draw a *second* cylindrical Gaussian surface S_2 with length L at $r > a$. Again, the field must be constant and normal to all points on this surface from symmetry, again the flux through the end caps must be zero even though the field on the end caps may not be. The flux integral is *identical*:

$$\begin{aligned} \phi_e &= \oint_{S_2} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{r}}\, dA \\ &= \phi_{\text{caps}} + E_r \int_C dA \\ &= E_r (2\pi r) L \end{aligned} \tag{2.91}$$

and in fact it will *always* be this algebraic form for a cylindrical problem, to the point where we will get bored writing this line out umpty times doing homework. Don't let that stop you! Do it every time, as when you know something well enough to be slightly bored writing it out, that's just about perfect, isn't it?

Again we can count up the charge inside S_2 on the thumbs of one hand. It is the total charge on the shell *inside the Gaussian surface of length L!* Which is, in fact (noting that dA for a cylindrical shell of radius a is $a d\theta\, dz$):

$$\begin{aligned} Q_{S_2} &= \int_S \sigma_0\, dA = \int_0^{2\pi} d\theta \int_{-L/2}^{L/2} a\sigma_0\, dz \\ &= 2\pi a L \sigma_0 \end{aligned} \tag{2.92}$$

which we *could* have done using our heads instead of calculus, but again this way you get to see how to do a two dimensional integral that separates into two trivial one dimensional integrals.

Finally, we write out Gauss's law and solve for E_r:

$$\begin{aligned} \phi_e = E_r(2\pi r L) &= \frac{Q_{S_2}}{\epsilon_0} \\ E_r &= \frac{2\pi a L \sigma_0}{2\pi L \epsilon_0} \frac{1}{r} \\ &= \frac{\sigma_0}{\epsilon_0} \frac{a}{r} \\ &= \frac{2k\lambda_0}{r} \end{aligned} \tag{2.93}$$

where I've used the fact that $\lambda_0 = Q_S/L = 2\pi a \sigma_0$ to help show that *the field of a cylindrically symmetric charge distribution outside that distribution* is the same as that of a *line of charge with the same net charge per unit length on its axis*.

Note well: The parameter L (which you *made up* when you drew your Gaussian surface) *cancels* from the problem. Of course it does! And a good thing, too!

In lecture your instructor will probably do a few more difficult problems – perhaps a solid cylinder of charge, or multiple cylindrical shells, or even a solid cylinder with a charge distribution like $\rho(r) = Ar$ where A is a constant! You should be able to do *any* problem with a cylindrical distribution of charge that you can integrate or sum inside any given Gaussian cylinder using this method.

Example 2.3.5: Planar: A sheet of charge

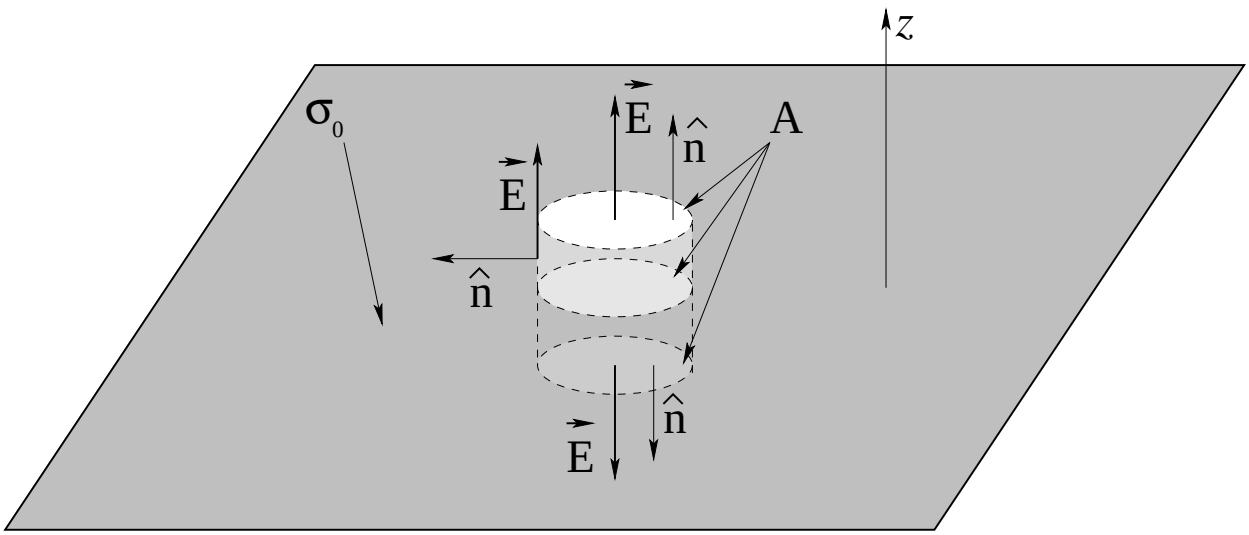


Figure 2.16: An (infinite) plane sheet of uniform charge per unit area σ_0. The Gaussian surface in this case is a simple "pillbox" symmetrically drawn so it intersects the sheet as drawn.

> Suppose you are given an infinite sheet of charge with a uniform charge per unit area σ_0. Find the field everywhere in space.

We solve this problem *exactly* like we did the two above. You (by now) should know the drill.

Here we only need to draw a single Gaussian surface as indicated in figure **??** above. We will again draw a *cylindrical* Gaussian surface *of length* z, but this time it must be symmetrically located so that it *symmetrically intersects* the plane of charge with $z/2$ of its length above and below the plane. This cylinder has an end-cap area of A which (like L in the previous problem) will *cancel* when we go to evaluate the field. We don't know what the field is on this surface yet, but we do know that on the *end-caps* it must lie parallel to $\vec{z}$ and be constant in magnitude and perpendicular to the end caps at all points. On the side of the cylinder the field may well vary with r, but it is *parallel* to this surface and therefore there is no net flux through it. Hence:

$$\begin{aligned} \phi_e &= \oint_S \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{z}}\, dA \\ &= \phi_{\text{side}} + 2E_z A \\ &= 2E_z A \end{aligned} \tag{2.94}$$

where you should note that we have *two* end caps, each of which contributes $E_z A$ to the flux.

The total charge inside this Gaussian surface is trivial:

$$Q_S = \int_A \sigma_0\, dA = \sigma_0 A \tag{2.95}$$

where there really isn't much of anything to integrate or evaluate.

Finally, we write out Gauss's law and solve for E_z:

$$\begin{aligned} \phi_e = 2E_zA &= \frac{Q_S}{\epsilon_0} = \frac{\sigma_0 A}{\epsilon_0} \\ E_z &= \frac{\sigma_0}{2\epsilon_0} \\ &= 2\pi k\sigma_0 \end{aligned} \tag{2.96}$$

where we note that the field is *uniform* – it doesn't depend on z, and of course it cannot depend on x and y either as every point is in the middle of an infinite plane! This last result is very important.

Note well: The parameter A (which you *made up* when you drew your Gaussian surface) *cancels* from the problem. Also note that this is exactly the result we got for the field on the axis of a disk of charge when we let the radius go to ∞. This gives us confidence that Gauss's Law *works!*

As before, in lecture your instructor will probably do a few more problems, perhaps a slab of charge of finite thickness or the field produced by *two* infinite sheets of charge, one with charge σ_0 and the other with charge $-\sigma_0$ (a model for a parallel plate capacitor that we will study in great detail shortly).

2.4: Gauss's Law and Conductors

2.4.1: Properties of Conductors

A conductor is a material that contains many "free" charges that are *bound to the material* so that they cannot easily jump from the conductor into a surrounding insulating material (where a vacuum is considered an insulator for the time being, as is air) but *free to move* within the material itself if any e.g. electrical field exerts a force on them.

In a typical conductor – for example a metal such as silver or copper – there is on average roughly one free electron *per atom* in the material. That is in the ballpark of 10^{24} free electrons per mole of metal, which in turn is somewhere between 10^4 and 10^5 Coulombs of free charge! As we discussed in class, two charges of one Coulomb each separated by one meter exert a force of 9×10^9 Newtons on each other, more than enough to *rip apart* any normal material (releasing roughly ten billion joules of energy) as they "explosively" rejoin. Consequently we have no hope of either removing *all* of the free electrons from (say) 50 or 60 grams of solid metal and separating them by any appreciable macroscopic distance, or adding enough electrons so that every atom has an extra one. From energy balance alone, it would require the entire energy output of a city-sized gigawatt electrical generator for more than a day to accomplish it, and the material itself would come apart in a cloud of superheated plasma long before we succeeded.

This means that we can consider the free charge in a macroscopic chunk of conductor to be 'inexhaustible'. As far as we're concerned, we can always add charge to a conductor, or take it away, or rearrange it as we please with fields and forces, and never run a risk of "saturating" the conductor's ability to supply still more free charge, at least not with the work we are willing to do (and pay for!) and while keeping the conductor itself intact.

Now let's think a moment about the "free" bit without worrying about actual equations – we'll use logic and reason to figure out what we should expect to see when we try to push free charges around inside (say) a metallic conductor. First, we have to try to imagine a force we can use to push on those free charges *with*. So far, we know of only two forces that can push on massive charged particles – gravitation and the electrostatic force. Obviously Earthly gravitation is much, much weaker than the electrostatic forces that bind the electrons to atoms and molecules, as we don't observe electrons "dripping" out of solid matter and falling to the ground under a conductor (and besides, with luck you did an in-class problem where you concluded at the end that the Coulomb force acting on a bound electron in (say) a hydrogen atom is some 10^{36} ***times stronger*** than the gravitational force between the electron and proton in the nucleus at any separation). Finally, it is a True Fact (well beyond the scope of the course to treat in detail) that the quantum interaction of an electron with protons or neutrons via the strong and weak *nuclear* forces is essentially irrelevant.

Consequently, if we want to exert a force on the free charges in an ***isolated conductor*** to move them *independent* of the underlying atomic/molecular lattice they are bound to (they will still fall with the object they are bound to as gravity acts on the protons, neutrons and electrons alike) ***it will have to be using the electric field itself*** (or, as we will see later, with the ***electromagnetic*** field as magnetism will also push free charges around, but *only if they are moving and never doing any actual work as it does so*).

Next, those charges are (by hypothesis) free to move and hence ***will accelerate in the direction of the net electrostatic field/force we apply to push them around*** to the extent permitted by their interaction with the underlying lattice of atoms or molecules that make up the conducting medium (something we will discuss in detail in Week/Chapter 5). They will continue to move (accelerating or in a steady state of motion) until they encounter the *boundaries of the isolated material*, where a ***potential energy barrier*** holds them inside of the conductor, exerting an electrostatic ***surface force*** perpendicular to the surface of the conductor sufficient to prevent any motion across those boundaries). At the boundaries, free charges build up and rearrange until they create a macroscopic electrostatic field of their *own*.

Empirically, this rearrangement of charge in an *isolated* conductor placed in an external electrostatic field only lasts for a very, very short time – nanoseconds to microseconds, depending on a variety of things we'll learn about later. By the end of this time, the conductor will reach a state we will call ***electrostatic equilibrium*** (which I will routinely abbreviate **'ESE'** as we will refer to this state quite frequently for the next few chapters) where the net free charge in the conductor ***stops moving around*** as all of the forces acting on it, including the surface force confining it to the material, *cancel*. Since the force that caused the rearrangement was, by the argument above, almost certainly caused by an external electrostatic field, and since *only* electrostatic fields exert non-negligible forces on the conduction charges, we can conclude that the charges have, at that time, arranged themselves in such a way that the total electrostatic force on the free charges in the material is *zero*.

At last we are ready to bring our threads of logic and reason above to a conclusion, one where GLE will play a crucial, and quantitative, role! When the conductor is in ELE, we can see that the following five propositions must be true, and of course, you should *learn* them, and more importantly, master the chain of reasoning we use to arrive at them!

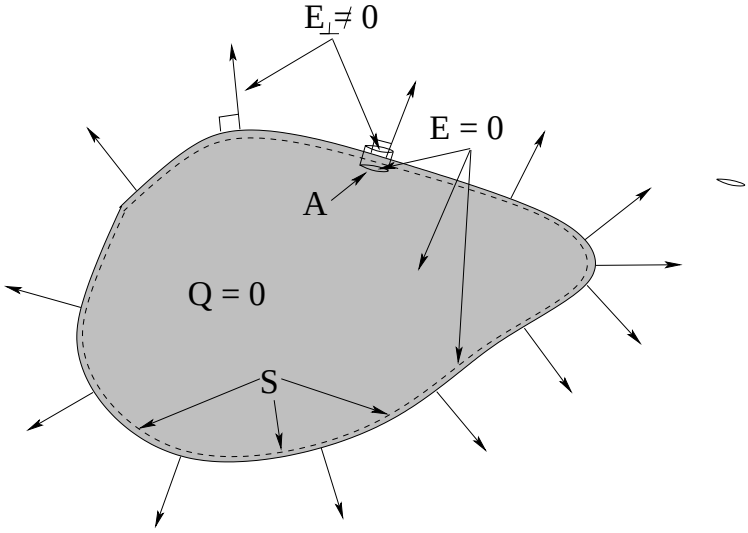


Figure 2.17: An arbitrary chunk of conducting material in electrostatic equilibrium can have no field inside, or else it wouldn't be in equilibrium. It can have no field tangent to its surface, or it wouldn't be in equilibrium. From these facts we can deduce several useful things about conductors in electrostatic equilibrium using Gauss's Law.

1 ***The electric field vanishes inside a conductor in ESE***. The argument for this is pretty much given in its entirety above. By definition, in ESE the free charges in the bulk inerior of the conductor are not moving/accelerating. Only forces exerted by a *nonzero* electric field are capable of making them move relative to the underlying lattice, and evidently this force is zero. Consequently:

$$\vec{\boldsymbol{E}}_{\text{inside}} = 0 \tag{2.97}$$

Note that when we leave our coarse-grained macroscopic world where a micrometer is "infinitesimal" and look at things at the atomic/molecular scale, $\vec{\boldsymbol{E}}$ really vanishes across the *first few layers of atoms*, a distance of a few nanometers, not at a mathematically precise "surface". At the *macroscopic* scale where we consider charge to be *continuously* distributed along lengths, surfaces, or volumes we will consider this layer a few tens of angstroms thick as being, for all practical purposes, an "infinitely" thin surface where the field "instantly" vanishes. Bear this in mind as we continue.

2 ***There is no net charge in the volume of matter inside a conductor in ESE***. This follows from **1** and from Gauss's Law applied "backwards". Consider the gaussian surface S drawn *just inside* the surface of the arbitrary "blob" of conducting material drawn in figure 2.17 above. For this surface or any *other* closed surface drawn entirely inside the conductor (that is, not containing any part of its surface or the space outside):

$$\oint_S (\vec{\boldsymbol{E}} = 0) \cdot \hat{\boldsymbol{n}} dA = 0 = \frac{1}{\epsilon_0} \int_{V/S} \rho_{\text{inside}}\, dV = \frac{Q_{\text{inside}}}{\epsilon_0} \quad \Rightarrow \quad \boxed{\boldsymbol{\rho}_{\text{inside}}, \boldsymbol{Q}_{\text{inside}} = \mathbf{0}} \tag{2.98}$$

This is "backwards" because instead of using knowledge of Q or ρ to find $\vec{\boldsymbol{E}}$, we use our knowledge of $\vec{\boldsymbol{E}}$ from **1** to conclude something important about ρ and/or Q!

3 ***All unbalanced charge placed or distributed on a conductor in ESE must reside on the surface***. This follows from **2**. We can certainly add charge or remove charge from

the conductor to give it a nonzero *net* charge. We can also put a neutral conductor in an external electric field, but then *some* unbalanced (not coarse-grained neutral) charge distribution on the conductor must somehow cancel that field on the interior so that **1** and **2** remain true in ESE.

Well, if this unbalanced charge isn't on the inside, it *must be on the surface*, just *outside* the dashed line representing the largest surface S we can construct that is completely "inside" the figure 2.17 (which is really a few nanometers inside its "true" surface) where it will form a coarse-grained *surface* charge density $\sigma \neq 0$ consisting of charges unbalanced only in a layer a few atoms/molecules thick.

4 ***There can be no field component parallel to the surface of a conductor in ESE***. In equation form, we write this:

$$\boxed{\vec{\boldsymbol{E}}_{||} = 0} \tag{2.99}$$

The argument is identical to that used to deduce rule **1**. Suppose that there were a nonzero electric field parallel to the surface of the conductor "just outside" of the conductor. The only thing that can cancel that field so that it goes to zero by the time one is *inside* the conductor is the field created by free charge inside the conductor itself. The electric field is *continuous* everywhere but at the location of true point charges, so the field outside *must* penetrate at least that surface layer a few atoms thick into the conductor before it can be completely cancelled in the coarse grained limit, but *even in this layer*, if it were macroscopically *nonzero*, it would act on the free charges there and push them around, *contradicting the assertion that the conductor is in ESE!*

From this we can conclude that by the time the charge on the isolated conductor stops rearranging (reaches ESE) when we put it in an electrostatic field or add a net charge to it, there can be no electric field component parallel to its surface *at* – "just" outside – that surface.

5 ***The electrostatic field just outside of a conductor in ESE is perpendicular to its surface or zero.*** Since the field at the surface of a conductor in ESE can't be *parallel* to the surface, the only remaining possibility is that $\vec{\boldsymbol{E}}_{\perp}$ may *not* be zero just *outside* the surface. However, from **1** above it *must* be zero *inside* (at least, inside that surface layer of charge a few atoms thick that we are treating as being "infinitesimally" thick in the coarse-grained limit).

This is an invitation to apply GLE to find a relation between the unbalanced surface charge distribution permitted by **3** and the allowed (possibly nonzero) electric field perpendicular to that surface just outside. Consider an "infinitesimally" thin (just thick enough to contain the entire surface charge layer we've been discussing above) Gaussian pillbox with inner surface "inside" the conductor where $\vec{\boldsymbol{E}} = 0$, and outer surface *just outside* like that drawn in figure 2.17 above. We already know that $\vec{\boldsymbol{E}}_{||} = 0$, so there can be no electric field flux through the sides of the pillbox. There is no flux through the bottom of the pillbox inside the conductor because the field is zero there. The only thing that contributes to the flux, then, is the top surface just outside the conductor. GLE for a pillbox with an (infinitesimally small) top surface A drawn parallel to the suface is thus:

$$\oint_S \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA = E_{\perp} A = 4\pi k_e Q_{\text{in } S} = \frac{\sigma A}{\epsilon_0} \tag{2.100}$$

Cancelling the arbitrary area A of the pillbox, we get:

$$\boxed{\vec{\boldsymbol{E}}_\perp = 4\pi k_e \sigma = \frac{\sigma}{\epsilon_0}} \tag{2.101}$$

In words: the field at the surface of a conductor is perpendicular to the surface and directly proportional to the surface charge density with constant of proportionality $4\pi k_e = 1/\epsilon_0$! Wherever there is an unbalanced surface charge, there will be a nonzero electric field normal to the surface, going *out* from the surface if σ is positive, *in* towards the surface if it is negative.

Note well that *all of these properties are for ESE only!* As we will shortly learn, conductors that carry nonzero currents are *not* in ESE and *do* have nonzero electric fields inside that *are* parallel to the surfaces. I often ask questions that test whether or not you understand this on exams, so be careful!

Example 2.4.1: Field and Charge Distribution of a Blob of Conductor

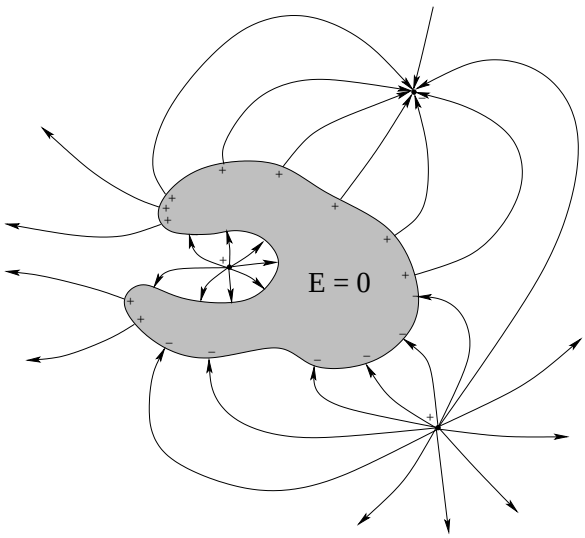


Figure 2.18: A conductor with an arbitrary shape near an external charge rearranges its charge into a surface charge that cancels the field inside and causes the field near the surface to be perpendicular to the surface.

Suppose we have an *arbitrary shape* of conducting material. As usual, we'll visualize this as an amoeboid blob of metal with no particular symmetry or shape so that we aren't tempted to use any "special" property of a regular shape like a sphere or cylinder in our analysis. It is *at rest* in the field produced by a number of nearby fixed point charges (in the plane of the figure) of either or both signs, and has been for some time.

What can we tell about the field inside the conductor, the charge distribution of the conductor, and so on using *just the principles enumerated above?* The following are possible *questions* you might be asked on a quiz or exam, with an explanation of the answers.

- Where is the field inside strongest? (The field inside is *zero everywhere*, trick question.)

- Given the conductor and the charges, can we sketch a guesstimate of the field in the plane of the figure? (Yes, done for you above. Note the use of the rule that the field lines enter or leave the surface of the conductor at right angles. Of course in reality the conductor and location of external charge could/would be three dimensional and everything could be more complicated...)

- Is the entire conductor electrically neutral? (No, charge on the *surface only* has re-arranged, with negative electrons being attracted to the positive charges and getting as "close as they can" to them (while still remaining as far apart as possible from each other, in competition) and leaving behind positive charges on the atoms as "close as possible" to the nearby negative charges ditto. The + and - signs on the figure represent a possible visualization of this surface charge, which is related to the field outside by:

$$E_{\perp} = 4\pi k_e \sigma$$

  from Gauss's Law plus our knowledge that the field vanishes inside.)

- Is the *interior* of the conductor electrically neutral? (Sure, it must be. If it weren't the charges there would create a field (see Gauss's Law!) and move away from one another until they reach the surface and become part of the surface charge distribution.)

- Can we tell just from the figure whether or not the conductor is overall electrically neutral (has a net charge or not)? (No, not really. The lines of force in the figure above suggest that it might be, but *we* drew them in response to the question above, right? So there isn't any real reason to rely on them. What we *do* know is that if it isn't neutral, all of the surplus charge will be located on the surface of the conductor, arranged in just the right way that the field lines leave the surface at right angles.)

Make sure that you understand the ideas underlying all of these answers.

Example 2.4.2: Two Thick Plates Plus Wires (Capacitor)

In the figure above, two conducting plates with facing area A, with wires attached to them are schematically illustrated. The plates are deliberately drawn to be *thick* and the gap between the plates is similarly exaggerated. We assume that the plates are ***large*** compared to this gap.

Suppose equal and opposite charges $\pm Q$ are placed on the plates (and prevented from flowing together through the conducting wires). We know that the field inside the shaded metal region must be zero once the plates are in electrostatic equilibrium. We also know that the charges have to spread out on the surface(s) of the conductors. Finally, we know that the oppposite charges will ***attract*** across the gap between the plates.

The charge distribution illustrated above, with the charges spread out uniformly on the facing surfaces of the plates as $\pm\sigma = \pm Q/A$ satisfies all of these conditions. As we have seen, the field of a single plane sheet of charge is $E = \dfrac{\sigma}{2\epsilon_0} = 2\pi k_e \sigma$, directed ***away*** from a positive surface charge density.

The field lines from the upper plate go up above the surface layer $+\sigma$ and down below it. Similarly the field lines go down above the surface layer $-\sigma$ and up beneath it. The idealized

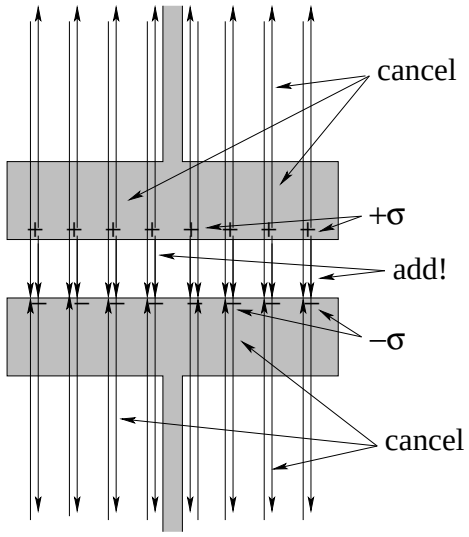


Figure 2.19: Opposite charges placed on two facing conducting plates spread out to form ***surface charge layers***. This is exactly what is needed to ***cancel the fields of the two layers in the plates themselves*** while ***adding together in the space between the plates.***

field lines from each surface charge layer go ***all he way to infinity***, where the total field is the vector sum of the two fields, one from the upper layer $+\sigma$, the other from the lower layer $-\sigma$.

As you can see in the figure, above $+\sigma$ the up field from the upper layer and the down field from the lower layer ***cancel***, making the field zero (as desired) everywhere in the metal plate above $+\sigma$. The same is true below the lower layer $-\sigma$. In between the plates, though, the field from the upper layer is down, the field from the lower layer is down *also* and hence the total field is:

$$E_{\text{tot}} = E_u + E_l = \frac{\sigma}{2\epsilon_0} + \frac{\sigma}{2\epsilon_0} = \frac{\sigma}{\epsilon_0}$$

down. The field runs from the positive surface layer to the negative surface layer and is zero everywere inside the bulk conductor and for that matter in the air above and below the plates!

This is an important example as finding this field in terms of $\sigma = Q/A$ is a required step for finding first the potential difference between the two plates (next chapter) and then the capacitance of this arrangement of conductors (the chapter after that).

Note well! The charges spread out on these surface ***must be equal and opposite!*** This is true even if one puts ***different*** charges on the two plates! You will work some examples for spherical conducting shells for homework and should pay attention to this happening there as well, and for the same reasons.

Creating Charged Objects

As noted at the beginning of week 1, the ability to demonstrate things like Coulomb's Law revolves around several things. One is the ability to accurately measure very small forces – this Coulomb was able to do with his personally invented torsional balance. The other was the ability to create controlled amounts of charge and place it on isolated conductors on his balance.

This section is intended to give you *some* idea of how one can generate charge (by means of friction or induction) and how one can then use it to generate like amounts of charge for experiments. The primary two means for the latter are charging by induction and charge transfer.

Charging by induction is illustrated below:

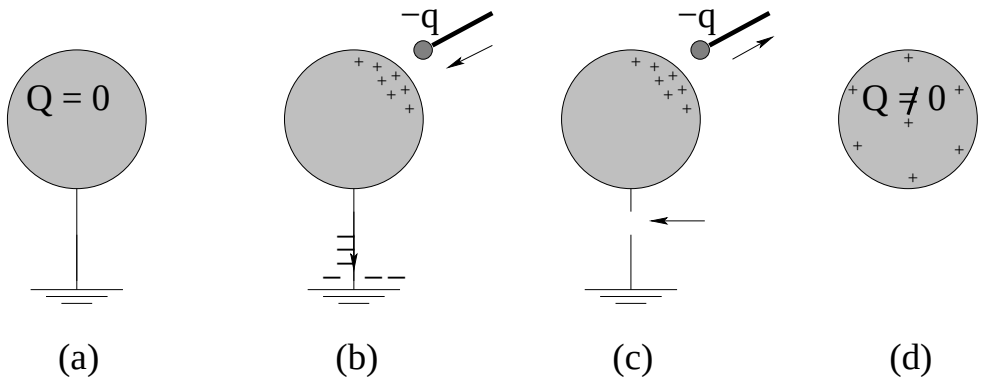


Figure 2.20: Charging by induction in four steps.

In the first panel (a), a neutral, spherical conductor is connected to "ground", which can be thought of as a ***really, really big conductor***, a reservoir of charge that generates essentially no additional field no matter how much charge you pull from it or deliver to it. Note well the symbol used for ground.

Second (b) a charged object (perhaps prepared by the triboelectric effect, rubbing a glass rod with silk to produce the negative charge shown or using a crude electrostatic generator) is brought near the conductor. There it attracts charge of the opposite sign and repels charge of the same sign which tries to get as far away as possible, which happens to be the ground.

Third (c) the connection to ground is removed, isolating the charge on the sphere, and the induction charge is removed, producing:

(d) a charged, isolated conducting sphere.

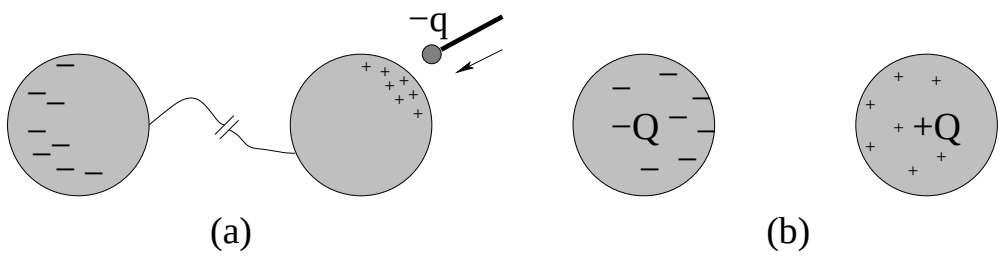


Figure 2.21: Charging by induction with no ground.

It is not strictly necessary to use the ground. You can also produce equal and opposite charges by using two spheres connected with a wire, bringing the charged object near one and pushing charge over to the other before disconnecting the wire as before. This is schematically illustrated in figure 2.21 above. Since the two objects began electrically neutral, they will have equal and opposite charges!

To produce the *same* charge on two identical conducting spheres, it suffices to charge one

sphere up as shown in figures 2.20 or 2.20 and then bring it into contact with an identical sphere. The charge then splits onto the two spheres symmetrically, leaving them both with half of the original charge. This process can be repeated with more spheres, producing a series of spheres with Q, $Q/2$, $Q/4$, $Q/8$ on them. This suffices to be able to demonstrate the needed bilinearity in charge in Coulomb's Law, provided only that one can measure very small forces and distances with some accuracy.

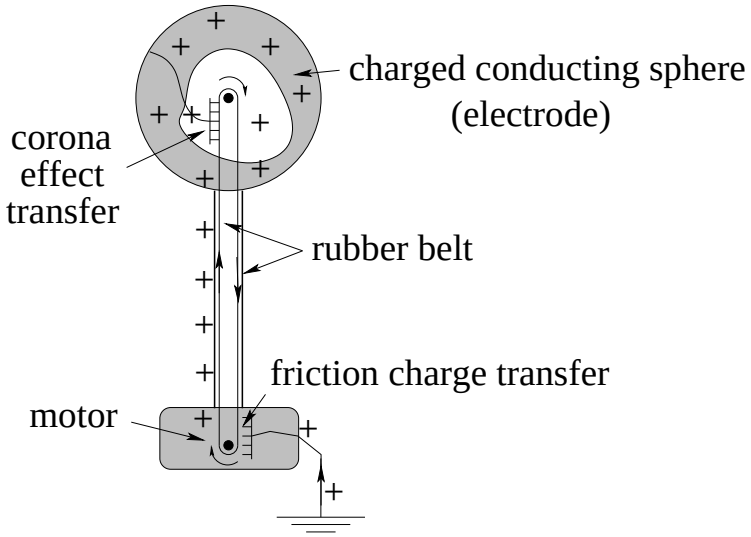


Figure 2.22: A Van de Graff Electrostatic Generator

Finally, it is possible to charge up a conducting sphere at the end of an insulating rod and move it *inside* of a hollow, conducting sphere and touch it to the larger sphere on the inside. Charge is immediately transferred and pushed to the *outside* of the larger sphere. The advantage of doing this is that one can do it over and over again, accumulating an ever-larger charge on the larger sphere! This is the basis of the ***Van de Graff*** generator illustrated in figure 2.22, which uses a flexible (rubber or silk) belt to continuously convey ***tribolelectrically generated charge*** picked up from ground to a hollow conducting sphere at the top.

Triboelectric charge is charge that comes from rubbing two materials together and transferring charge preferentially from one to another using simple friction (tribology in physics and engineering is, recall, the study of friction) depending on the relative *electronegativity* of the materials being rubbed together. By making the rollers of the e.g. rubber belt of different materials and/or physically rubbing the rubber belt with a soft material, one can generate a charge on the rubber at the bottom, push it up on the *insulating* belt through a hole in the top spherical conductor on the belt, and pull it off near the top roller with a plate covered with sharp points near the belt via the *corona effect* discussed in the chapter on dielectrics and capacitance.

Inside, a wire transfers it to the sphere, where it immediately moves to the outside surface of the sphere. One has to push further charge up through the hole against the force exerted by the charge already on the sphere, so the motor at the bottom has to *do work* in order to increase or maintain the charge on the sphere.

Van de Graff generators were the basis of the very first "atom smashing" particle accelerators used to probe nuclear structure. They are still in use today in research accelerators[59]

[59] Duke University has a high-resolution tandem Van de Graff accelerator as of the time of this writing – I helped

They were quickly largely replaced by e.g. cyclotrons – described elsewhere in this text – and other accelerators capable of achieving more than the 1-30 MeV particle energies they can produce. While Van de Graff generators were for a time used or considered for the productions of nucleotides used in nuclear medicine, I was able to find no real evidence that they are currently in an sort of medical production environment. The much more compact cyclotron, on the other hand, has almost become a standard piece of hospital equipment, because many of the most useful isotopes have very short half-lives (deliberately!) and hence have to be produced right next to where they will be used (as close as "down the hall") in order for the isotopes not to decay below useful levels during the time required for transportation.

to design its beam optics as a project in my senior year at Duke as an undergraduate.

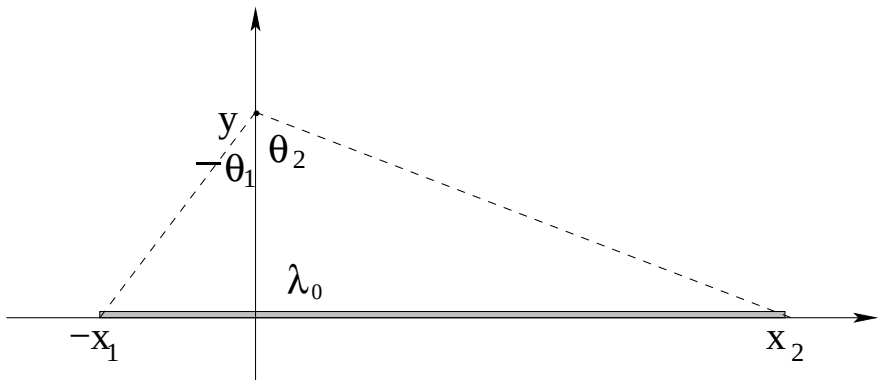
Homework for Week 2

Problem 1.

Physics Concepts

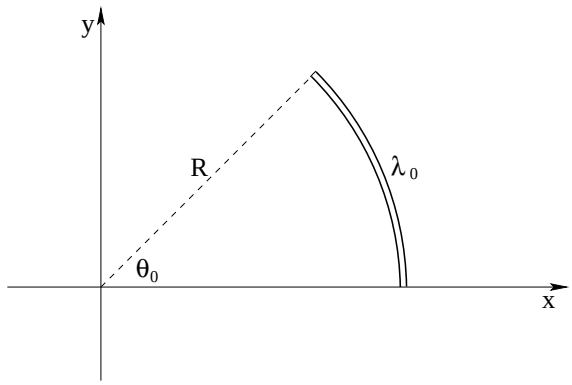
Make this week's physics concepts summary as you work all of the problems in this week's assignment. Be sure to cross-reference each concept in the summary to the problem(s) they were key to. Do the work carefully enough that you can (after it has been handed in and graded) punch it and add it to a three ring binder for review and study come finals!

Problem 2.



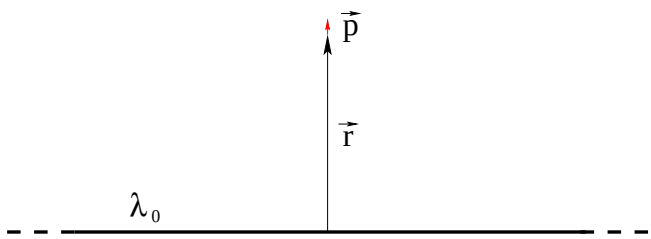
A uniform line of charge with charge per unit length λ_0 runs from $-x_1$ to x_2 *as shown* on the x axis. Find ***both components of the electric field*** at the (arbitrary) point y on the vertical axis indicated on the figure. You may express your answer in terms of $-\theta_1$ and θ_2 (as shown) instead of $-x_1$ and x_2 if you would prefer.

Problem 3.



An arc of linear charge density λ_0 and radius a is centered on the origin and subtends an angle θ_0 as shown. Find the electric field ***vector*** at the origin. Note that this requires finding magnitude *and* direction, or (easier) else finding the field's cartesian components.

Problem 4.



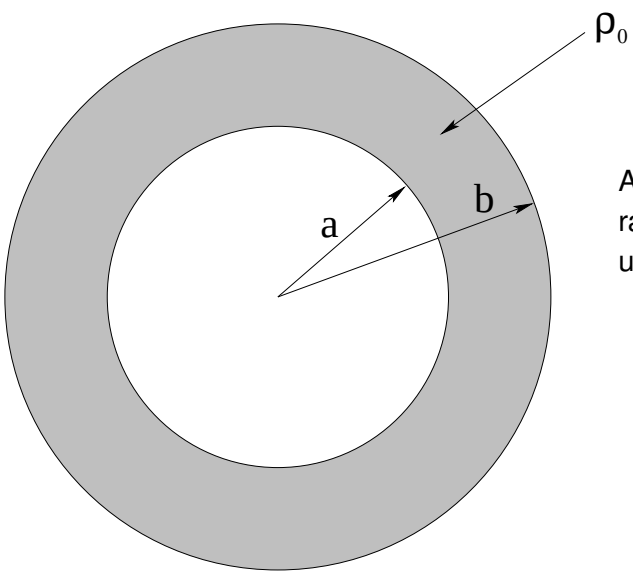
A 'point dipole' $\vec{\boldsymbol{p}}$ is located a distance r from an infinitely long line of charge with a uniform linear charge density $+\lambda_0$. Assume that the dipole is aligned with the field produced by the line charge. Determine the force acting on the dipole. Is it attracted to or repelled by the line?

Note that there are two ways to do this problem. One uses calculus and

$$F_r = -\frac{dU}{dr} = \frac{d\vec{\boldsymbol{p}} \cdot \vec{\boldsymbol{E}}}{dr}$$

The other assumes a finite size dipole $\vec{\boldsymbol{p}} = q\vec{\boldsymbol{\ell}}$ and uses e.g. the binomial expansion of the force in the limit $\ell \ll r$ as you did on your homework in the previous chapter to arrive at the result for a 'point dipole'. You may want to look back at those problems as you do this, and compare the difficulty of the two methods.

Problem 5.

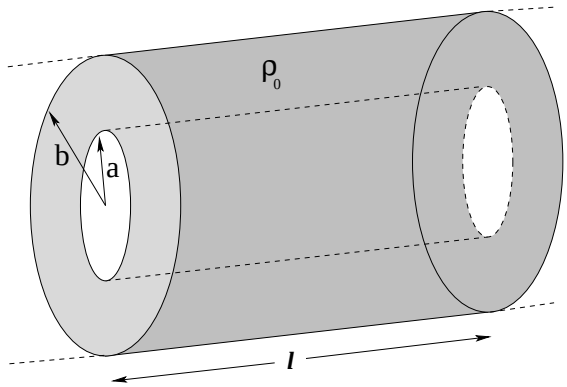


A thick, nonconducting spherical shell of inner radius a and outer radius b has a uniform volume charge density $\rho(r) = \rho_0$.

a) Find the total charge of the shell.

b) Find the electric field everywhere.

Problem 6.



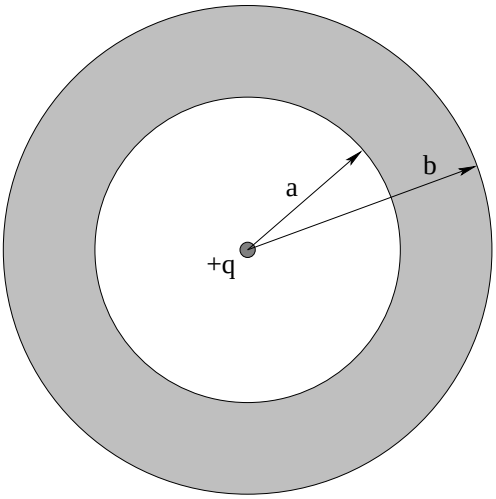
An infinitely long, thick, nonconducting cylindrical shell of inner radius a and outer radius b carries a uniform volume charge density $\rho(r) = \rho_0$.

a) Find the total charge in a chopped-off section of the infinite cylinder of (finite) length ℓ.

b) Find the electric field everywhere.

c) Let $a = 0$. Find the electric field (now that of a *uniform cylinder of charge*) everywhere.

Problem 7.



A spherical ***conducting*** shell with *zero net charge* has inner radius a and outer radius b. A point charge q is placed at the center of the shell (the origin) as shown.

a) Use Gauss's Law and the properties of conductors in equilibrium to find the electric field in the three regions:

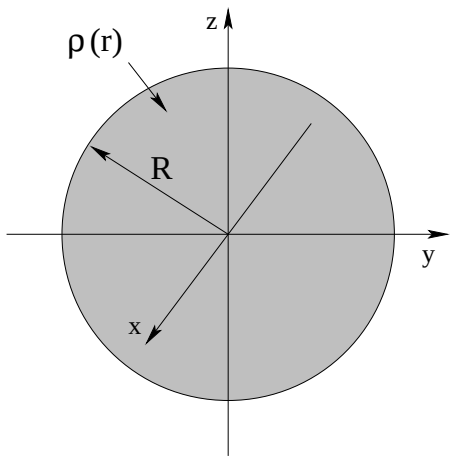
   I: $(r < a)$ **II:** $(a < r < b)$ **III:** $(b < r)$

b) Find the charge density on the inner and outer surfaces of the shell.

Problem 8.

A conducting neutral sphere of radius R is placed in a uniform electric field $\vec{\boldsymbol{E}} = E_0\hat{z}$. Using Gauss's Law and the properties of conductors in equilibrium, *draw a qualitatively correct representation* of the electric field that results. Also indicate on the figure the qualitative distribution of charge on the surface of the conductor one might expect as its charge polarizes in response to the external field. Is there more charge near the "equator" or the poles?

Problem 9.



Suppose you have a solid sphere with a radius R and a spherically symmetric charge density $\frac{dQ}{d\mathcal{V}} = \rho(r)$. Use Gauss's Law to find the electrostatic field(s) at all points in space for:

a) A *uniform* charge density $\rho(r) = \rho_0$ for $r \leq R$.

b) A *non-uniform* charge density $\rho(r) = \rho_1 \frac{r}{R}$.

In both cases, show that the field outside of the sphere is $E_r = \frac{k_e Q_{\text{tot}}}{r^2}$.

Note that this problem requires you to do an *integral* to evaluate the total charge inside a Gaussian surface. I'll help you set it up. The volume of a differentially thin spherical shell is its area $A = 4\pi r'^2$ times its thickness dr':

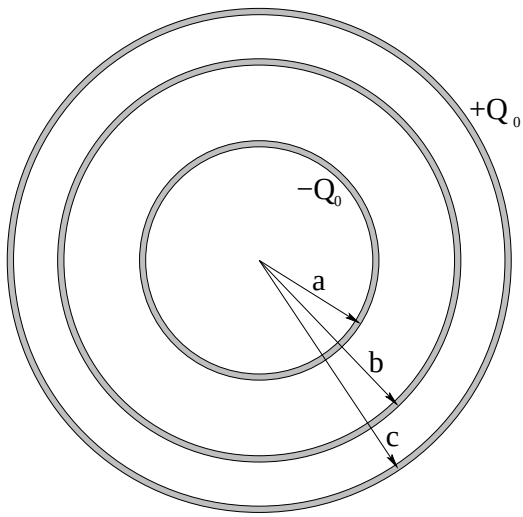
$$d\mathcal{V} = 4\pi r'^2 dr'$$

The (differential) charge in this shell is therefore:

$$dQ = \rho(r')\, d\mathcal{V}\rho(r') = 4\pi r'^2 dr'$$

where you have to substitute in the $\rho(r)$ given above for at least the second case. If you integrate this from $r' = 0 \to r$ you will have $Q(r)$, the total charge inside the radius r. Note that I use r' instead of r as the variable to integrate over so you can make r (itself a variable for the GLE part of the solution) a limit of integration – remember how that works?

Problem 10.

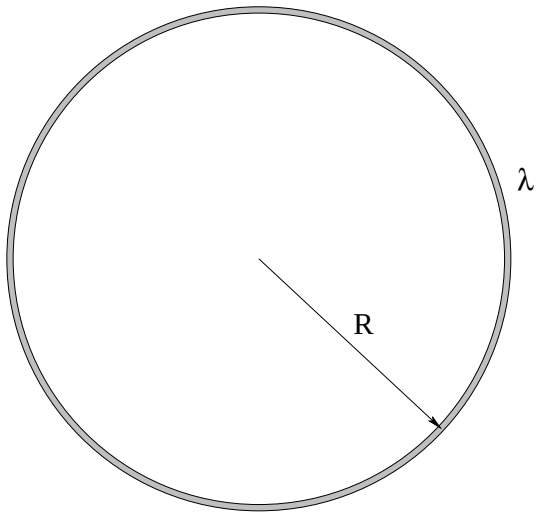


Consider three "thin" concentric conducting spherical shells with radii $a < b < c$ respectively. Initially all three shells are neutral. Then a negative charge $-Q_0$ is placed on the innermost sphere, a matching positive charge $+Q_0$ is placed on the outermost sphere, and the arrangement allowed to come to equilibrium.

a) Find the electric field everywhere and plot it. You will probably find this easier to do if you let each shell have a small (relative to a) finite thickness as drawn above.

b) Make a table showing the net charge on the inner and outer surfaces of each conducting shell.

Problem 11.



The electric field vanishes inside a uniform spherical shell of charge because the shell has exactly the right geometry to make the $1/r^2$ field produced by opposite sides of the shell cancel according to the intuition we developed from our derivation of Gauss's Law. It isn't a general result for arbitrary symmetries, however.

Consider a *ring* of charge of radius R and linear charge density λ. Pick a point P that is in the plane of the ring but not at the center.

a) Write an expression the field produced by the small pieces of arc subtended by opposed small angles with vertex P, along the line that bisects this small angle.

b) Does this field point towards the nearest arc of the ring or the farthest arc of the ring?

c) Suppose a charge $-q$ is placed at the center of the ring (at equilibrium). Is this equilibrium stable[60]?

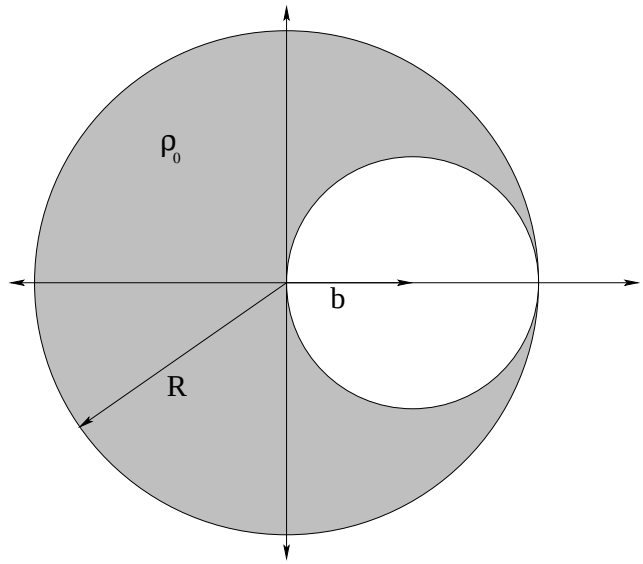
d) Suppose the electric field dropped off like $1/r$ instead of $1/r^2$. Would you expect the electric field to vanish in the plane inside of the ring? Would this be a good form for the electric field in Edwin Abbot's novel *Flatland* so that they could have a Gauss's Law too[61]?

Problem 12.

A uniformly charged nonconducting sphere of radius a is centered on the origin and has a uniform charge density $\rho(r) = \rho_0$.

[60] As a parenthetical aside, note that this is the problem with the ringworld described in Larry Niven's famous *Ringworld* series of science fiction novels, as gravitational attraction has the same form as the electrostatic attraction discussed in this problem.

[61] Alternatively, could a flatlander speculate that reality was really three dimensional because of the apparent *failure* of an expected $1/r$ force law? Questions such as this are highly relevant to modern field theorists hoping to infer extra/hidden dimensions.



a) Show that at a point within the sphere a distance r from the center the electric field is given by:

$$\vec{\boldsymbol{E}} = \frac{\rho_0 \vec{\boldsymbol{r}}}{3\epsilon_0} = \frac{4\pi k \rho_0 \vec{\boldsymbol{r}}}{3}$$

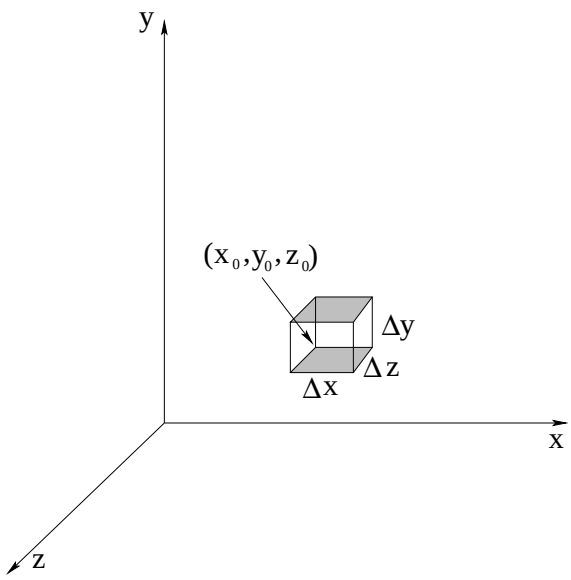
b) Material is removed from the sphere to create a spherical cavity of radius $b = a/2$ with center at $x = b$ on the x axis (shown above). Show that the electric field inside the cavity is *uniform* and equal to:

$$\vec{\boldsymbol{E}} = \frac{\rho_0 \vec{\boldsymbol{b}}}{3\epsilon_0} = \frac{4\pi k \rho_0 \vec{\boldsymbol{b}}}{3}$$

in magnitude (where $\vec{\boldsymbol{b}} = b\hat{\boldsymbol{x}}$).

Hint: By far the easiest way to attack this problem is to imagine that the "hole" is made up of a sphere of uniform charge density $-\rho_0$ and radius b that is *superposed* on the uniform sphere of charge density ρ_0 and radius a. In that way the two charge densities cancel and leave "the cavity", while you can easily find the fields using the results of part (a) with a bit of algebra. Also, *draw big pictures* of the spheres. You have to add vectors in the hole! If you don't make a big sphere with a hole large enough to draw vectors in, it's going to be really hard to visualize what's going on accurately enough to guide you when you try to add up the field. If you do a really *good* picture, you may see the *trivial* way to do the addition that actually makes this problem rather *easy* (given (a)) instead of a matter of adding up vector components the hard way!

Advanced Problem 13.



Consider a *small* gaussian surface in the shape of a cube with faces parallel to the xy, xz, and yz planes sitting in region where there is a continuous electric field. Let the corner nearest the origin be located at $\vec{r}_0 = (x_0, y_0, z_0)$ and the cube edge lengths be $\Delta x = \Delta y = \Delta z$ in the directions parallel to the different axes.

Since the electric field is continuous, each component of the field can be expanded in a Taylor series:

$$\begin{aligned}
\vec{\boldsymbol{E}}(\vec{\boldsymbol{r}}_0 + \Delta\vec{\boldsymbol{r}}) \quad = \quad & \\
& \left(E_x(\vec{\boldsymbol{r}}_0) + \Delta x \left.\frac{\partial E_x}{\partial x}\right|_{\vec{\boldsymbol{r}}_0} + \Delta y \left.\frac{\partial E_x}{\partial y}\right|_{\vec{\boldsymbol{r}}_0} + \right. \\
& \qquad \left. \Delta z \left.\frac{\partial E_x}{\partial z}\right|_{\vec{\boldsymbol{r}}_0} + \ldots \right) \hat{\boldsymbol{x}} + \\
& \left(E_y(\vec{\boldsymbol{r}}_0) + \Delta x \left.\frac{\partial E_y}{\partial x}\right|_{\vec{\boldsymbol{r}}_0} + \Delta y \left.\frac{\partial E_y}{\partial y}\right|_{\vec{\boldsymbol{r}}_0} + \right. \\
& \qquad \left. \Delta z \left.\frac{\partial E_y}{\partial z}\right|_{\vec{\boldsymbol{r}}_0} + \ldots \right) \hat{\boldsymbol{y}} + \\
& \left(E_z(\vec{\boldsymbol{r}}_0) + \Delta x \left.\frac{\partial E_z}{\partial x}\right|_{\vec{\boldsymbol{r}}_0} + \Delta y \left.\frac{\partial E_z}{\partial y}\right|_{\vec{\boldsymbol{r}}_0} + \right. \\
& \qquad \left. \Delta z \left.\frac{\partial E_z}{\partial z}\right|_{\vec{\boldsymbol{r}}_0} + \ldots \right) \hat{\boldsymbol{z}} +
\end{aligned} \tag{2.102}$$

where we only keep/show first order terms.

Noting that $\Delta A = \Delta x \Delta y = \Delta x \Delta z = \Delta z \Delta y$ (depending on the side) and that $\Delta V =$

$\Delta x \Delta y \Delta z$, show that the net electric flux *out* of this box is:

$$\sum_{\text{sides}} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}}\, \Delta A = \phi_{\text{net}} = \left(\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} \right) \Delta V = \vec{\boldsymbol{\nabla}} \cdot \vec{\boldsymbol{E}}\, \Delta V$$

Note well, to get this result you need to eliminate certain components in the full expansion. To accomplish this, you will need to neglect any term that is ***second order*** in Δx, Δy, or Δz.

This is justified by taking the differential limit: $\Delta x \to dx$, etc. Then Gauss's Law as we have thus far learned it becomes the following vector differential form:

$$\sum_{\text{sides}} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}}\, dA = \vec{\boldsymbol{\nabla}} \cdot \vec{\boldsymbol{E}}\, dV = \frac{\rho}{\epsilon_0} dV$$

or

$$\vec{\boldsymbol{\nabla}} \cdot \vec{\boldsymbol{E}} = \frac{\rho}{\epsilon_0} \tag{2.103}$$

Congratulations! You've just derived Gauss's Law in its ***vector differential*** form (and, incidentally, have derived the divergence theorem for vector fields if we extend the sums above back to integrals by summing over all the little differential cubes in an extended volume with interior surface contributions cancelling out). We won't use this this semester, but it is very important to *start* to think about how the one (integral) form is equivalent to the other (differential) form, as the latter turns out to be very useful!

Week 3: Potential Energy and Potential

- The change in electrostatic potential energy moving a charge between two points in the field of other charges is:

$$\Delta U(\vec{\boldsymbol{x}}_0 \to \vec{\boldsymbol{x}}_1) = -\int_{\vec{\boldsymbol{x}}_0}^{\vec{\boldsymbol{x}}_1} \vec{\boldsymbol{F}} \cdot d\vec{\boldsymbol{x}} \tag{3.1}$$

  where $\vec{\boldsymbol{F}}$ is the total force due to all other charges.

- The vector electrostatic force can be found from the the potential energy function by taking its negative *gradient*:

$$\vec{\boldsymbol{F}} = -\vec{\boldsymbol{\nabla}} U \tag{3.2}$$

- For charge density distributions with "compact support" (ones we can draw a ball around, basically) we by convention define the zero of the potential energy function to be at ∞:

$$U(\vec{\boldsymbol{x}}) = -\int_{\infty}^{\vec{\boldsymbol{x}}} \vec{\boldsymbol{F}} \cdot d\vec{\boldsymbol{x}} \tag{3.3}$$

  For point charges q_1 and q_2, it is just:

$$U(\vec{\boldsymbol{x}}_1, \vec{\boldsymbol{x}}_2) = \frac{kq_1q_2}{|\vec{\boldsymbol{x}}_1 - \vec{\boldsymbol{x}}_2|} \tag{3.4}$$

- Since the potential energy is just a scalar and satisfies the superposition principle, we can evalute the total energy of a system of point charges as:

$$U_{\text{tot}} = \frac{1}{2}\sum_{i \neq j} \frac{kq_iq_j}{|\vec{\boldsymbol{x}}_i - \vec{\boldsymbol{x}}_j|} \tag{3.5}$$

  (there is a similar integral expression for continuous charge distributions we will address later) where the 1/2 is to compensate for double counting in the sum.

- The electrostatic *potential* produced by a charge q is a one-body scalar field defined by:

$$V(\vec{\boldsymbol{x}}) = \lim_{q_0 \to 0} \frac{U(\vec{\boldsymbol{x}})}{q_0} \tag{3.6}$$

  so that the potential of a point charge in coordinates centered on the charge is just:

$$V(\vec{\boldsymbol{r}}) = \frac{kq}{r} \tag{3.7}$$

- The potential is to the field as the potential energy is to the force, so:

$$V(\vec{\boldsymbol{x}}) = -\int \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{x}} + V_0 \tag{3.8}$$

with V_0 and arbitrary constant of integration, used to set a suitable zero of the potential energy. For compact charge distributions:

$$V(\vec{\boldsymbol{x}}) = -\int_{\infty}^{\vec{\boldsymbol{x}}} \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{x}} \tag{3.9}$$

and

$$\vec{\boldsymbol{E}} = -\vec{\boldsymbol{\nabla}} V \tag{3.10}$$

- The potential of a charge distribution can obviously be evaluated by superposition:

$$V_{\text{tot}}(\vec{\boldsymbol{x}}) = \sum_i \frac{kq_i}{|\vec{\boldsymbol{x}} - \vec{\boldsymbol{x}}_i|} \tag{3.11}$$

or

$$V_{\text{tot}}(\vec{\boldsymbol{x}}) = \int \frac{kdq_0}{|\vec{\boldsymbol{x}} - \vec{\boldsymbol{x}}_0|} = \int \frac{k\rho(\vec{\boldsymbol{x}}_0)d^3r_0}{|\vec{\boldsymbol{x}} - \vec{\boldsymbol{x}}_0|} \tag{3.12}$$

- Conductors at electrostatic equilibrium are *equipotential*. We can therefore speak of the *potential difference* between two conductors in electrostatic equilibrium where it doesn't matter what path we use to go from one conductor to the other. This also means that if we charge one isolated conductor to some potential and then connect it to another isolated conductor, charge will flow until the two conductors (now one) are at the *same* potential, a process called *charge sharing*.

- In a strong enough electric field, *dielectric breakdown* occurs and insulators "suddenly" become conductors (e.g. lightning in air). Strong fields are often induced in the vicinity of a sharp conducting point, causing a slower *corona effect* discharge that is the basis for lightning rods.

This completes the chapter/week summary. The sections below illuminate these basic facts and illustrate them with examples.

3.1: Electrostatic Potential Energy

The electrostatic force is *conservative*. That is, the work done moving a charge between any two points in an electrostatic field is independent of the path taken. For conservative forces we can define the *change in potential energy* to be the negative work done by the electrostatic force moving between two points:

$$\Delta U(\vec{\boldsymbol{x}}_0 \to \vec{\boldsymbol{x}}_1) = -\int_{\vec{\boldsymbol{x}}_0}^{\vec{\boldsymbol{x}}_1} \vec{\boldsymbol{F}} \cdot d\vec{\boldsymbol{x}} \tag{3.13}$$

The corresponding relation between the potential energy thus defined and the force is (as usual):

$$\vec{\boldsymbol{F}} = -\vec{\boldsymbol{\nabla}} U \tag{3.14}$$

Consequently we see that we could equally well define the electrostatic potential energy in terms of an *indefinite* integral and an *arbitrary constant of integration*:

$$\Delta U(\vec{\boldsymbol{x}}) = -\int \vec{\boldsymbol{F}} \cdot d\vec{\boldsymbol{x}} + U_0 \tag{3.15}$$

that effectively sets the point where the potential energy is zero.

By convention, for charge densities that have *compact support* – ones that one can draw a ball of finite radius (however large that radius might be) so that it *completely contains* all of the charge – we define the potential energy to be zero at ∞, just as we did for the gravitational potential energy:

$$\Delta U(\vec{\boldsymbol{x}}) = -\int_{\infty}^{\vec{\boldsymbol{x}}} \vec{\boldsymbol{F}} \cdot d\vec{\boldsymbol{x}} \tag{3.16}$$

(so that U_0 is zero, if you prefer). We remain free to choose a different zero, however, in any problem where doing so is computationally convenient.

Using the relations above, it is easy to show that the potential energy of two point charges is:

$$U = \frac{kq_1q_2}{|\vec{\boldsymbol{x}}_1 - \vec{\boldsymbol{x}}_2|} \tag{3.17}$$

which again looks very much like that for gravity as might be expected.

One important advantage of working with the potential energy is that it is a *scalar*. To find the total potential energy of a collection of charges, we just *add it up pairwise*:

$$U_{\text{tot}} = \frac{1}{2}\sum_{i\neq j} \frac{kq_iq_j}{|\vec{\boldsymbol{x}}_i - \vec{\boldsymbol{x}}_j|} \tag{3.18}$$

Note that in this sum the $1 \to 2$ interaction is counted *twice*, once as q_1q_2 and once as q_2q_1. We only wish to count it once, so we divide the result by 1/2. Another way to deal with this issue is to order the sum so that we simply never do a pair twice:

$$U_{\text{tot}} = \sum_{i<j} \frac{kq_iq_j}{|\vec{\boldsymbol{x}}_i - \vec{\boldsymbol{x}}_j|} \tag{3.19}$$

This stands for "sum over all q_j and all q_i such that $i < j$" which excludes all the self-energy $i = j$ terms. Good thing, too, since they are all infinite!

3.2: Potential

The good thing about potential energy is that it is a scalar and easier to evaluate than the *vector* force or field. However, it isn't terribly easy! It is still a two-body interaction term and requires us to do a nasty double sum (that becomes an even nastier double integral) when we have a large collection of charges.

A couple of weeks ago we introduced the idea of the *field* to eliminate two body computations for electric force and to give us the comfort of an apparent action-at-a-distance *cause* of the electric force. Let us do exactly the same thing here. We will define the electrostatic

potential to be a scalar field of "potential energy per unit charge" that is the *cause* of a charged particle placed in it having a potential energy.

The formal definition of the potential is that it is the potential energy of a small test charge q_0 interacting with all the other charges that create the potential, per unit test charge, in the limit that this small test charge vanishes:

$$V(\vec{\boldsymbol{x}}) = \lim_{q_0 \to 0} \frac{U(\vec{\boldsymbol{x}})}{q_0} \tag{3.20}$$

Note that this strange-seeming condition ensures that the test charge itself doesn't perturb the charge distribution that produces the potential.

The SI units for potential are:

$$1 \text{ Volt} = \frac{1 \text{ Joule}}{1 \text{ Coulomb}} \tag{3.21}$$

If we apply this rule compute the potential at $\vec{\boldsymbol{x}}$ produced by a point charge q at the origin of coordinates, we get:

$$V(\vec{\boldsymbol{x}}) = \lim_{q_0 \to 0} \frac{1}{q_0} \frac{kqq_0}{|\vec{\boldsymbol{x}} - 0|} = \frac{kq}{r} \tag{3.22}$$

where $r = |\vec{\boldsymbol{x}}|$. Alternatively we could use the definition of the field relative to the force to define:

$$V(\vec{\boldsymbol{x}}) = -\int \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{x}} + V_0 \tag{3.23}$$

For charge distributions with compact support, we by convention pick the zero of potential at ∞ so that:

$$V(\vec{\boldsymbol{x}}) = -\int_{\infty}^{\vec{\boldsymbol{x}}} \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{x}} \tag{3.24}$$

In many cases (especially when we start to treat conductors more thoroughly in later chapters) we will be interested in *potential differences*. If the field is known and well behaved, they can be easily computed by means of:

$$\Delta V(\vec{\boldsymbol{x}}_1 \to \vec{\boldsymbol{x}}_2) = -\int_{\vec{\boldsymbol{x}}_1}^{\vec{\boldsymbol{x}}_2} \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{x}} \tag{3.25}$$

We can invert these relations to obtain:

$$\vec{\boldsymbol{E}} = -\vec{\boldsymbol{\nabla}} V \tag{3.26}$$

which in some cases will give us a relatively easy path to find the field. If the potential is relatively easy to find by (say) superposition (because it is a straight scalar sum or integral over the potentials of all the contributing charges) then one can find the field by doing relatively easy derivatives instead of sums or integrals over vector components.

Note that this relation gives us a new way to write the strength of a field in SI units as volts per meter. Note also that there is a precise analogy between force and potential energy and field and potential. Finally, note that once we know the potential produced by a collection of

fixed charges, we can compute the potential energy of a charge q placed in the potential *subject to the condition* that the presence of the charge in the potential does not cause significant rearrangement of the charges that create that potential as:

$$U = qV \tag{3.27}$$

This will not always be the case! In fact, if we were picky we'd say that it is almost never the case in nature, because atoms aren't "solid" objects and inevitably distort in the presence of the field of the perturbing charge. However, that doesn't really stop us from using this expression; we merely have to compute the potential energy in the *self-consistent* perturbed potential of the other charges. It does make it a bit more difficult, though.

3.3: Superposition

As we noted in the previous section, a major motivation for introducing potential is that it is a scalar quantity that we can evaluate by doing sums that don't involve the complexity of vector components or charge-charge interactions. The rule for finding the potential of a collection of charges is simple: We just add up the scalar potential of each (point-like) charge independent of all the rest!

This is once again the *superposition principle* for electrostatics, now applied to the scalar potential:

$$V_{\rm tot}(\vec{\boldsymbol{x}}) = \sum_i \frac{kq_i}{|\vec{\boldsymbol{x}} - \vec{\boldsymbol{x}}_i|} \tag{3.28}$$

In words, the potential at a point in space is the simple (scalar) sum of the individual potentials of all the charges that contribute to that total potential.

As before, when we are working at scales where there are many many elementary point charges contributing to the potential, we can coarse grain average. That is, we can look at a volume ΔV that is large enough to contain sufficient charge for a smooth average charge density to result that is also small enough that we can sum over it as if it is the integration volume element dV (or ditto for surface or linear distributions with elements dA and dx respectively).

Then the sum becomes:

$$\begin{aligned} V_{\rm tot}(\vec{\boldsymbol{x}}) &= \int \frac{k\ dq_0}{|\vec{\boldsymbol{x}} - \vec{\boldsymbol{x}}_0|} \\ &= \int \frac{k\ \rho(\vec{\boldsymbol{x}}_0)\ d^3r_0}{|\vec{\boldsymbol{x}} - \vec{\boldsymbol{x}}_0|} \qquad \text{volume} && (3.29) \\ &= \int \frac{k\ \sigma(\vec{\boldsymbol{x}}_0)\ d^2r_0}{|\vec{\boldsymbol{x}} - \vec{\boldsymbol{x}}_0|} \qquad \text{area} && (3.30) \\ &= \int \frac{k\ \lambda(\vec{\boldsymbol{x}}_0)\ dr_0}{|\vec{\boldsymbol{x}} - \vec{\boldsymbol{x}}_0|} \qquad \text{line} && (3.31) \end{aligned}$$

3.3.1: Deriving or Computing the Potential

The rules above give us two distinct ways to evaluate the potential in any given problem, and we must look at the problem carefully to assess which one is best.

a) If the field is known, varies only in one dimension, and is integrable in some system of coordinates, we can integrate

$$-\int E_x dx$$

to find the potential. For all practical purposes in this course, problems involving the symmetric distributions of charge whose fields we can find using Gauss's Law are precisely the ones where it is likely to be most convenient to evaluate the potential in this way.

It is *necessary* to use this approach to find the potential differences of a non-compact charge density distribution such as an infinite line or infinite sheet. This is because the sum of the potential of an infinite amount of charge (however it is distributed) is infinite, which is in turn why we restrict the use of the superposition forms of the potential that vanish at ∞ to compact charge distributions.

b) If the field is not known or discoverable from Gauss's Law and/or is not "one dimensional" in the sense that we can easily find a line to integrate over where the vector components of the field don't enter in a non-trivial way, we will probably be better off computing the field directly from the superposition principle – summing or integrating all of the contributions to the potential from all the point charges or point-like elements of a charge distribution to find the total.

Note that both of these approaches *will yield the same answer* for charge distributions with compact support within the inevitable constant V_0 for *all* problems to which they are consistently applied. In fact, even for non-compact distributions they will yield the same answer for the part that varies with the coordinates of the point once one "renormalizes" the limiting form of the superposition answer by subtracting the appropriate infinite constant. That's because the negative gradient of the two forms must, of course, return the same field!

3.4: Examples of Computing the Potential

Example 3.4.1: Potential of a Dipole on the x-axis

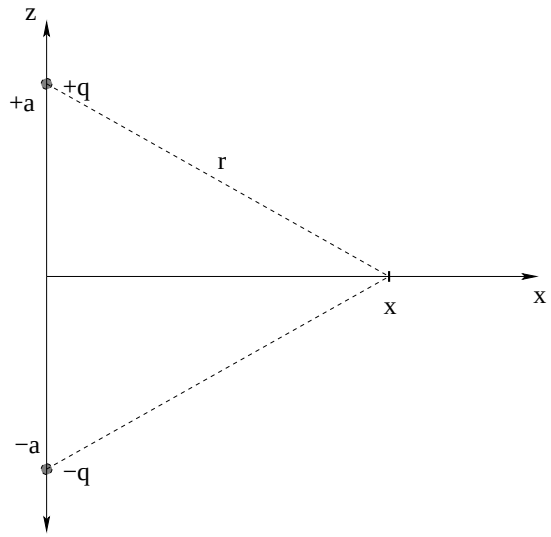


Figure 3.1: A simple dipole aligned with the z-axis.

> This is the same dipole studied in the the chapter on field. Find the *potential* at an arbitrary point on the x-axis.

This problem is deceptively simple. We know from the superposition principle that the potential is:

$$\begin{aligned} V(x) &= \sum_{i=1}^{2} \frac{k_e q_i}{r_i} \\ &= \frac{k_e q}{(x^2+a^2)^{1/2}} - \frac{k_e q}{(x^2+a^2)^{1/2}} = 0 \end{aligned} \tag{3.32}$$

This is absolutely correct – the potential of a dipole vanishes on the *entire plane* that symmetrically bisects the line connecting the charges.

The "deception" occurs when we try to compute the *field* by using $\vec{\boldsymbol{E}} = -\vec{\boldsymbol{\nabla}} V$. We are ever so tempted to go e.g.:

$$E_z = -\frac{dV}{dz} = -\frac{d0}{dz} = 0 \tag{3.33}$$

which is simple, easy, and *wrong!* The problem is that even though the function $V(x, y, z)$ is zero at a point that does *not* mean that its *slope* is zero at the point! We have to use L'Hopital's Rule to evaluate a derivative at a point where its lower order derivatives or value are zero.

What this means is that we have to evaluate the function for $V(x, y, z)$ *near* but not *on* the point where the function is zero, take the desired derivative, and then let the parameter that describes that nearness go to zero. In this case, we need to find $V(x, z)$ for some *small* z (near zero), take the derivative, and let the value of z in the derivative go to zero. See if you

can draw pictures to verify the following algebra, for a point $z \ll a \ll x$ above the point on the x-axis.

$$V(x,z) = \frac{k_e q}{(x^2 + (a-z)^2)^{1/2}} - \frac{k_e q}{(x^2 + (a+z)^2)^{1/2}} \tag{3.34}$$

Now we can differentiate:

$$\begin{aligned} E_z &= -\frac{d}{dz}\frac{k_e q}{(x^2 + (a-z)^2)^{1/2}} + \frac{d}{dz}\frac{k_e q}{(x^2 + (a+z)^2)^{1/2}} \\ &= -\frac{k_e q(a-z)}{(x^2 + (a-z)^2)^{3/2}} - \frac{k_e q(a+z)}{(x^2 + (a+z)^2)^{1/2}} \end{aligned} \tag{3.35}$$

NOW we can let $z \to 0$ to find out what the field is on the x-axis (adding and cancelling terms as necessary, and substituting $p_z = 2qa$ in for the dipole moment):

$$\begin{aligned} E_z &= -\frac{2k_e qa}{(x^2 + a^2)^{3/2}} \\ &= -\frac{k_e p_z}{(x^2 + a^2)^{3/2}} \end{aligned} \tag{3.36}$$

Compare this to equation (1.26)! Hmmm, looks the same[62]! And it wasn't *that* difficult, although it was certainly more difficult than we might have expected. To see how really *easy* it was, consider. We actually just obtained the *exact* E_z field for all points in space, since the answer is azimuthally symmetric and we could rotate the answer to tell us the field in planes other than the xz plane! And the E_x field is equally easy to find.

It will turn out that Cartesian coordinates suck in so many ways when doing physics problems. Physics is if anything naturally spherical or cylindrical – nature is only rarely rectilinear. Let's redo the potential problem above, but not let's find the potential at an *arbitrary point in space* in *spherical polar coordinates*. Remember, the math section has a lovely little review of Cartesian, Cylindrical and Spherical coordinate systems – the big three one needs to work with in this course – in case you have never seen spherical coordinates before (or don't remember them, effectively the same thing).

[62] Allowing, of course, for the change in the name of the vertical axis...

Example 3.4.2: Potential of a Dipole at an Arbitrary Point in Space

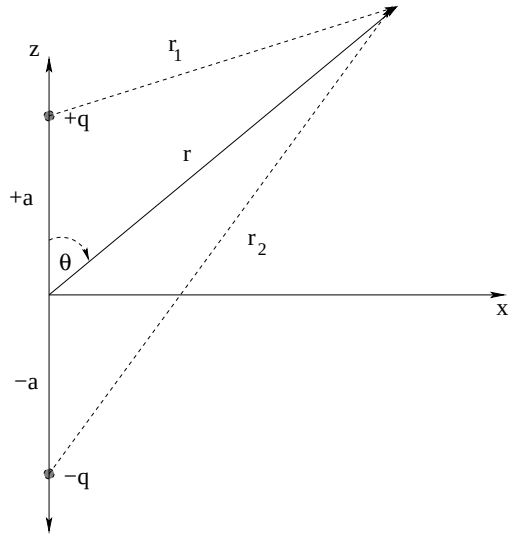


Figure 3.2: A simple dipole aligned with the z-axis, in a spherical coordinate system.

Find the potential of this dipole at an arbitrary point $P = (r, \phi, \theta)$. Because the problem is manifestly *azimuthally symmetric* the answer cannot depend in any way on ϕ (the azimuthal/longitude coordinate), so we might as well label the point $P = (r, \theta)$ in the plane of the figure, where the answer can be azimuthally rotated by ϕ about the z-axis to any other plane without changing the form of the answer.

The potential in this problem is extremely easy to find *if you can remember the law of cosines:*

$$r_1 = +\sqrt{r^2 + a^2 - 2ar\cos(\theta)} \tag{3.37}$$

$$r_2 = +\sqrt{r^2 + a^2 + 2ar\cos(\theta)} \tag{3.38}$$

so that the potential can be read off by inspection:

$$V(r, \theta) = \frac{k_e q}{(r^2 + a^2 - 2ar\cos(\theta))^{1/2}} - \frac{k_e q}{(r^2 + a^2 + 2ar\cos(\theta))^{1/2}} \tag{3.39}$$

Of course, if you *don't* remember the law of cosines, you should visit the math chapter and learn to derive it in two or three lines so you don't ever forget it again, as we will use it fairly often and you don't want this to be an obstacle to your learning!

To find the field *now*, one can take the gradient of this exact result. However, actually taking gradients is beyond the immediate scope of this course, so just bear in mind that you *can* (and if you are a physics major, almost certainly sooner or later *will*) and otherwise forget it. Doing so isn't particularly simple in any event because of the fairly complicated denominators (although it is still much easier than finding the field directly).

Consider what happens, though, when one looks at the potential at a point $r \gg a$, so far away that the dipole looks like a "point object". To find the potential then, we must use the

binomial expansion to factor out the leading r dependence and to move the complicated stuff from the denominator to the numerator (losing the square roots in the process). That is:

$$\begin{aligned}
\lim_{r\gg a} V(r,\theta) &= \frac{k_e q}{(r^2+a^2-2ar\cos(\theta))^{1/2}} - \frac{k_e q}{(r^2+a^2+2ar\cos(\theta))^{1/2}} \\
&= \frac{k_e q}{r}\left\{(1-2\frac{a}{r}\cos(\theta)+\frac{a^2}{r^2})^{-1/2} - (1+2\frac{a}{r}\cos(\theta)+\frac{a^2}{r^2})^{-1/2}\right\} \\
&= \frac{k_e q}{r}\left\{(1+\frac{a}{r}\cos(\theta)-\frac{a^2}{2r^2}+...) - (1-\frac{a}{r}\cos(\theta)-\frac{a^2}{2r^2}+...)\right\} \\
&= \frac{k_e q}{r}\left\{2\frac{a}{r}\cos(\theta)+\mathcal{O}\left(\frac{a^3}{r^3}\right)\right\} \\
&\approx \frac{k_e 2qa}{r^2}\cos(\theta) \\
&\approx \frac{k_e p_z}{r^2}\cos(\theta) = k_e\frac{\vec{\boldsymbol{p}}\cdot\hat{\boldsymbol{r}}}{r^2}
\end{aligned} \tag{3.40}$$

where $\hat{\boldsymbol{r}}$ is a unit vector in the $\vec{\boldsymbol{r}}$ direction. (We used our freedom to rotate the coordinate system so that $\vec{\boldsymbol{p}}$ points in an arbitrary direction instead of $\vec{\boldsymbol{z}}$ to guess the last result.)

This *is* a very simple form and is a very important one as well! This last equation is the ***completely general potential of a point dipole*** at a point $P = (r,\theta,\phi)$ measured relative to the dipole center (and with θ measured from the dipole axis). Note that the answer is azimuthally symmetric and doesn't depend on ϕ, as one expects. Taking the gradient of *this* to find the field (when you eventually try it) is actually pretty *easy*.

We dwell so much on dipoles because they are the most common and important microscopic configuration of charge that produces fields outside of atoms. Atoms are roughly spherically symmetric and tend to be electrically neutral in isolation. However, atoms are easily *polarized* by any applied field, including molecular fields. There are molecules (such as the ubiquitous water molecule) that have permanent electric dipole moments. Speaking as one big bag of (mostly) water to another, those little electric dipoles can organize in some pretty amazing ways! We will continue to explore dipole models until we wrap the whole notion up as a macroscopic property of matter called its *dielectric permittivity* in the next chapter.

From these two examples it should be simple enough to find the potential at a point due to any reasonable number of discrete charges provided only that you can do the coordinate geometry needed to find the distance(s) from the charges to the point of observation. The pythagorean theorem, the (more general) law of cosines: things like that are thus your best friends in evaluating potentials of point charges because once you know the distances you just sum $k_e q/r$ for all of those charges.

It's a bit harder to do a continuous distribution of charge. Let's look at a couple of continuous problems and move on to using the field itself (evaluated with Gauss's Law) to integrate to the potential or potential difference.

Example 3.4.3: A ring of charge

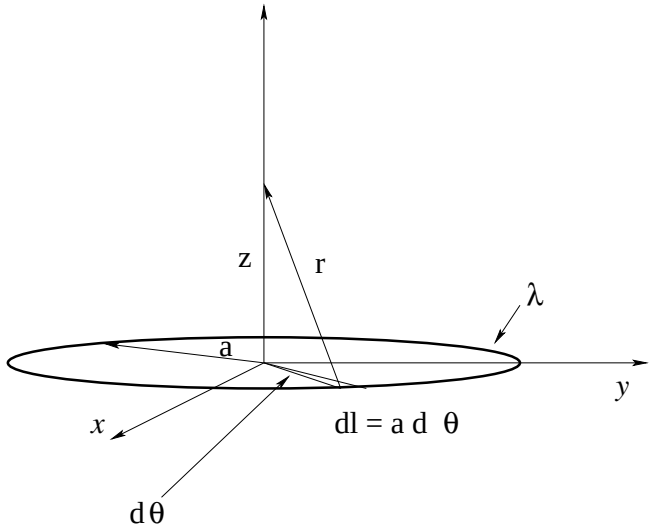


Figure 3.3: A ring of charge in the xy-plane, concentric with the z-axis.

Suppose you are given a ring of charge with charge per unit length λ and radius a on the xy-plane concentric with the z-axis. Find the potential at an arbitrary point on the z axis.

Although there is a quick and easy answer to this problem (that will be apparent at the end, if not at the beginning) we will work through this problem in detail to illustrate the general methodology of finding a potential by integrating over a continuous distribution of charge. The steps are:

a) In suitable coordinates, define a differential "chunk" of the charge. In this problem, that would be a differential-size arc segment of the ring.

b) Determine the differential charge of the chunk as "the charge of the chunk is the charge per unit whatever times the differential whatever of the chunk" where 'whatever' might be length, area or volume (in this case length).

c) Write a simple expression in suitable coordinates for the differential *potential* produced at the point of interest by the differential (point-like) chunk of charge:

$$dV = \frac{k_e\, dq}{r}$$

where r is the distance from the chunk to the point of observation. Note well that this is a *scalar* integral, making it relatively simple!

d) Integrate both sides. The left hand side becomes $V(\vec{r})$ at the point of observation (in suitable coordinates). The right hand side becomes the algebraic expression of the potential (the answer).

e) Simplify, if appropriate or required.

f) If one wishes to find the field from the potential, remember e.g.

$$E_z = -\frac{dV}{dz}$$

Beware L'Hopital's Rule! That is, if differentiating someplace that the function itself vanishes (or its functional dependence on certain coordinates vanishes) be sure that you differentiate at a general point *near* the limit point and *then* take the limit!

Let's step through this.

$$dl = a\ d\theta \tag{3.41}$$

defines a differential chunk of the ring. Its charge is:

$$dq = \lambda\ dl \tag{3.42}$$

The differential potential of this chunk at a point on the z-axis is:

$$dV(z) = \frac{k_e\ dq}{r} = \frac{k_e \lambda a\ d\theta}{(z^2 + a^2)^{1/2}} \tag{3.43}$$

We integrate over all of the chunks of charge that make up the ring by integrating θ from 0 to 2π:

$$\begin{aligned} V(z) = \int dV &= \int_0^{2\pi} \frac{k_e \lambda a\ d\theta}{(z^2 + a^2)^{1/2}} \\ &= \frac{k_e (2\pi a)\lambda}{(z^2 + a^2)^{1/2}} \\ &= \frac{k_e Q}{r} \end{aligned} \tag{3.44}$$

where we used the fact that $2\pi a\lambda = Q$, the *total charge of the ring!*

This final answer we can easily *understand* and might have even guessed without doing an integral. *All* of the charge of the ring is the *same distance r* from the point of observation, and potential depends *only* on this distance (not on direction) so the potential is just k_e times the total charge divided by that distance.

If we do indeed try to find the electric field by differentiating this last result:

$$\begin{aligned} E_z &= -\frac{d}{dz}\frac{k_e(2\pi a)\lambda}{(z^2 + a^2)^{1/2}} \\ &= \frac{k_e(2\pi a)\lambda z}{(z^2 + a^2)^{3/2}} \\ &= \frac{k_e Q z}{(z^2 + a^2)^{3/2}} \end{aligned} \tag{3.45}$$

Compare this to equation (2.17) above. Hmmm, looks like they are the same! However, evaluating the potential integral and then taking its derivative seems (to me, at any rate) to be *much easier* than doing the integral to find the field directly, with all of its components, and that's *before* we evaluated the E_x and E_y fields explicitly.

Note that we can exploit the insight we gained from this problem in a variety of ways to answer certain questions concerning the potential "by inspection". For example:

- A ring of charge Q a distance $R = (a^2 + z^2)^{1/2}$ from the point of observation;
- An arc of charge Q that has angular width θ and radius R, at the center of curvature;
- A hemispherical shell of charge Q with a radius R, at the center of the (hemi)sphere;
- Six charges each with charge $Q/6$ arranged in a hexagon that has a distance $2R$ between opposing corners, at the center;
- A single charge Q a distance R from the point of observation;

all produce a potential $k_e Q/R$ at the point of observation indicated! In all these cases a total charge of Q is arranged in various ways a distance R from the point of observation. In potential direction doesn't matter, so all of the potentials of all of the charges that make up these systems add to the one simple result.

Example 3.4.4: Potential of a Spherical Shell of Charge

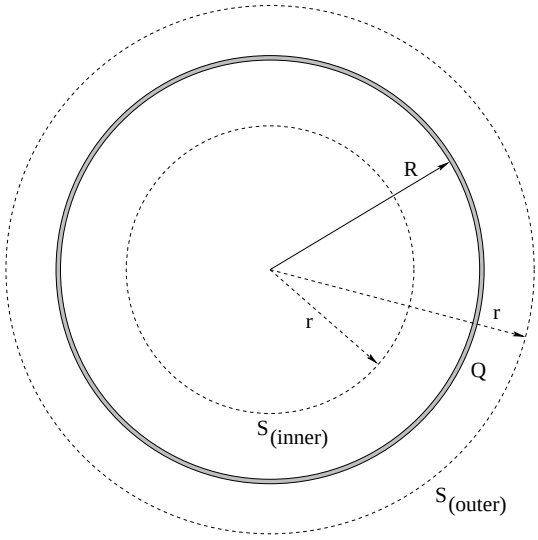


Figure 3.4: A spherical shell of charge of radius R.

> Suppose you are given a spherical shell of radius R of uniformly distributed charge Q. Find the field and the potential at all points in space.

If we want to find the potential produced by a spherical shell (or other spherical distribution of charge) and try to find it by direct integration of the potential of all the charges that make up the shell, we'll quickly discover that while it is easy to write down the integral we need to solve in some system of coordinates, it isn't so easy to *do* the integral. It's still possible – good students of calculus or students who just want a challenge can tackle it with a reasonable chance of success – but it isn't terribly easy. It's a *useful* example, though, useful enough that I include it in the book after this "easy way" example, for those very students who want to give it a try on their own and then have some way to check or correct their work.

On the other hand, finding the *electric field* from Gauss's Law is *very* easy (and is done in detail in Week 2 above, so we won't repeat the steps here). Try it on your own to make sure that you get:

$$\begin{aligned}\vec{\boldsymbol{E}} &= 0 \qquad (r < R) \\ \vec{\boldsymbol{E}} &= \frac{k_e Q}{r^2}\hat{\boldsymbol{r}} \qquad (r > R)\end{aligned}$$

in sphere-centered spherical coordinates. We recall that the potential of any charge distribution with compact support can be found from the field by directly integrating the field according to:

$$V(\vec{\boldsymbol{r}}) = -\int_{\infty}^{\vec{\boldsymbol{r}}} \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{l}} \tag{3.46}$$

In this case, we integrate piecewise from the outside in to find the field outside and inside of the sphere, accordingly. Outside:

$$V(\vec{\boldsymbol{r}}) = -\int_{\infty}^{r} \frac{k_e Q}{r^2} dr = \frac{k_e Q}{r} \tag{3.47}$$

for all $r > R$. Inside:

$$V(\vec{\boldsymbol{r}}) = -\int_{\infty}^{R} \frac{k_e Q}{r^2} dr - \int_{R}^{r} 0 \; dr = \frac{k_e Q}{R} \tag{3.48}$$

which is *constant* everywhere inside the sphere! This not only makes sense, we'll make this into a *rule*. Any volume where the electrical field vanishes has a *constant potential* – we call such a region *equipotential*. We'll talk about equipotential regions below when discussing conductors in electrostatic equilibrium (which are, as you can probably already see, equipotential).

A spherical shell of charge thus produces a potential *outside* that looks like the potential of a point charge at the origin to match its field that looks like that of a point charge at the origin. *Inside*, its potential is constant, the value it had on the shell itself coming in from the outside.

Now, a bit of warning based on my many years of teaching this class. For some of you, the first time you see a problem like this on a quiz with a region where the field is zero, the Devil is going to whisper into your ear "C'mon, dude. The field in these is zero, so the potential in there must be zero too. Put down zero and let's move on." Unfortunately, if you listen to the Devil, you'll be condemned to Physics Quiz Hell, because this would be *wrong!* Remember that the electrical field is basically the derivative of the potential. The derivative of *any constant* is zero, not just the *particular* constant whose *value* is zero.

Think of it in terms of the tops of mesas, flat mountains. Anyplace that is "flat" in potential has no field. A charge placed there doesn't gain energy moving around. But that doesn't mean that the *height* of the mesa is sea-level, or that one doesn't have to climb a steep slope from sea-level to reach the flat part. Similarly, we may have to do quite a bit of work to push a test charge from infinity to the edge of a spherical shell of charge, but once we go inside the field vanishes and we can move it anywhere without doing work. The potential inside is constant, but that constant has to reflect the *total* work done coming in from infinity (per unit charge) and is not particularly likely to be *zero*.

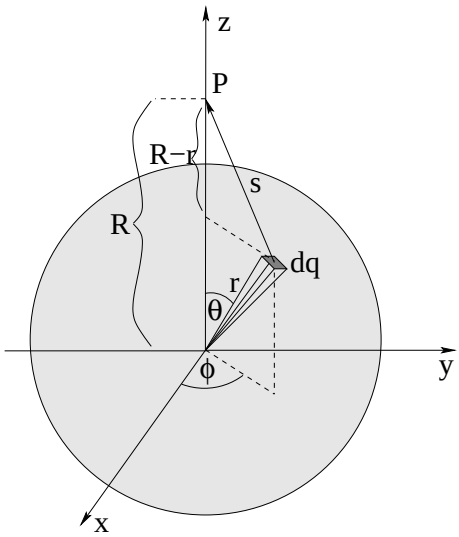


Figure 3.5: Geometry for finding the potential of a uniform spherical ***shell*** of constant charge density σ by direct integration.

Example 3.4.5: Advanced: Spherical Shell of Charge

Consider figure 3.5. You should recognize it has being almost exactly the same geometry as was used to integrate to find the (much more difficult) *electric field* of the spherical shell last week in a similarly advanced example. In a way, it would be a lot easier to just do these two examples in the opposite order, as it is a lot easier to integrate to find the potential than the field in the first place, and once we have done so we can always find the field by differentiating.

As before, we lose nothing by putting a point P at a distance R from the origin. We consider the charge dq of a tiny patch dA on the surface of the sphere, and write down the potential of this patch at P:

$$dV = \frac{k_e dq}{s} = \frac{k_e \sigma r^2 \, d\cos(\theta) \, d\phi}{(R^2 + r^2 - 2Rr\cos(\theta))^{1/2}} \tag{3.49}$$

We integrate both sides, the right hand side over the entire solid angle:

$$V = \int dV = \int \frac{k_e dq}{s} = \int_{-1}^{1} \int_{0}^{2\pi} \frac{k_e \sigma r^2 \, d\cos(\theta) \, d\phi}{(R^2 + r^2 - 2Rr\cos(\theta))^{1/2}} \tag{3.50}$$

We can do the ϕ integral immediately and factor out all the constants:

$$V = 2\pi r^2 \sigma k_e \int_{-1}^{1} \frac{d\cos(\theta)}{(R^2 + r^2 - 2Rr\cos(\theta))^{1/2}} \tag{3.51}$$

This is much easier to integrate than the vector relation of the field chapter example:

$$\begin{aligned}
V &= 2\pi r^2 \sigma k_e \int_{-1}^{1} \frac{d\cos(\theta)}{(R^2 + r^2 - 2Rr\cos(\theta))^{1/2}} \\
&= \frac{2\pi r^2 \sigma k_e}{-2Rr} \int_{-1}^{1} \frac{-2Rr d\cos(\theta)}{(R^2 + r^2 - 2Rr\cos(\theta))^{1/2}} \\
&= \frac{2\pi r^2 \sigma k_e}{-2Rr} 2\left(R^2 + r^2 - 2Rr\cos(\theta)\right)^{1/2}\Big|_{-1}^{1} \\
&= \frac{2\pi r^2 \sigma k_e}{-2Rr} 2\left((R-r) - (R+r)\right) \\
&= \frac{2\pi r^2 \sigma k_e}{-2Rr}(-2r) \\
&= \frac{k_e(4\pi r^2 \sigma)}{R} = \frac{k_e Q}{R}
\end{aligned} \tag{3.52}$$

Much, much easier!

Example 3.4.6: Potential of a Uniform Ball of Charge

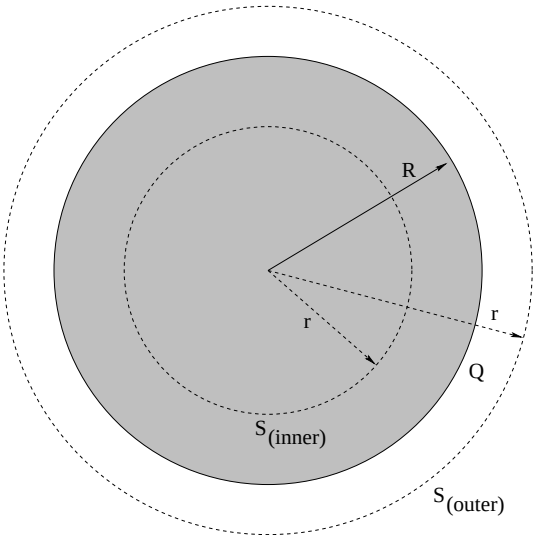


Figure 3.6: A solid sphere of uniform charge density ρ and radius R.

Find the field *and* the potential at all points in space of a solid insulating sphere with uniform charge density ρ and radius R.

If you will recall, finding the field of a solid sphere of charge is *both* an example in the text above and was a homework assignment a couple of weeks ago – so by now you should have gone over it repeatedly and made it your own. The result was:

$$E_r = \frac{k_e\left(\frac{4\pi R^3 \rho}{3}\right)}{r^2} = \frac{k_e Q}{r^2} \qquad r > R$$

and

$$E_r = k_e\left(\frac{4\pi\rho}{3}\right) r = \frac{\rho r}{3\epsilon_0} \qquad r < R$$

for the exterior and interior of the sphere (where we used $4\pi k_e = 1/\epsilon_0$ in the last equation just so you don't completely forget this relation as we prefer to work with k_e but one day you'll need to be able to work with ϵ_0). So just to humor me, get out paper and prove (to yourself, if nobody else) that you can still get this result, starting with Gauss's Law and *without looking*.

With the field(s) in hand, we now recapitulate the reasoning of the previous example. The distribution of charge has compact support, so we can integrate in from infinity to find the potential (relative to infinity):

$$\begin{aligned} V(r) &= -\int_\infty^r \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{l}} = -\int_\infty^r E_{r>R} dr \\ &= -\int_\infty^r k_e Q\, r'^{\,-2} dr' \\ &= \frac{k_e Q}{r} \qquad r > R \end{aligned} \tag{3.53}$$

and we find, as hopefully you had already anticipated, that the potential of the solid sphere *outside* was that of a point charge with the same total charge at the origin, in perfect correspondance with the field.

The place things get more interesting is when we try to evaluate the potential *inside* the sphere. The potential is defined as an integral in from ∞, but the *field changes functional form* at $r = R$. We therefore have to do the integral *piecewise*, doing first the integral from ∞ to R, then from R to r. This is why we wrote out both terms in the spherical shell example above, even though the field inside was zero (and so was that part of the integral) – we want to get in the habit of *always* doing the integral piecewise and simply being happy when one or another piece is zero, rather than either expecting it or forgetting that this is what we are really doing. Thus:

$$\begin{aligned} V(r) &= -\int_\infty^r \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{l}} = -\int_\infty^R E_{r>R} dr - \int_R^r E_{r<R} dr \\ &= -\int_\infty^R k_e \left(\frac{4\pi R^3 \rho}{3}\right) r'^{\,-2} dr' - \int_R^r k_e \left(\frac{4\pi\rho}{3}\right) r' dr' \\ &= k_e \left(\frac{4\pi R^2 \rho}{3}\right) + k_e \left(\frac{2\pi\rho}{3}\right) \left\{R^2 - r^2\right\} \\ &= 2\pi k_e \rho R^2 - k_e \left(\frac{2\pi\rho}{3}\right) r^2 \qquad r < R \end{aligned} \tag{3.54}$$

Let's think a teensy bit about this result, and then plot it (as we did for the field) to help us remember it, as (recall) the uniform ball of charge is the basis of the simplest model for an atom and hence the key to easily understanding lots of things such as polarization, ionization, and more. First of all, note that the potential is (by the meaning of integrals in the first place) the *area* under the $E_r(r)$ curve from r to ∞. $\vec{\boldsymbol{E}}$ is continuous but not smooth (look back at figure 2.14 and note the cusp at $r = R$), but $V(r)$ is continuous and *smooth* at $r = R$ – the function and its first derivative match at the point, although the second derivatives differ. Outside the potential drops off like $1/r$, a monopolar potential that corresponds to the monopolar field. Inside, the potential *increases like an upside down quadratic* all the way to the origin, where it has its maximum value!

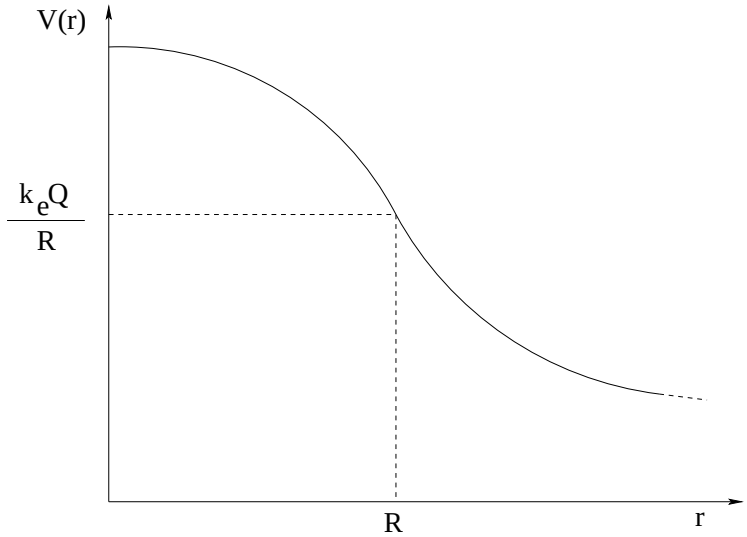


Figure 3.7: The potential produced by a uniform sphere of charge both inside and outside, as a function of r.

Example 3.4.7: Potential of an Infinite Line of Charge

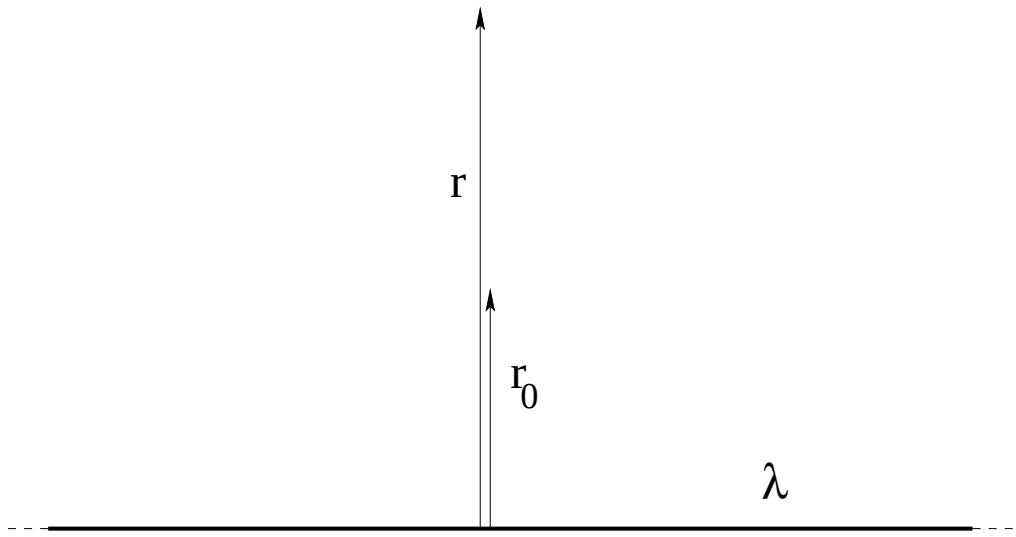


Figure 3.8: An "infinitely long" line of uniform charge density λ.

> Find the field *and* the potential *relative to the reference radius* r_0 at all points in space around an infinite line of charge. Explore the necessity of a reference point (because the indefinite integral is infinite at 0 and ∞).

As before, we will assume that you already know and can easily show that the *field* of an infinite straight line of charge is:

$$\vec{\boldsymbol{E}} = \frac{2k_e\lambda}{r}\hat{\boldsymbol{r}}$$

in cylindrical coordinates, so that $\hat{\boldsymbol{r}}$ points directly away from the line. In fact, you should be able to show this *two ways* – using Gauss's Law (very easy) and by direct integration (much harder).

We can thus equally easily write down an expression for the potential at a distance r from the line:

$$V(r) = -\int_{\infty}^{r} \frac{2k_e\lambda}{r'} dr' = -2k_e\lambda\,(ln(r) - \ln(\infty)) = \infty - 2k_e\lambda\ \ln(r) \tag{3.55}$$

Oops. Looks like our potential is *infinite*. That's a problem...

To solve it, we compute the potential not relative to infinity but to some particular radius r_0:

$$V(r) = -\int_{r_0}^{r} \frac{2k_e\lambda}{r'} dr' = -2k_e\lambda\,(ln(r) - \ln(r_0)) = -2k_e\lambda\ \ln\left(\frac{r}{r_0}\right) \tag{3.56}$$

where we use the convenient property of natural logs: $\ln(a)+\ln(b) = \ln(ab)$ to simplify the final expression. If we let $r_0 = 1$ (in whatever units we are considering this can be further simplified to:

$$V(r) = -2k_e\lambda\ \ln(r) \tag{3.57}$$

but this *obscures the units* – recall that the argument of any function with a power series expansion e.g. $\ln$ *must be dimensionless*, so the "r" in this is the *ratio* of r in the units of choice to "1" in the unit of choice. Note well that this does not matter whenever we compute *potential difference*, which is the quantity that will be the most important one in the next chapter/week:

$$\Delta V(r_1 \to r_2) = -\int_{r_1}^{r_2} \frac{2k_e\lambda}{r'} dr' = 2k_e\lambda \ln\left(\frac{r_1}{r_2}\right) \tag{3.58}$$

where the natural log is *negative* (recall) when $r_1 < r_2$ so $r_1/r_2 < 1$. This makes *sense!* Note well that the potential *decreases* when we move *away* from the line in the direction of the field (as the potential energy decreases when we move in the direction of its associated conservative force).

On your own, show that we also get this expression if we form $\Delta V(r_1 \to r_2) = V(r_2) - V(r_1)$ using *any* of the forms for $V(r)$ given above (even the one with ∞ in it, as long as we are permitted to subtract $\infty - \infty = 0$, which of course is not necessarily or generally true but which *can* be true as the setting of the zero of the potential).

3.4.1: Potential of an Infinite Plane of Charge

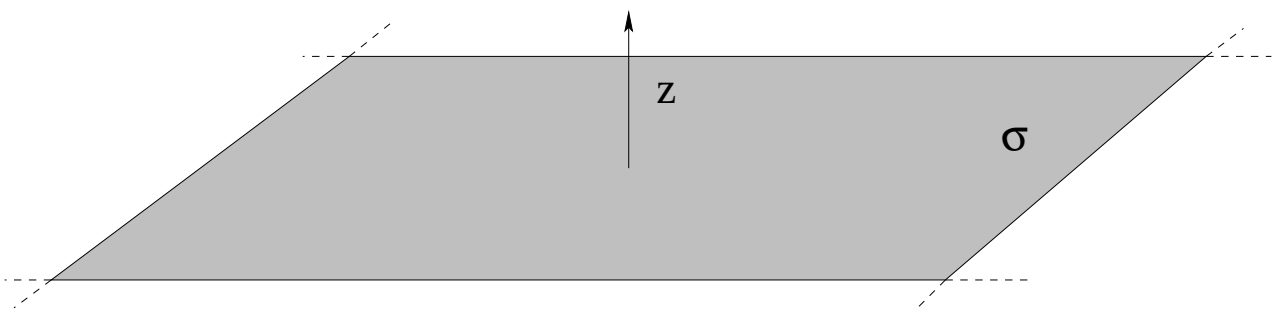


Figure 3.9: An "infinite" plane of uniform charge density σ.

Find the field *and* the potential *relative to the plane itself* at all points in space around an infinite plane of charge. Explore the necessity of a finite reference point (where e.g. $z = 0$ is the most convenient) because the potential integrated *in* from ∞ is clearly infinite.

Using Gauss's Law (or taking the limit of e.g. a disk on its axis) you can easily show that the electric field a distance z above an infinite plane of charge with charge density σ is:

$$E_z = 2\pi k_e \sigma$$

(pointing away from the plane symmetrically on both sides) independent of z. That is, the plane of charge creates a *uniform* electric field that reaches from the plane to (in principle) ∞ without change.

If we try to evaluate the potential at a finite point z relative to ∞ we get into trouble once again because the charge distribution is non-compact:

$$V(z) = -\int_{\infty}^{z} 2\pi k_e \sigma \, dz = \infty - 2\pi k_e \sigma z \tag{3.59}$$

We feel uncomfortable with infinite quantities, so we either subtract away the infinity with a new (infinite) constant of integration, or just measure the potential difference relative to some other zero. A common, and convenient one (that leads to the same result as throwing away the infinity is $z = 0$, on the plane itself. Interestingly, this is still well defined!

$$V(z) = -\int_{0}^{z} 2\pi k_e \sigma \, dz = 0 - 2\pi k_e \sigma z = -2\pi k_e \sigma z \tag{3.60}$$

Again we will most often be interested in computing potential differences rather than potentials in the subsequent chapters, especially for non-compact charge distributions. We note that the functional variation with z is such that the potential *decreases* when one moves away from the plane; this is the most important thing to keep in mind when trying to assign or check the sign of the potential (or potential difference). The field *always* points in the direction of decreasing potential.

3.5: The Potential Energy of Charge Distributions

3.5.1: The Potential Energy of Multiple Point Charges

In our initial discussion above, we went from finding the electrostatic potential energy of a charge in the field of other charges to the "potential" – a scalar field that exists at all points in space due to the presence of charges and is the "cause" of the electrostatic potential energy of a (test) charge placed at that point in space. We wrote down a couple of equations that each represented the potential energy of a collection of discrete charges using the superposition principle – we added up the potential energy for *each pair of charges in the collection* (counted only once!) to get the total potential energy:

$$U_{\text{tot}} = \frac{1}{2} \sum_{i \neq j} \frac{k_e q_i q_j}{|\vec{\boldsymbol{x}}_i - \vec{\boldsymbol{x}}_j|} = \sum_{i<j} \frac{k_e q_i q_j}{|\vec{\boldsymbol{x}}_i - \vec{\boldsymbol{x}}_j|} \tag{3.61}$$

where the inequality in the latter expresssion effectively causes us to count each ij pair only once while the former just counts them in both orders but divides by two.

One way to think about the "count each pair only once" rule is to think of the energy as being associated with a sort of *bond* between the two charges – there is only one "bond" in

between the two ends. We actually do the same thing when we think of the work stored in a stretched spring with masses at both ends – we don't count it twice just because two objects will *share* that potential energy if the masses are released.

These formulas are simple enough, to be sure, but we should without any doubt do at least one simple example just to see how this works.

Example 3.5.1: The Potential Energy of Four Charges in a Square

Suppose we have a four identical charges q arranged in a square, and would like to compute their total potential energy. There are a variety of ways we could draw a picture to represent this – one such way is presented in figure 3.10, where we place the four charges at the corners with coordinates $(\pm a, \pm a)$ in the x-y plane.

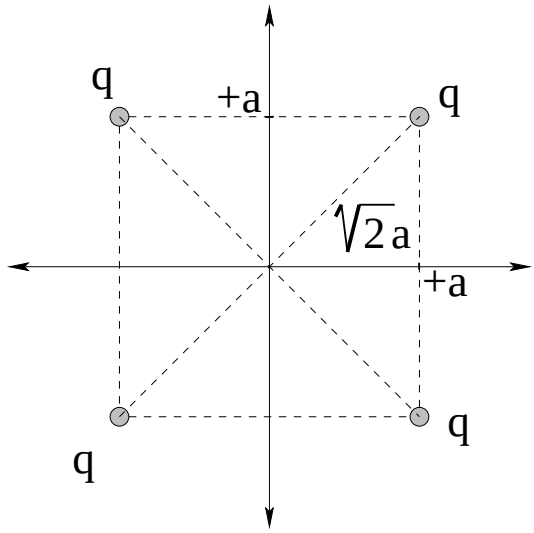


Figure 3.10: Four identical charges q arranged on the corners of a square with sides of length $2a$, centered on the origin in the x-y plane. Note that the length of a *diagonal* of this square is $2\sqrt{2}a$.

We could just write down the answer by inspection, but I'm going to walk you through a particular way of *understanding* this result. Suppose we start with no charges anywhere closer than ∞. In that case, *there is no field* anywhere in space and *it costs us no work* to move a charge q from ∞ and locate it at (say) the lower left corner.

Now we want to bring in the *second* charge, but when we do, we have to do work against the field/force produced by the charge that is already there. As we showed above, the work ***I*** have to do ***against*** that charge to end up with it at the upper left hand corner is:

$$W_{\mathrm{me},12} = U_{12} = \frac{k_e q^2}{2a} \tag{3.62}$$

where I introduced dummy indices to help us keep track of what charge we are working on and which corner we are going to put it on (1234 starting at the lower left hand corner).

To bring in a third charge, ***I*** have to do work against the fields of *both* of these charges. The work I do is *stored* as potential energy in the system, the same way the work *I* do lifting a

book of mass m off the ground by a height H *against* gravity is stored as the potential energy of the book, mgH. In this case, the work done is to put the third charge in the upper right hand corner is:

$$W_{\mathrm{me},13} + W_{\mathrm{me},23} = U_{13} + U_{23} = \frac{k_e q^2}{2\sqrt{2}a} + \frac{k_e q^2}{2a} \tag{3.63}$$

(you can seem I'm just using $k_e q^2/r$ for whatever the r is between the final resting points of the charges).

Finally, bringing in the fourth charge to the lower right hand corner is clearly:

$$W_{\mathrm{me},14} + W_{\mathrm{me},24} + W_{\mathrm{me},34} = U_{14} + U_{24} + U_{34} = \frac{k_e q^2}{2a} + \frac{k_e q^2}{2\sqrt{2}a} + \frac{k_e q^2}{2a} \tag{3.64}$$

Now when I add them, *sure*, I get exactly what I expected and could have written down directly:

$$U_{\mathrm{tot}} = \sum_{i=1}^{j-1} \sum_{j=2}^{4} \frac{k_e q_i q_j}{|\vec{\boldsymbol{x}}_i - \vec{\boldsymbol{x}}_j|} = U_{12} + U_{13} + U_{14} + U_{23} + U_{24} + U_{34} \tag{3.65}$$

where we *do not count "self-energy"*, that is, any sort of U_{ii} contribution. Self-energy is a tricky subject! We'll work on a specific model for it in a bit, and in the process encounter one of the "mysteries" of physics that points along the path to a consistent (eventually) field theory.

Summing up the terms, we get:

$$\boxed{U_{\mathrm{tot}} = 2\frac{k_e q^2}{a} + \frac{k_e q^2}{\sqrt{2}\,a}} \tag{3.66}$$

Solving the problem in this way emphasizes one critical point:

> **The potential energy of a collection of charge equals the work required to assemble it bringing the charge in from ∞.**

This will guide us as we seek to compute the potential energy of of *continuous* charge distributions.

3.5.2: The Potential Energy of Continuous Charge Distributions

We are now ready to compute the potential energy of a *continuous* distribution of charges. We'll start by generalizing the sum rules from before by coarse graining and treating each little chunk of charge as a point charge. This is illustrated in figure 3.11.

We start by applying the usual ritual to express the charges in terms of differential volumes:

$$dq_1 = \rho(\vec{\boldsymbol{r}}_1) d^3 r_1 \qquad\qquad dq_2 = \rho(\vec{\boldsymbol{r}}_2) d^3 r_2 \tag{3.67}$$

where $d^3 r$ is another way of writing "the volume element in three dimensions for coordinate $\vec{\boldsymbol{r}}$". We use it because in this particular context, we're about to discover a potential V inside an integral and it wouldn't do to confuse it with a volume element written as dV!

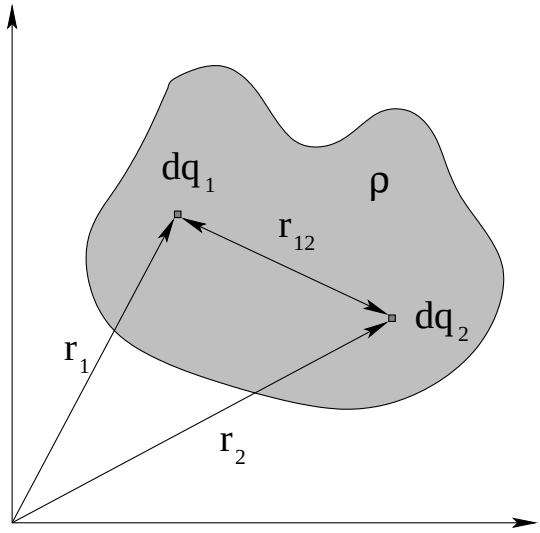


Figure 3.11: Two differentially small chunks of a charge distribution $\rho(\vec{r})$, separated in space by the distance r_{12}.

Now we can easily write the (differential) potential energy of *just these two chunks!*

$$dU = \frac{k_e dq_1 dq_2}{r_{12}} = k_e \frac{\rho(\vec{r}_1) d^3 r_1 \rho(\vec{r}_2) d^3 r_2}{|\vec{r}_1 - \vec{r}_2|} \tag{3.68}$$

We have to integrate both sides of this over the entire distribution, but we have a problem! If we fix, say, $\vec{r}_2$ and compute the potential energy of just dq_2 at this point in the field of all of the *other* charge in the distribution, we still have to sum over all of the chunks dq_2, using a second integral. But if we integrate over *both* $\vec{r}_1$ *and* $\vec{r}_2$ to get every chunk in the field of every other chunk, we'll count each pair twice! This means that for this to work, we'll have to divide by one half:

$$U_{\rm tot} = \int dU = \frac{1}{2} \int_{\mathcal{V}_1} \int_{\mathcal{V}_2} \frac{k_e dq_1 dq_2}{r_{12}} = \frac{1}{2} \int_{\mathcal{V}_1} \rho(\vec{r}_1) d^3 r_1 \left\{ \int_{\mathcal{V}_2} \frac{k_e \rho(\vec{r}_2) d^3 r_2}{|\vec{r}_1 - \vec{r}_2|} \right\} \tag{3.69}$$

where both integrals are over the entire support volume $\mathcal{V}$ of the charge distribution and are labelled with the coordinate one is integrating over to keep all of this straight.

We can make this just a bit simpler if we identify the second integral in the big $\{\}$ above:

$$V(\vec{r}_1) = \left\{ \int_{\mathcal{V}_2} \frac{k_e \rho(\vec{r}_2) d^3 r_2}{|\vec{r}_1 - \vec{r}_2|} \right\} \tag{3.70}$$

so:

$$U_{\rm tot} == \frac{1}{2} \int_{\mathcal{V}_1} V(\vec{r}_1) \rho(\vec{r}_1) d^3 r_1 \tag{3.71}$$

In words, one way of evaluating the electrostatic potential energy of a charge distribution is to find the potential of that distribution ***at an arbitrary point inside*** by integrating dV over the distribution, find the potential energy of a tiny chunk of that distribution at that point, sum the chunks (with the integral) over the entire distribution, and then (recognizing that this double counted the sum) ***divide by two***. This clearly corresponds to one of the two forms for the potential energy as a sum over discrete charges. But what of the other form?

The second form, as we saw in the example above, is the "build a distribution" approach. We start with no charge, then bring a tiny chunk dq to (say) the origin for free – this "for free" corresponds again to not counting self-energy, something we'll talk a bit about below once we have solved a key example. Then we bring in a second chunk, counting the increase in potential energy as the work we do brinnging that charge in. We bring in the third, fourth, etc, only instead of doing this with discrete charges, we do this with differential chunks in a suitable coordinate system!

Formulating this algebraically in a way that doesn't depend on a specific coordinate frame is a bit tricky. That's because we have to break the two integrals involved into ***two disjoint pieces*** – an "interior" volume that represents the amount of charge brought in *so far* and an "exterior" part that represents the integral of all of the charge *outside* of that volume. By partitioning the integral in this way, we avoid double counting the same way we avoid it with the double sum where one index (say j) runs from 1 to N and the other index (i) runs from 1 to $j-1$, with no self-energy, so $i<j$ for all terms in the sum. With hopefully obvious notation, then:

$$U_{\text{tot}} = \int_{\mathcal{V}_{1,\text{exterior to 2}}} \rho(\vec{\boldsymbol{r}}_1) d^3 r_1 \int_{\mathcal{V}_{2,\text{interior to 1}}} \frac{k_e \rho(\vec{\boldsymbol{r}}_2) d^3 r_2}{|\vec{\boldsymbol{r}}_1 - \vec{\boldsymbol{r}}_2|} \tag{3.72}$$

Both integrals are over the entire volume $\mathcal{V}$ of the support of ρ, note well! They are simply arranged so that as each value of (say) $\vec{\boldsymbol{r}}_2$ is included in the evaluation of $V(\vec{\boldsymbol{r}}_1)$, that point is subsequently *excluded* from the integral over $\vec{\boldsymbol{r}}_2$.

If you are reading this and trying to learn *both* what this means so you understand it *and* how to have a faint hope of putting it into practice solving problems, you are very likely feeling pretty insecure right about now. Hopefully you do get the *idea* – we do both integrals (somehow) in such a way that we avoid double counting, great, fine, super, but *how the heck do we do that?*

The easiest way to get the rest of the way there is (as usual) to work a good example!

Example 3.5.2: Potential Energy of a Uniform Ball of Charge

Let's find the potential energy of a *uniform ball of charge* with total charge Q and radius R. We will use the method that implements the disjoint integral above and (more importantly) *interpret* this as summing the work required to assemble the ball of charge from infinity – the "build a ball" method, if you like.

Figure 3.12 shows just such a ball at an intermediate stage of being built, in *spherical coordinates* centered on the ball. At the instant portrayed, a uniform charge density ρ has filled the ball of eventual radius R (dashed sphere) out to the radius r (light grey sphere). We have grabbed a handful of charge $dQ = \rho d\mathcal{V}$ at infinity and pushed it, working *against the field of the ball* in to form a differentially thin layer with volume $d\mathcal{V}$ on the surface of this ball.

Now, the field of the ball of charge (so far) is, from GLE, using a (red) gaussian surface S of radius r' outside of the ball:

$$\oint_S \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA = E_r 4\pi r'^2 = \frac{1}{\epsilon_0} \int_{\mathcal{V}/S} \rho d\mathcal{V} = \frac{4\pi\rho r^3}{3\epsilon_0} \tag{3.73}$$

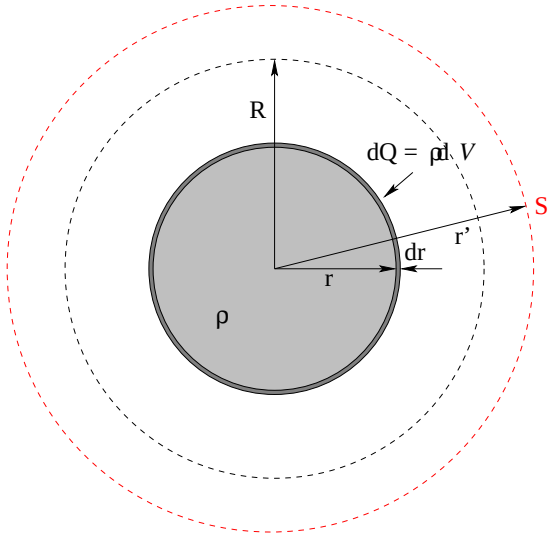


Figure 3.12: A ball of charge in the process of being built. So far it has accumulated out to a radius r, and the figure illustrates a (dark shaded) thin layer of charge being *added* to the ball.

(**Note Well** the difference between r and r'!) where:

$$\rho = \frac{3Q}{4\pi R^3} \tag{3.74}$$

so:

$$E_r(r') = \frac{k_e Q \frac{r^3}{R^3}}{r'^2} = \frac{k_e Q_{\text{ball so far}}}{r'^2} \tag{3.75}$$

expressed in terms of the givens Q and R. In this:

$$Q_{\text{ball so far}}(r) = \frac{r^3}{R^3} Q$$

is the *fraction* of the total charge inside r. At this point, if you have *mastered* the homework and examples done so far I could probably have just written the last two equations down and started here and you would have followed the argument, but it never hurts to practice using GLE as we go.

Next, we find the work *we* do pushing the charge dQ in from ∞ *against* this field to the radius r. We push in the negative r' direction, and $F_r(\vec{r}') = dQ E_r(r')$, so:

$$dW = -dQ \int_\infty^r \frac{k_e Q \frac{r^3}{R^3}}{r'^2} dr' = dQ \frac{k_e Q \frac{r^3}{R^3}}{r} = \frac{k_e Q r^2}{R^3} dQ \tag{3.76}$$

As we have argued before, the potential energy we add to the system equals the work *we* do adding this charge. We recognize the fraction multiplying the dQ at the end as just $V(Q_{\text{ball so far}}, r)$. We also use our usual litany to write:

$$dQ = \rho\, d\mathcal{V} = \frac{3Q}{4\pi R^3} \times 4\pi r^2 dr = \frac{3Q r^2 dr}{R^3} \tag{3.77}$$

$$dU = dW = V(Q_{\text{ball so far}}, r)\, dQ = \frac{k_e Q r^2}{R^3} dQ = \frac{3 k_e Q^2 r^4 dr}{R^6} \tag{3.78}$$

We have now reduced the *two* 3D integrals in the formal algebraic expression to *one one dimensional integral* that (I sincerely hope) *makes sense!* If you like, the increase in potential energy brought about by increasing the thickness of the charged ball of radius r *so far* by a differential amount dr is:

$$dU = V\,dQ \tag{3.79}$$

where V is the potential at the surface of the ball *so far* and dQ is the charge added *at that radius*. V of the ball at its surface is the interior integral at the *specific* radius r, and we're about to do the exterior integral to add up the work required to build the ball out of many infinitesimal layers, like an onion!

Now it is easy!

$$\boxed{U_{\rm tot} = \int dU = \frac{3k_eQ^2}{R^6}\int_0^R r^4 dr = \frac{3}{5}\frac{k_eQ^2}{R}} \tag{3.80}$$

Example 3.5.3: Potential Energy of a Uniform Ball of Charge – Second Method

Let's do this example a second time, but this time we'll use the potential inside the ball of the *whole* ball times dQ and divide the result by 2. We'll start with the potential inside a ball of charge, computed as an example above in equation 3.54:

$$V(r) = 2\pi k_e\rho R^2 - k_e\left(\frac{2\pi\rho}{3}\right)r^2 = \frac{3k_eQ}{2R} - \frac{k_eQr^2}{2R^3} \tag{3.81}$$

Now we have to multiply this by dQ (in spherical polar coordinates as usual) and divide by 2 to form dU:

$$dU = \frac{1}{2}V(r)\,dQ = \frac{1}{2}\left(\frac{3k_eQ}{2R} - \frac{k_eQr^2}{2R^3}\right)\times\frac{3Q}{R^3}r^2dr \tag{3.82}$$

where I've simplifed $\rho\,d\mathcal{V}$. We have two terms, then, to integrate from 0 to R to cover the ball. First:

$$U_1 = \frac{9k_eQ^2}{4R^4}\int_0^R r^2dr = \frac{3k_eQ^2}{4R}$$

and second:

$$U_2 = -\frac{3k_eQ^2}{4R^6}\int_0^R r^4dr = \frac{3k_eQ^2}{20R}$$

so that:

$$\boxed{U_{\rm tot} = U_1 + U_2 = \frac{3k_eQ^2}{R}\left(\frac{1}{4} - \frac{1}{20}\right) = \frac{3k_eQ^2}{5R}} \tag{3.83}$$

as we got before.

Note that this method is arguably *slightly more difficult* than adding up the work required to build the ball, once you factor in the work involved in both finding $V(r)$ for $r < R$ and doing the resulting integrals. For spheres, at least, it is easy enough to jump straight to $V(r) = k_eQ_{\rm ball\ so\ far}(r)/r$, evaluate the charge in the ball, and just *write down* the integral to be done in one step, because GLE provides rules for the field and potential outside of a spherical ball.

Example 3.5.4: Potential Energy of a Spherical Shell of Charge

Before we go on to a final topic of interest regarding the potential energy of a spherical distribution of charge with compact support, let's find the potential energy of a simple spherical shell of charge Q at radius R. Now we *know* that $E_r = k_e Q/r^2$ for $r > R$ and:

$$V(R) = \frac{k_e Q}{R} \tag{3.84}$$

This lets us find U using the rule with the factor of $\frac{1}{2}$ in it in *one step*:

$$\boxed{U_{\text{tot}} = \frac{1}{2}V(R)Q = \frac{k_e Q^2}{2R}} \tag{3.85}$$

There are several other ways of getting this result, including a "build a ball" solution that actually uses the *first* method, summing up the work needed to add small increments of charge dq to a ball that already has the charge q:

$$dW = dU = V(q)dq = \frac{k_e q}{R}dq \tag{3.86}$$

so:

$$\boxed{U_{\text{tot}} = \int dU = \int_0^Q \frac{k_e q}{R}dq = \frac{k_e Q^2}{2R}} \tag{3.87}$$

Note that in this latter case we actually end up integrating over $q\,dq$, not r, but we end up getting the same result. It turns out that this approach is slightly easier to understand when we think about charging up a conductor, which is the primary topic of the next chapter.

3.5.3: Self-energy of a 'Point Charge'

Up to now, we have ignored the so-called *self-energy* of a point charge – why, when we sum over the potential energy of all pairs of charges do we not include a term that describes a charge interacting with *itself*, the *diagonal* terms omitted in the discrete sum rule? Also, why does it turn out to be "OK" (or at least, give us reasonable answers) when we do *not* omit this double counting when we do a double *integral* over a smooth charge distribution, presumably counting dq interacting with itself at zero-ish distance?

The answer is both complicated and extremely interesting. To understand it, we have to build a *model* for a charged particle. Or rather, we just finished building something that would work very well *as* a model for a charged particle – a charged ball with charge Q and radius R. We see that when we do so, we end up with a potential "self-energy" of the ball equal to something like:

$$U_Q = \frac{3k_e Q^2}{5R} \text{ or } \frac{k_e Q^2}{2R} \text{ or } \sim \mathcal{O}(1) \times \frac{k_e Q^2}{R}$$

The first result was for a uniform ball of charage. The second was for a spherical shell with the same charge and radius. Presumably other *similar* ways of distributing charge Q inside a ball of radius R will differ in the dimensionless numerical factor, but will likely *scale* like a constant of order unity times $k_e Q^2/R$!

Now we can determine what happens when we imagine compressing a given finite charge Q into smaller and smaller balls! As $R \to 0$, we see that $U(R \to 0) \to \infty$! If we forget the factor of $3/5$, or $1/2$ (which depends on the *details* of the charge distribution) and focus on the rest, we can compute a couple of extremely interesting quantities that give us insight into nuclear physics, certain properties of electrons, and field theories involving ***pointlike charged particles*** in general!

Consider a model for a *proton* – where we know $Q = +e$ and $R \approx 10^{-15}$ meters (one fermi) – as a ball of charge. If one computes $k_e e^2/R$ in eV, one gets +1.44 MeV (try it!) This is the *order of magnitude* of the energy bound up in the electrostatic field of the charge of a proton. This energy is *repulsive* – the nucleus wants to *blow apart* and release the electrostatic potential energy in the form of e.g. kinetic energy of its constituents, which we believe to be three *quarks* if you look back at the first chapter!

Why don't the quarks fly apart? They have to be bound together by *even stronger forces* than this remarkably strong electrostatic repulsion! The so-called "strong nuclear forces" that glue all of this charge together (with *gluons*, yet) must be much stronger than electrostatic forces to make the *total* energy negative or a proton would not be a stable bound state, and they are. Electronic energy levels in atoms are scale eV, nuclear energy levels are scale MeV (and higher) which explains why stars burn slowly and release far, far more energy than can be explained by "atomic" electronic bonding (conventional burning). Nuclear fusion releases on the order of *ten million times* as much energy per fusion event than does e.g. burning one carbon atom into carbon dioxide.

To consider the electron, we require a "true fact" (that is, fortunately, fairly common knowledge): Mass and energy are interchangeable, and the "rest mass" of an object corresponds to a "rest energy" of mc^2 where $c = 3 \times 10^8$ meters/second is the speed of light. Now we suppose that an electron's rest mass is all due to its electrostatic energy of confinement, the energy tied up in the charge e confined to *some* radius, and we seek that radius, which we will call "the classical radius of the electron"[63] . This is the same computation as above, only backwards – we know the energy already, we know k_e and the charge $-e$, we solve for r_e. If you do this, using $U \sim mc^2 = 0.5$ MeV for an electron, one gets 2.8×10^{-15} meters. Note well that this is somewhat *larger* than the size of a proton (as the electron has less energy). The classical radius of the electron turns out to be an important quantity in determining the properties of electromagnetic radiation from point charges.

However, this leaves us with a serious problem. We can imagine a proton made up of quarks bound with a force so strong that it overcomes the electrostatic force trying to blow the quarks apart, and can even find some evidence that such a force exists and that the proton is a composite particle. However, when we examine the electron in very high energy collisions, we find no evidence of "structure" that might be expected to exist if it were a composite particle. Furthermore, in all cases the electron behaves (within the bounds of *quantum* mechanics) as if it were a *true point-like particle with zero*[64] *radius!*

[63]Wikipedia: http://www.wikipedia.org/wiki/Classical Electron Radius.

[64]This is true only in a qualified sense. It turns out that the vacuum itself is *polarizable* in the strong fields that exist very, very close to the location of the electron in quantum field theories. Sufficiently strong fields split the vacuum into electron-positron pairs that *screen* the electrostatic field at very short ranges, soften the divergence in self-energy, and have observable side effects so we really believe that this happens. However, *all* the elementary charged partcles *have* to be pointlike, because if they were not we would *need* an additional interaction to hold

3.6: Conductors in Electrostatic Equilibrium

Last week we learned together, Gauss's Law and the notion of equilibrium combine to give us important information about *conductors* – material with an "inexhaustible" supply of charged particles such as electrons that are free to move within the conductor and behave like an "electrical fluid". In particular, we determined that $\vec{\boldsymbol{E}} = 0$ inside a conductor in electrostatic equilibrium and that $\vec{\boldsymbol{E}}_{||} = 0$ at the surface, so that any electrical field immediately outside its surface must be perpendicular to the surface.

This suffices to show that conductors are *equipotential* – the potential difference between any two points in the conductor or on its surface is:

$$\Delta V = -\int_{\vec{\boldsymbol{x}}_0}^{\vec{\boldsymbol{x}}_1} \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{x}} = 0 \tag{3.88}$$

Note that this doesn't mean that the potential of the conductor is *zero*, only that it is a *constant*. That is consistent:

$$\vec{\boldsymbol{E}} = -\vec{\boldsymbol{\nabla}} V_0 = 0 \tag{3.89}$$

when V_0 is any constant.

This also permits us to make an important observation. For any arrangement of (say two) isolated conductors with sufficient symmetry that we can put an arbitrary charge on either of them and not have their interaction break the symmetry of the charge's redistribution, we can compute the *potential difference* between the conducting pair as a function of the charge difference between them. This potential difference will turn out to be proportional to the charge transferred and will only otherwise depend on the *geometry* of their arrangment. In the next chapter this will be the basis of the notion of *capacitance*.

3.6.1: Charge Sharing

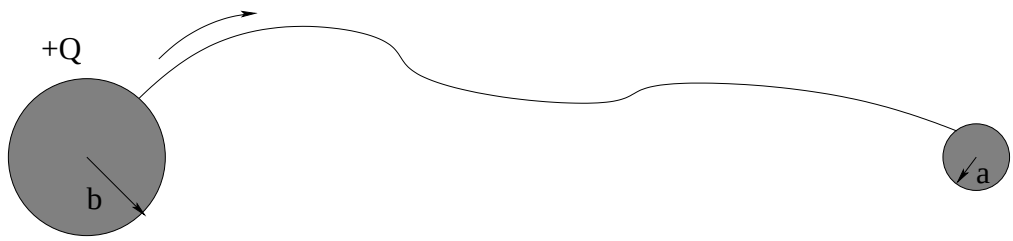


Figure 3.13: Charge sharing between two distant conductors connected by a wire. They become equipotential, with charge transferred (shared) between them to make it so.

Here is an important example of equipotentiality. Suppose one has two conducting spheres, one with radius a and one with radius b such that $a \ll b$ (as seen in figure **??** above. Let us further suppose that the spheres are very distant from one another so that the field of one is

them together against the electrostatic repulsion, as the *non*-elementary proton is in fact held together! Instead, we need a consistent quantum field theory of the elementary particles, which is just great in theory but extremely elusive in practice and, fortunately, well beyond the bounds of this course.

very weak in the vicinity of the other (so that very little charge redistribution occurs if one or the other is charged up). We begin by imagining that we have put a charge Q on sphere b.

In that case it is easy to see or show that:

$$V_b = -\int_{\infty}^{b} E_r dr = \frac{kQ}{b} \tag{3.90}$$

everwhere inside sphere b while

$$V_a = 0 \tag{3.91}$$

on the other sphere. There is clearly a potential difference between the two spheres. Now imagine that we connect the two with a thin conducting wire. They form a single conductor and therefore quickly *equalize* their potentials as charge flows from b to a.

Charge is conserved. They will reach equilibrium when:

$$\frac{k(Q-q)}{b} = \frac{kq'}{b} = \frac{kq}{a} \tag{3.92}$$

where q is the net charge transferred from b to a and q' is the remaining charge on b. This can be rewritten as:

$$\frac{q}{q'} = \frac{a}{b} \tag{3.93}$$

The smaller the sphere the smaller the fraction of charge on it, which makes sense since the *ratio* of charge to radius must be the same.

Now, however, we compute the *radial field at the surface* of the two conductors. It is:

$$E_a = \frac{kq}{a^2} \tag{3.94}$$

$$E_b = \frac{kq'}{b^2} \tag{3.95}$$

If we take the ratio of the *field strengths* we get:

$$\frac{E_a}{E_b} = \frac{q}{q'}\frac{b^2}{a^2} = \frac{b}{a} \tag{3.96}$$

and conclude that the *field is much stronger on the surface of the smaller conductor*. In fact, it becomes *infinite* in the limit that $a \to 0$ relative to a finite b.

What this tells us is that the field in the vicinity of a conductor in electrostatic equilibrium at some non-zero potential is *much stronger at sharp points* than it is on smooth surfaces with a large radius of curvature. This has important consequences, as we shall see!

3.7: Dielectric Breakdown

Insulators are not ever perfect, because electrons as charge carriers are not bound to the conducting substrate by an infinite potential energy barrier. In a sufficiently large field electrons are torn from their parent atoms and insulators “suddenly” become conductors, a process called *dielectric breakdown*. Lightning is a spectacular example of dielectric breakdown in nature.

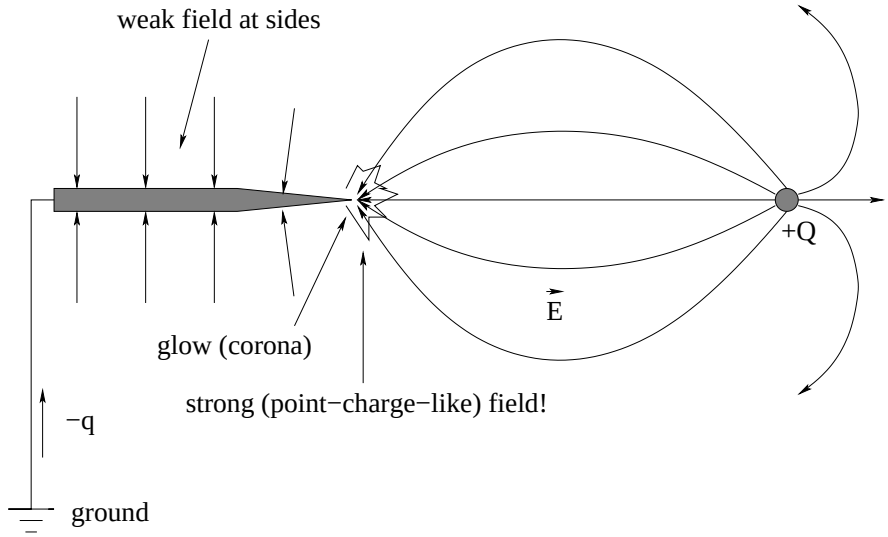


Figure 3.14: External charge $+Q$ induces a charge -q on the sharp tip of a nearby conductor. Electric fields lines leave the tip at right angles, producing a field that looks like that of a very large *point charge* which is extremely strong very close to the tip. This in turn ionizes nearby air molecules, creating the *corona* (and spraying/repelling negatively charged ions out into the air where they are attracted to $+Q$ and eventually neutralize it).

The way lightning (or any sort of arc discharge) works is that charge builds up on clouds and/or the ground to create a large potential difference. At some point the field strength associated with this potential difference becomes great enough that the force it exerts on electrons exceeds the force binding the electrons to their parent atoms in the insulator (or alternatively, they get enough potential energy to overcome the potential energy barrier that confines them). At first only a few electrons get away, and are quickly accelerated by the field as they get over the confining potential barrier.

These electrons in turn collide with other nearby atoms, tranferring momentum to them and knocking still more electrons loose. A cascading chain reaction occurs that heats the atoms in the path of the ever increasing flow of charge and knocks still more charge loose to join that flow. In a fraction of a second, the superheated air becomes a white-hot *plasma* that conducts electricity quite well and the enormous charge difference between ground and cloud or cloud and cloud neutralizes in a burst of millions of ampere's of current. Bang! Zap! Ouch!

It is important to remember whenever working with high voltages that *few materials* are terribly good insulators against the strong fields associated with large potential differences over a short distance. That is, if you get close enough to a high voltage line it will simply arc over and electrocute you. It may well arc through a piece of glass or plastic and kill you. Wood is an insulator for ordinary voltages but conducts more than enough to kill you if you try to touch a high voltage power line with a stick.

Note also that if one approaches a conductor with a charge, one *induces* a charge on the part of the conductor nearest the charge. If that part happens to be a sharp point, the properties of charge sharing on an equipotential conductor create an *extremely strong field* in the immediate vicinity of the point as illustrated in figure 3.14. This is the basis of Prokop Diviš' (and, independently a short time later, of Benjamin Franklin's) most important inventions, the

lightning rod[65]

The field generated by the induced charge at the tip of a sharp, conducting, grounded point in the field of a highly charged cloud overhead can easily be strong enough to ionize air molecules in the immediate vicinity of the tip and make them conduct! The ionized air molecules transfer electrons from molecule to molecule while still in the vicinity of the tip, and these electrons emit light as they rebind to molecules. This "crowning" glow at the tip is called the *corona*; a related phenomenon observed sometimes on ships at sea in storms (where entire ship masts can flicker and glow as stormclouds approach) is called Saint Elmo's Fire[66]

Dielectric breakdown also occurs when a sharp point is deliberately charged to a high voltage instead of being connected to ground *near* a source of strong field. One can easily observe this light in the classroom or lab, visible in the dark as a faint blue-violet glow emitted by the air around the point of a thumbtack attached to an electrostatic generator (works best on dry days with the lights off).

The molecules that are charged by colliding with the tip are then *strongly repelled* by the field of the tip itself. They literally spray away from it, carrying charge and momentum and flowing towards the inducing charge on e.g. the overhead cloud. This is a process called *corona discharge*, and is important in the design/function of most modern copiers or printers. The virtue of Franklin's discovery is that because clouds usually approach buildings relatively slowly, this spray of charge has just the right sign to be attracted to the cloud that is causing it instead of gradually returning to ground! As it arrives, it can *gradually* neutralize the cloud's charge (of either sign) before the field builds up to where the catastrophic dielectric breakdown known as lightning does neutralizes the cloud to ground *all at once* – through your barn or building – starting fires and killing humans and horses in the process.

In the event that this slow neutralization fails, the grounding conductor attached to the tip attracts the resulting lightning stroke and provides it with a path to ground that is not through the building structure itself. As the linked Wikipedia article indicates, this protection is not absolute as may be enough current in a direct strike to turn the conductor itself into a plasma and start fires, or the bolt may fork and take several paths to ground including ones through the structure. It is also worth noting that there is (still) considerable debate on the optimal curvature or "pointiness" of the tip of a lightning rod, with the modern consensus being a somewhat rounded tip with a larger radius of curvature works better than a sharp point with a very small one.

A second popular classroom demonstration of the spray of charged particles from a high voltage point in air involves attaching a freely pivoted conducting "whirlygig" with two sharp points connected to a high voltage generator, bent so that the charged air repelled from their tips exerts a torque (via Newton's Third Law, of course) that makes the arms spin. This is a kind of "ionic jet" and makes it clear that there is a *substantial* flow of charged matter away from the sharp points as long as they are charged up to high voltage!

[65]Wikipedia: http://www.wikipedia.org/wiki/Lightning Rod. .
[66]Wikipedia: http://www.wikipedia.org/wiki/St Elmos Fire. .

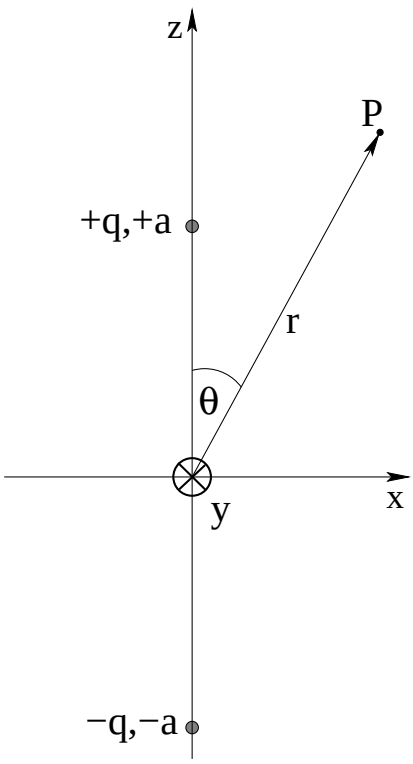
Homework for Week 3

Problem 1.

Physics Concepts

Make this week's physics concepts summary as you work all of the problems in this week's assignment. Be sure to cross-reference each concept in the summary to the problem(s) they were key to. Do the work carefully enough that you can (after it has been handed in and graded) punch it and add it to a three ring binder for review and study come finals!

Problem 2.



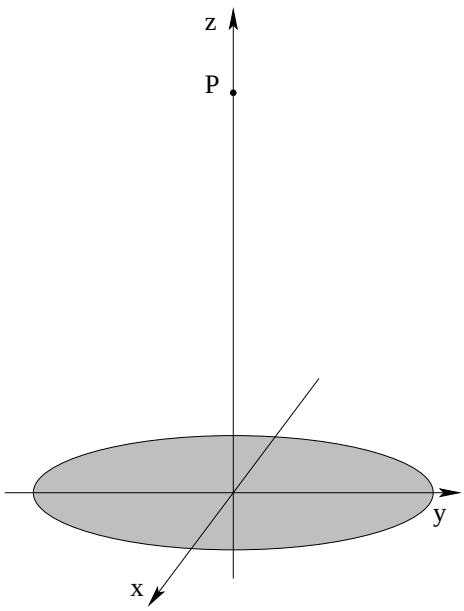
Suppose you have an electric dipole: charge q at position $z = a$ on the z-axis and charge $-q$ at $z = -a$.

a) Write an exact expression for the electrostatic potential of the dipole at point P located at $\vec{r} = (r, \theta)$. Note that the potential must be ϕ-independent because of azimuthal symmetry.

b) Expand your answer to a) for $r \gg a$ to leading surviving order and express the answer in terms of the magnitude of the (z-directed) dipole moment, $p_z = 2qa$.

c) Note that the potential is identically zero on the xy-plane. Suppose one slides an (infinite) thin *grounded conducting sheet* in between the two charges. This requires no work and does not alter the fields or potentials in either half-space above or below it. Now imagine removing the charge below this plane. Does doing so change the fields or potentials in the upper half space (recall that the conductor *screens* the two spaces)? Using the insight gained from thinking about this, compute the force of attraction between an isolated charge q a distance a away from a grounded conducting sheet.

Problem 3.



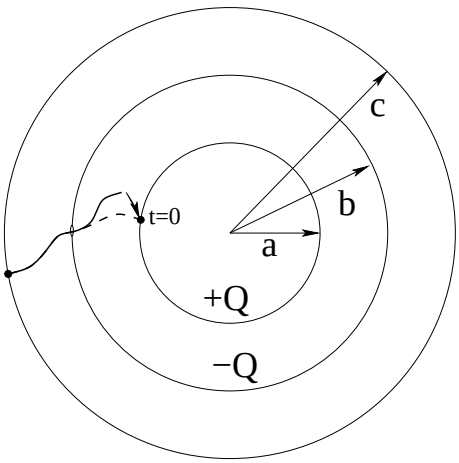
a) Find the potential of a thin disk of radius R covered with a surface charge density σ at a point $P = (0, 0, z)$ on its axis of symmetry as shown.

b) Then expand the result to leading order in the two limits $R \gg z$ and $z \gg R$ and interpret/identify the potentials in both of these cases.

c) Use $E_z = -\dfrac{dV}{dz}$ to show that your answers to part b) lead to reasonable/expected expressions for the electric field in these two limits.

Problem 4.

Three thin conducting spherical shells have radii $a < b < c$ respectively. Initially the shell with radius a has a charge $+Q$ and the shell with radius b has a charge $-Q$. At $t = 0$, you connect the shells with radii a and c using a thin wire that passes through a tiny (insulated!) hole through the middle shell and wait for the charges on all three to reach a new equilibrium where the charges on each shell are Q_a, Q_b and Q_c respectively. Find:



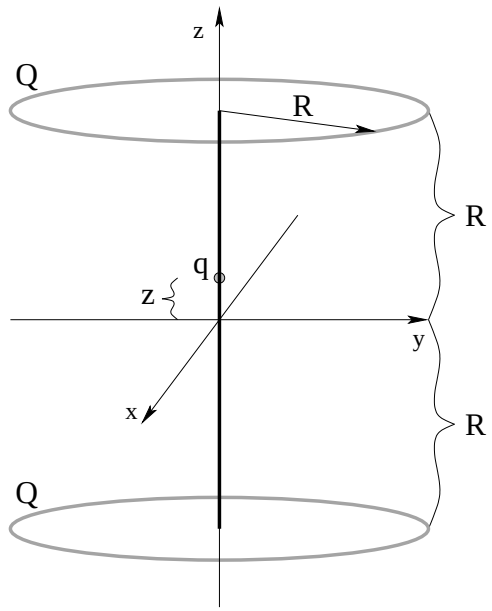
a) The potential at all points in space in terms of Q_a, Q_b and Q_c.

b) The charges Q_a, Q_b, Q_c on all three shells in terms of the given Q.

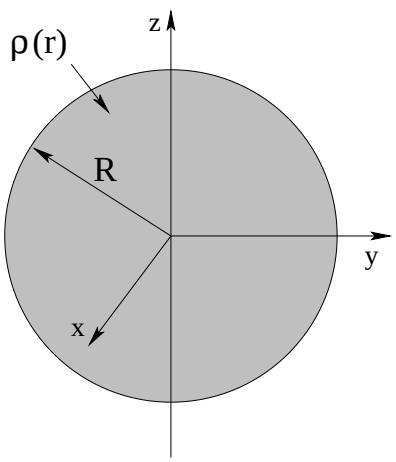
Problem 5.

Two rings of charge Q and radius R (uniformly distributed) are located at $z = \pm R$ and have the same (z) axis. If a small bead of mass m with charge q is threaded on a frictionless string along the z axis and displaced a small distance $+z_0 \ll R$ from the origin, it will oscillate *approximately* harmonically.

Expand the potential energy of the bead interacting with the rings and use the result to find ω, the angular frequency of oscillation. Write down a general expression for $z(t)$.



Problem 6.



Consider the two possible charge density distributions inside a sphere of radius R of charge:

a) A *uniform* charge density: $\rho(r) = \rho_0$

b) A *non-uniform* charge density: $\rho(r) = \rho_1 \dfrac{r}{R}$

(both zero for $r > R$). Find $E_r(r)$ (from Gauss's Law) for the electrostatic field(s) at all points in space (both $r \leq R$ and $r > R$) for the two charge density distributions given below. Then find the electrostatic potential(s) $V_a(r)$ and $V_b(r)$ at all points in space for the two cases by directly integrating $E_r(r)$ in from $r = \infty$.

Note well that you found the E-fields *already* in a homework problem last week – review/redo your solutions as part of *this* solution to help solidify your understanding!

Problem 7.

Use Gauss's Law and direct integration of the field to ompute the potential difference(s) ΔV between:

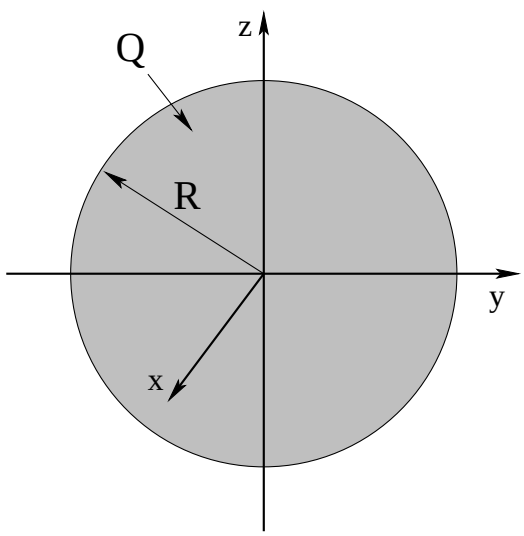
a) Two conducting spherical shells of radius a and b with a charge $+Q$ on the inner one and charge $-Q$ on the outer one.

b) Two "infinitely long" conducting cylindrical shells of radius a and b with a charge per unit length $+\lambda$ on the inner one and charge per unit length $-\lambda$ on the outer one.

c) Two "infinite" conducting sheets of charge, one with charge $+\sigma$ on the xy plane and with with charge $-\sigma$ parallel to the first one but at $z = d$.

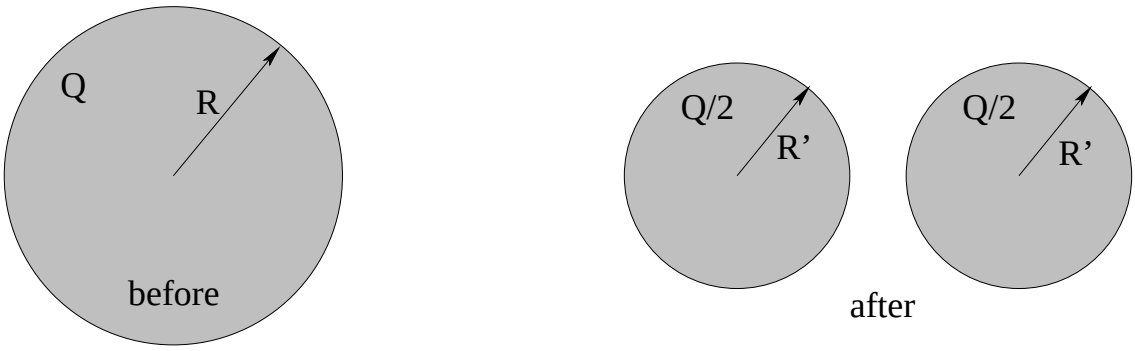
Great! Now you've done *almost all the work* required to understand *capacitance* in our next chapter!

Problem 8.



Find the potential energy of a uniform ball of charge of total charge Q and radius R. You have two ways you might try this so far: integrating the work required to *assemble* the charge against the field of the charge already present, or (equivalently) integrating $dU = V\,dq$ in layers. In the next chapter you'll learn a third way that would work: integrating the electric field energy density over all space.

Advanced Problem 9.



The total electrostatic potential energy of a uniform ball of charge with charge Q and radius R is:

$$U_Q = \frac{3}{5}\frac{k_e Q^2}{R}$$

Let's try to use this to understand a little bit about ***nuclear fission*** by pretending that this is a model for (say) a Uranium 235 nucleus with 92 protons and total charge:

$$Q = 92e = 92 \times 1.6 \times 10^{-19} \approx 1.5 \times 10^{-17} \text{ coulombs}$$

bound together in a nucleus with 143 neutrons inside a radius of approximately $R = 7.4 \times 10^{-15}$ meters.

Imagine that this charge Q is distributed uniformly in an *incompressible fluid* inside the spherical nucleus. Now imagine that sphere splits into two identical, smaller spheres, each with half of the charge and the *same density*. Then:

a) Find the radius R' of these two spheres when each sphere has half of the total charge Q.

b) Find the total electrostatic energy of these two spheres once they have stabilized and are separated by a large distance.

c) Compare the answer to the answer from the previous problem. Was energy released? What form would you expect this energy to take?

Week 4: Capacitance

- Conductors *store charge* and as they do so, their *potential* (difference) *increases* relative to ground.

- If we arrange two conductors in a symmetric way and do *work* to transfer charge from one to the other (leaving behind an equal charge of the opposite sign) we call the arrangement a *capacitor* – a device for storing energy in the electrostatic field.

- The capacitance of the arrangement is defined to be:

$$C = \frac{|\Delta Q|}{|\Delta V|} \tag{4.1}$$

  or, the capacitance *is* the amount of charge we can store that creates a potential difference of one volt between the conductors. ***Note the absolute value bars*** – capacitance is given as a ***positive quantity***.

- The SI units of capacitance are called ***farads*** where:

$$1\text{F} = \frac{1 \text{ Coulomb}}{1 \text{ Volt}} \tag{4.2}$$

  A farad is an *enormous* capacitance. Typical values for capacitors in devices range from picofarads to microfarads, although one can actually buy one farad capacitors for special projects these days. ***Large capacitors are dangerous!*** Especially when strung together to make a large capacitor at high voltage! Anything over a few hundred microfarads at a potential of 100+ volts or so can be lethal!

  You should be able to *derive* the following quantities (from Gauss's Law, integration of potential difference, dividing into the presumed total charge):

- Parallel plate capacitor:

$$C = \frac{\epsilon_0 A}{d} \tag{4.3}$$

  where A is its cross sectional area and d is the separation of the plates.

- Cylindrical capacitor:

$$C = \frac{2\pi L \epsilon_0}{\ln(b/a)} \tag{4.4}$$

  where a is the outer radius of the inner conductor, b the inner radius of the outer conductor, and L is its length (where we assume $L \gg (b-a)$).

- Spherical capacitor:

$$C = 4\pi\epsilon_0 \frac{ab}{(b-a)} \tag{4.5}$$

where a is the outer radius of the inner conductor and b the inner radius of the outer conductor.

- Energy stored in a capacitor:

$$U = \frac{1}{2}QV = \frac{1}{2}CV^2 = \frac{1}{2}\frac{Q^2}{C} \tag{4.6}$$

where the first form is the simplest to understand.

One question that is very important is *where* is all this energy stored in the capacitor? The "best" answer will be: in the electric field! If we write the energy in terms of the electric field, we find that the *energy density of the electric field* is given by:

$$\eta_e = \frac{1}{2}\epsilon_0 E^2 \tag{4.7}$$

- Adding capacitors in parallel:

$$C_{\text{tot}} = C_1 + C_2 + ... \tag{4.8}$$

- Adding capacitors in series:

$$\frac{1}{C_{\text{tot}}} = \frac{1}{C_1} + \frac{1}{C_2} + ... \tag{4.9}$$

- Dielectrics are *insulators* that *polarize* when placed in an electric field. This builds up a surface charge that *reduces* the electric field inside the material – it *displaces* it from its usual value. For "weak fields" this reduced field is:

$$\vec{\boldsymbol{E}} = \frac{\vec{\boldsymbol{E}}_0}{\epsilon_r} \tag{4.10}$$

where $\vec{\boldsymbol{E}}_0$ is the external field, $\vec{\boldsymbol{E}}$ is the field inside the dielectric, and $\epsilon_r \geq 1$ is the *relative permittivity* (also called the *dielectric constant* κ in many "standard" physics textbooks, although this usage has been deprecated as being too ambiguous) and is characteristic of the material.

One can consistently describe both conductors and insulators in terms of their dielectric properties by evaluating their *permittivity* (relative to the vacuum permittivity ϵ_0 we've used so far) and using it to compute the electric field inside the material:

$$\epsilon = \epsilon_r \epsilon_0 \tag{4.11}$$

This is the *actual* permittivity of the material, and in the general case of a time dependent applied electric field is a complex-valued function of frequency, leading (eventually) to a consistent description of *resistance* and Ohm's Law, and to *dispersion* and the rainbow!

- Dielectrics perform three important functions in the engineering of capacitors:

  a) They physically separate the plates (which, recall, experience a possibly strong force of attraction).

b) They reduce the field in between the plates, which reduces the potential difference, which increases the amount of charge one can store per volt – the capacitance. If the material *fills* the space between the plates you should be able to (easily) show that:

$$C = \epsilon_r C_0 \tag{4.12}$$

where C_0 is the capacitance without the dielectric.

c) They prevent *dielectric breakdown*, so the physical separation of the plates d can be much smaller (and the capacitance much larger) at some design voltage.

4.1: Capacitance

In the previous chapter we noted that *conductors in electrostatic equilibrium are equipotential*. If you imagine charging up any given conductor, every new bit of charge we add to it spreads itself out the same way. One expects the field produced at its surface to scale up or down proportional to the amount of charge on the conductor but not change its basic shape. As a consequence, one expects the *potential* produced by the conductor to be proportional to its total charge at all points in space, in particular inside the equipotential conductor itself.

This has been apparent in all of our Gauss's Law examples up to now. For example, a conducting sphere of radius R, charged with a total charge Q, has a field:

$$E_r = \frac{k_e Q}{r^2} \qquad (r > R) \tag{4.13}$$

$$= 0 \qquad (r < R \text{ inside the conductor}) \tag{4.14}$$

If we integrate this to find the potential everywhere in space we get:

$$V = -\int_{\infty}^{r} \frac{kQ}{r^2} dr$$

$$= 0\frac{k_e Q}{r} \qquad (r \geq R) \tag{4.15}$$

The conductor is *equipotential*, so the potential inside is the same as at its surface:

$$V = \frac{k_e Q}{R} \qquad (r < R) \tag{4.16}$$

We have seen how just *knowing* this solution for spherical shells, or the equivalent solution for cylindrical shells, can greatly improve our ability to solve problems quickly and easily by using superposition of these once-and-for-all solutions instead of trying to explicitly integrate the fields across all the different forms it might take in a problem with several conducting shells, although of course one will get the same answer either way.

Our discussion of capacitance *begins* with the observation that in this case (and the others we can solve, and other "odd" shaped conductors that we cannot) the potential of the conductor is *directly proportional to* the total charge on the conductor, and that the parameters in the potential besides the charge are k_e and things that describe its geometry, such as its physical dimensions and shape.

We could thus define a quantity we might call the "volticitance" of the conductor $\mathcal{V}$ so that (in the case of this example):

$$V = \mathcal{V}Q \tag{4.17}$$

with

$$\mathcal{V} = \frac{k_e}{R} = \frac{1}{4\pi\epsilon_0 R} \tag{4.18}$$

However, we often use conductors in particular arrangements to *store charge*. In general, we would like to be able to store a *lot* of charge on them with only a *small* potential difference. We thus seek instead a measure of the *capacity* of the conductor to store charge at any given voltage:

$$Q = CV = \left(\frac{1}{\mathcal{V}}\right) V = (4\pi\epsilon_0 R)V \tag{4.19}$$

where we have introduced the *capacitance*, the constant of proportionality that depends only on the geometry of the conductor.

To be specific, we define the *capacitance* of an arrangement of conductors used to store charge to be:

$$C = \frac{Q}{V} \tag{4.20}$$

where V is the potential difference across the arrangement as a function of the common charge Q used to create it. In the case of our example, the capacitance of an isolated conducting sphere is:

$$C = 4\pi\epsilon_0 R \tag{4.21}$$

In general the *SI units* of capacitance are easily remembered (as always) from the defining relation:

$$1 \text{ Farad} = \frac{1 \text{ Coulomb}}{1 \text{ Volt}}$$

and it is useful to remember that the dielectric permittivity of free space is:

$$\epsilon_0 = 8.85 \times 10^{-12} \text{ farads/meter}$$

which we should *also* recognize as being the natural units of ϵ_0 (or $1/k_e$) times a *length*.

Although we might have occasion to refer to the capacitance of an isolated conductor used (for example) as the storage ball on a VandeGraff generator, we will *almost always* use capacitance in the context of *specific arrangements* of *two conductors* that are designed and intended *just* to store charge in this way. Those three arrangements are:

- A **parallel plate** capacitor. This is our template model, and you should thoroughly learn it as it is quite simple and informative.
- A **cylindrical shell** capacitor.
- A **spherical shell** capacitor.

The latter two are primarily useful as teaching models, as you know everything you need to know in order to compute their capacitance from Gauss's Law and the definition of potential difference. Let's examine these three cases in some detail.

4.1.1: Computing the Capacitance: the Parallel Plate Capacitor

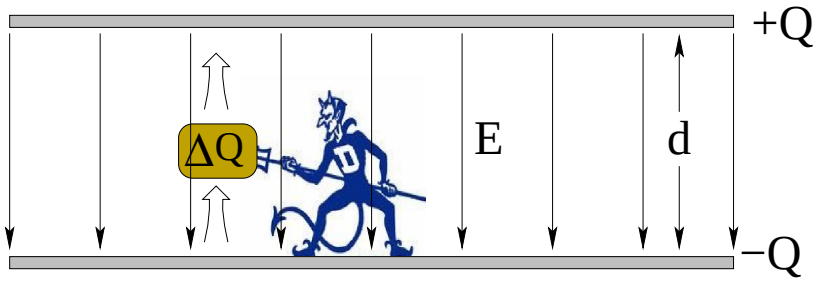


Figure 4.1: An "ideal" parallel plate capacitor of cross-sectional area A and plate separation d.

In figure 4.1 two parallel, flat, conducting plates are arranged so that they are separated by an insulating (empty/vacuum) gap d. A metaphorical "blue devil" armed with a metaphorical micro-pitchfork (that is, a still undefined process we will discuss later) forks up charge from one plate and shoves it, working against an ever increasing electric field, over to the other plate, eventually creating (after doing an amount of work that we will of course calculate shortly) the situation portrayed, with a charge $+Q$ on the lower plate and $-Q$ on the upper plate. We will invariably assume that a charged capacitor has the *same magnitude* of opposing charges on the two plates – in the static limit this is an exact result[67].

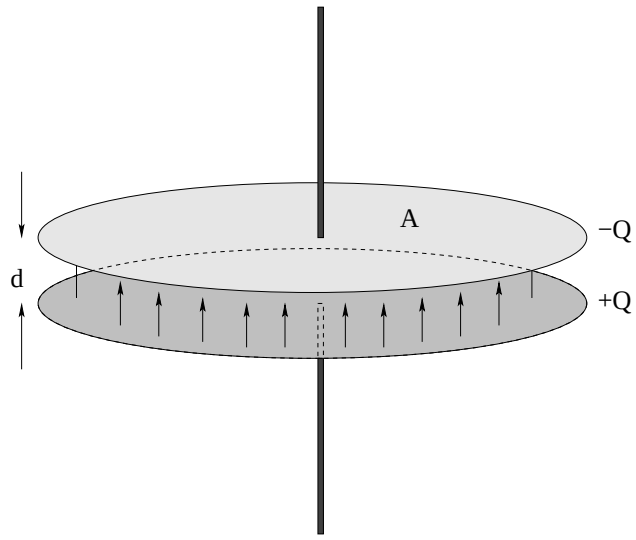


Figure 4.2: An "ideal" parallel plate capacitor of cross-sectional area A and plate separation d. In order for us to use Gauss's Law to compute the electric field between the plates, the condition $d \ll \sqrt{A}$ should hold.

Drawing all of this is a bit much, so we will idealize the figure as shown in figure 4.2, seen from a perspective that shows us the cross-sectional area A and the field between the plates. The two wires connected to the upper and lower plates are used to charge them up or connect them into a circuit.

[67] Why? Consider the properties of a conductor in electrostatic equilibrium, which requires perfect cancellation of the fields inside the conductors just inside the opposing surfaces...

We will name this arrangement a ***parallel plate capacitor*** – this is our *archetypical* capacitor, and finding the capacitance of other geometries, even when some dielectric material is inserted between the plates instead of a vacuum, will follow *exactly the same steps* illustrated below. This means that you should pay careful attention to those steps, as they reinforce pretty much everything learned in the first three chapters and will help to keep you from forgetting any of it as we move on to new material!

To compute the capacitance we execute the following steps, in order, *every time!*

a) Compute the electric field at all points in space, but in particular in between the plates, using a mix of Gauss's Law and the superposition principle. The field will, of course, be directly proportional to Q. We will idealize the field at the edges of the plates, something that is permissible if $d \ll \sqrt{A}$ and that in any event will not substantively affect their potential difference.

b) Compute the potential difference between the plates. Like the field, this will depend on the charge Q transferred from one plate to the other. Note well that we will always be computing a potential *difference* but we will often be lazy and write it as V, not bothering to add the Δ as in ΔV. It just makes the algebra a bit simpler, and keeps us from having to do the same thing for Q vs ΔQ.

c) Form the capacitance, $C = Q/V$. Note that the Q will *always cancel out* and leave us with something that depends on ϵ_0 and the *geometric* parameters of the plate. Pay close attention to the dimensions and units, as you will need to be able to tell if your answers to problems "make dimensional sense" on the fly!

So here are the steps. First we note that the charges distribute themselves (approximately) uniformly on the facing surfaces of the two plates, getting as close together as they can. This forms two equal and opposite sheets of charge with charge per unit area $\pm\sigma = \pm Q/A$. Applying Gauss's Law to either one of them, say the lower, we get:

$$\begin{aligned} \oint_S \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA &= 4\pi k_e Q_{\rm inS} \\ |E_z| 2A &= \frac{\sigma A}{\epsilon_0} \\ E_z = \frac{\sigma}{2\epsilon_0} = 2\pi k_e \sigma & \end{aligned} \tag{4.22}$$

(pointing away from the sheet of charge above and below it). We get exactly the same for the upper plate, except that the field points *toward* the negative sheet of charge.

We then apply the superposition principle using figure 4.3 as a guide. Above and below both sheets, the fields produced by the upper and lower charges *cancel*, as e.g. field from the upper one (in green) points down and the field from the lower one (in red) points up, and the fields have equal magnitudes. In between the plates, the field from the upper plate points up and so does the field from the lower one – the two fields *add*. Thus we obtain a total field of:

$$E_z = 4\pi k_e \sigma = \frac{\sigma}{\epsilon_0} \tag{4.23}$$

directed *upwards* between the plates, as drawn, and conclude that $E_z = 0$ should hold above and below the plates, at least for "infinitely wide" plates. Note well that this field is automagically

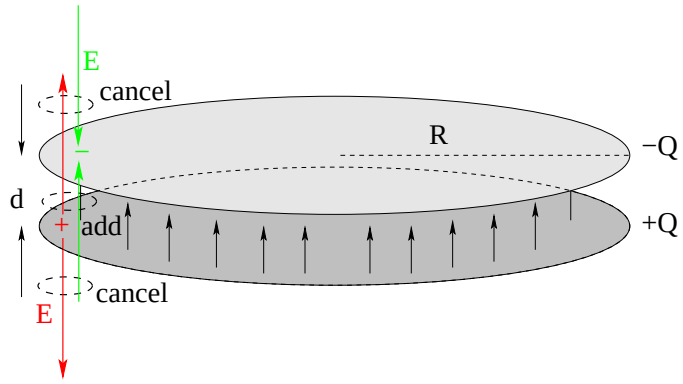


Figure 4.3: The $\vec{E}$-field in between the two oppositely charged plates *adds*, while above and below it *cancels*.

zero inside the conducting metal of the plates themselves and in the wires above and below the plates! Our assumption of charge distributing itself in two uniform sheets is *consistent* as it leads to the field vanishing inside the conductor, as we expect.

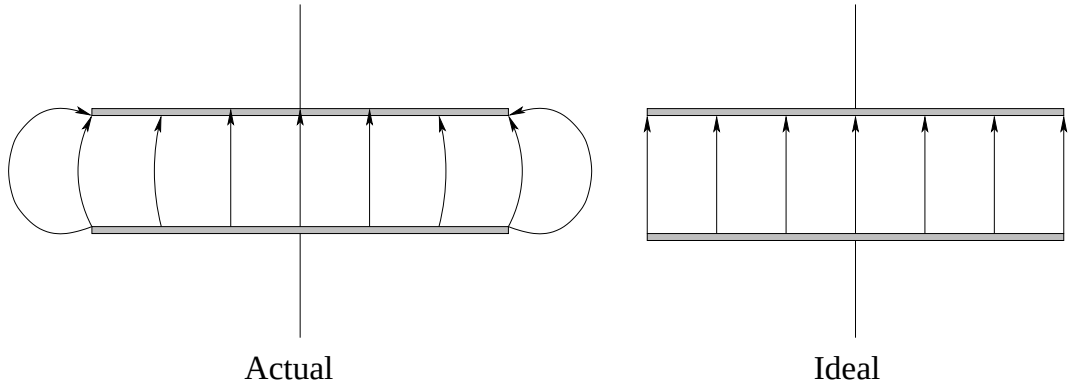


Figure 4.4: Fringe fields at the edge of an actual pair of parallel plates carrying opposite charge compared to the idealized field that vanishes sharply at the edge and is uniform in between the plates. Note that the field, and hence the potential difference, is almost identical in most of the volume between the plates.

Of course, our plates cannot *really* be infinite in area. What happens at the *edges* of the plate? There, the field "bulges" out from between the plates and forms curved field lines that resemble those of an ***electric dipole*** (because after all, the plates *do* form an electric dipole of a peculiar form). However, this "fringing field" *rapidly falls off in magnitude compared to its strength between the plates*, so much so that we won't go far wrong if we assume that it is *zero* – so that the electric field is effectively confined to the volume directly in between the facing plates. In this course we will therefore *always* idealize this by asserting that the $\vec{E}$-field "vanishes" just beyond the edges of the plates and is perfectly uniform in between, even though this isn't *precisely true*. This situation is portrayed in figure 4.4

With the fields in hand, it is but the work of a moment to compute the potential difference of the upper plate relative to the lower (or vice versa):

$$V = \Delta V = -\int_0^d E_z\, dz = -4\pi k_e \sigma d = -\frac{Qd}{\epsilon_0 A} \tag{4.24}$$

Note that the integral we computed is *negative*, which simply means that the upper plate is at a lower potential than the lower plate (consistent with the field pointing from the lower to the upper plate).

We are ready to form the capacitance. Our potential difference is negative, but when we form the capacitance we by convention make it a positive number – obviously the capacitance is symmetric and we can charge the plates in either direction, so there is no point in giving it a sign. We correspondingly form:

$$C = \frac{|Q|}{|V|} = \frac{Q}{\frac{Qd}{\epsilon_0 A}} = \frac{\epsilon_0 A}{d} \tag{4.25}$$

Note well the dependence of this *archtypical* capacitance on the dimensions of the capacitor. The *dielectric permittivity of free space* ϵ_0 appears on top and clearly has SI units (above others) of farads per meter. The capacitance varies *with* the cross-sectional area of the facing plates and *inversely with* their separation. Bigger plates (more area) means bigger capacitance; closer plates (smaller separation) also means bigger capacitance.

This is an important enough result that you should probably try to remember it *as well* as being able to derive it in detail, following all three steps outlined above. Note that this is a *great* problem to practice because this *one* problem requires you to use Gauss's Law for the electric field, the superposition principle, the definition of potential (difference) in terms of an integral of the field, the definition of capacitance, and a certain amount of common sense as far as idealization of the plate fields and the self-consistent distribution of charge in static equilibrium.

We'll now quickly indicate the key step for cylindrical and spherical capacitors, but without presenting *all* of the steps. Your very first homework problem is to fill in the missing steps *yourself*, creating "perfect" derivations of the capacitance for conducting plates with all three Gauss's Law geometries. Don't forget to draw your own figures!

Example 4.1.1: Cylindrical Capacitor

Given two concentric cylindrical conducting shells of length L and radii a and b such that $\delta = b - a \ll L$, find their capacitance. This is pictured in figure 4.5, although the figure exaggerates the size of a and b relative to each other or L. Usually the shells would be very close together, effectively trapping the field in between them everywhere but quite close to their edges.

Solution: As before, assume that they are charged up to $+Q$ on the inner and $-Q$ on the outer, perhaps by means of work done by our little blue devil dude and his charged-particle pitchfork. This puts a charge per unit length of $\pm\lambda = \pm Q/L$ on the inner and outer shell, respectively.

Again, we assume that $L \gg b - a$ so that we can use Gauss's Law to find the $\vec{\boldsymbol{E}}$-field:

$$E_r = \frac{2k_e\lambda}{r} \qquad a < r < b$$

in between the cylindrical shells and $E_r = 0$ otherwise – both for $r < a$ and $r > b$ – and as before we *neglect* the fringing fields that we expect to bulge out at the ends of the cylinders).

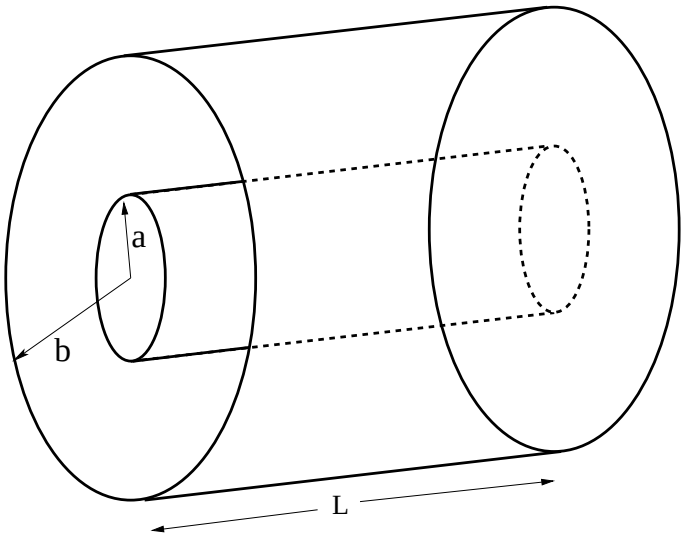


Figure 4.5: A cylindrical capacitor of length L, inner radius a, and outer radius b.

Then:

$$V = \Delta V = -\int_a^b E_r\, dr = -2k_e\lambda \ln\left(\frac{b}{a}\right) = -\frac{1}{2\pi\epsilon_0}\frac{Q}{L}\ln\left(\frac{b}{a}\right) \tag{4.26}$$

This is negative because we integrated from inside out (in the direction of the field). We could just as easily have integrated from outside in and gotten a positive potential difference. As always, the only thing that matters is that the potential must decrease when moving in the direction of the field.

The capacitance is now easy:

$$C = \frac{Q}{|V|} = \frac{2\pi\epsilon_0 L}{\ln\left(\frac{b}{a}\right)} \tag{4.27}$$

which has the right units – ϵ_0 times a length. Still, it isn't at all obvious that this has the limiting form of $\epsilon_0 A/d$. You are asked to show that it does, after all, have this form for homework. You might want to remember that $\ln(1+x) \approx x$ for $x \ll 1$ is the limiting form of the power series expansion for the natural log function when you get to this part of the first problem.

Example 4.1.2: Spherical Capacitor

Given two concentric spherical conducting shells with the radius of the inner one a and the outer one b such that $\delta = b - a$, find their capacitance. This is pictured in figure 4.6.

Solution: At this point, the steps should be familiar. Imagine a charge $\pm Q$ on the inner and outer shell respectively, put there by our intrepid devil. From Gauss's Law:

$$E_r = \frac{k_e Q}{r^2} \qquad a < r < b$$

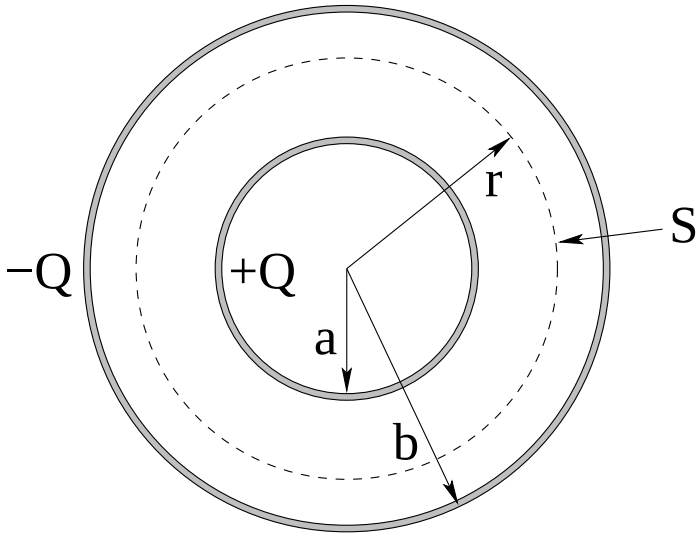


Figure 4.6: A spherical capacitor with iinner radius a, and outer radius b.

and $E_r = 0$ otherwise, with *no* idealization or fringing fields. From this we trivially find:

$$\begin{aligned} V &= \Delta V = -\int_b^a E_r dr \\ &= k_e Q \left\{ \frac{1}{a} - \frac{1}{b} \right\} \\ &= k_e Q \left\{ \frac{b-a}{ab} \right\} \\ &= \frac{1}{4\pi\epsilon_0} Q \left\{ \frac{b-a}{ab} \right\} \end{aligned} \tag{4.28}$$

This time I cleverly integrated from the outside in, *recognizing* that this would give me a positive potential difference as I integrate *against* the direction of the field. Now finding the capacitance is easy:

$$C = \epsilon_0 \frac{4\pi ab}{b-a} \tag{4.29}$$

where I've deliberately arranged it this way as a hint as to how to proceed to answer the "limiting form" part of the first homework problem.

4.2: Energy of a Charged Capacitor

It's time to compute how much work our little devil dude does shovelling charge from one plate over to the other. Imagine that he starts with the plates uncharged. The first pitchfork full of charge $\Delta Q \Rightarrow dQ$ that he moves over is "free". There is no field to push against yet. The second one, however, he must push against the field of the first one. The third one he must push against the field of the total charge of the first two. And so on.

Suppose he has been shovelling for a while on a capacitor C (where the particular geometry of the capacitor *does not matter* as long as we know the capacitance) and at this moment

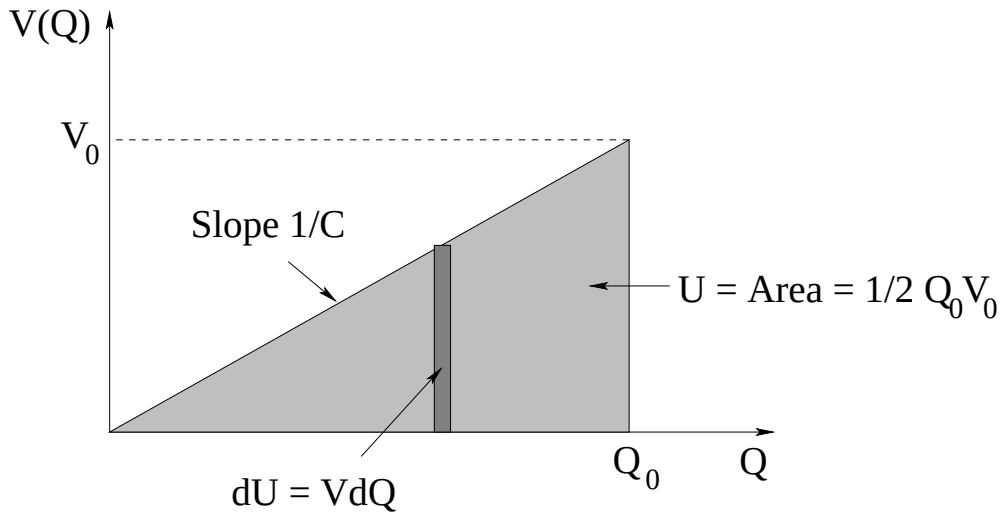


Figure 4.7: The energy as the area underneath the curve $V(Q) = Q/C$.

the total charge on capacitor plates is $\pm Q$, so that:

$$V = \frac{Q}{C} \tag{4.30}$$

is the potential difference between the plates. Then the *next* chunk of charge that he moves over with his little pitchfork (against the field/potential difference of the charge that is already there) requires him to do the *work of the devil*[68]:

$$dW_{\text{devil}} = dU_{\text{capacitor}} = V\,dQ \tag{4.31}$$

This is illustrated in figure 4.7. The work the *blue devil* does charging up the plates is *equal* to the change in the potential energy of the charged plates[69]. We therefore write:

$$dU = V\,dQ = \frac{Q}{C}\,dQ \tag{4.32}$$

and can easily *integrate both sides* to find the total energy stored on the capacitor when we begin with *no* charge and charge it up to a total charge Q_0:

$$U = \int dU = \frac{1}{C}\int_0^{Q_0} Q\,dQ = \frac{1}{2}\frac{Q_0^2}{C} \tag{4.33}$$

We can thus easily write the total energy stored *three ways*:

$$U = \frac{1}{2}\frac{Q_0^2}{C} = \frac{1}{2}CV_0^2 = \frac{1}{2}V_0Q_0 \tag{4.34}$$

(where note, we use $Q_0 = CV_0$ to go from the first to the second, then use it again to go to the third). The particular one of these that you end up using in any given problem most often depends on the givens – in some problems, you'll know Q and C; in others Q and V. My usual

[68]Metaphorically speaking, of course...

[69]Think of the work *you do* lifting a book over your head being equal to the *increase* in its gravitational potential energy – the work done by gravity, or the electric field in the case of the capacitor, is the opposite of the work done by you or the devil.

advice to students is to be *certain* to learn at least *one* of these form *plus* $C = Q/V$ – then it is easy to find the other two as needed.

Of these, the third form is perhaps the best one to learn as it has a very simple graphical interpretation. If we plot $V(Q) = Q/C$, we get a *straight line of slope* $1/C$. The integral of $dU = V\,dQ$ is just the *area* under this straight line at the particular values Q_0 and $V_0 = Q_0/C$. This, in turn, is just the area of a triangle – one half the base times the height. Which is, as you can easily see in figure 4.7, $U = \frac{1}{2}Q_0V_0$.

4.2.1: Energy Density

A very important question to ask is: just where *is* all of this energy in the capacitor stored? We did a lot of work charging up the capacitor, and all of the work we can get back comes from charge we've stored in this way being driven *by the electric field of the charge itself* back into equilibrium as the separated charges neutralize and the field collapses. It is therefore *reasonable* to guess that the energy is stored *in the electric field we create* as we rearrange the charge in the first place.

Can we write the energy of the capacitor in terms of the field strength? Yes we can! For simplicity, we'll as usual in this chapter consider the parallel plate capacitor to see how. In *this* course, we will then limit ourselves to verifying that *this works* in every case we can compute directly from the potential as well as from the electric field energy density, that is, that the result is *consistent* with the energy computed for e.g. spherical or cylindrical capacitors, or with just the energy stored creating a uniform ball of charge or spherical shell of charge. This isn't quite a proof that it is general, but it certainly seems as though it makes it more likely. We will defer to your *next* (more advanced) course in electrodynamics to derive the result more precisely, where energy conservation in electromagnetism is known as *Poynting's Theorem*[70] .

Consider, then, the energy stored in a parallel plate capacitor and write it in terms of the electric field strength:

$$\begin{aligned} U &= \frac{1}{2}CV^2 = \frac{1}{2}\frac{\epsilon_0 A}{d}(Ed)^2 \\ &= \frac{1}{2}\epsilon_0 E^2(Ad) = \frac{1}{2}\epsilon_0 E^2 \times \text{ (Volume where field isn't zero)} \end{aligned} \tag{4.35}$$

where Ad is the volume of the region in between the plates where the field is nonzero and constant in our idealized picture (neglecting fringing fields). If we divide both sides of this equation by the volume, we obtain:

$$\eta_e = \frac{dU}{dV} = \frac{1}{2}\epsilon_0 E^2 \tag{4.36}$$

the *energy density of the electromagnetic field*.

Now, as noted, we have no good reason *yet* to think that this is general and holds for varying electric fields, but it certainly might, so we try it to see if it does. Let's apply it to the

[70] Wikipedia: http://www.wikipedia.org/wiki/Poynting's Theorem. We could *almost* do the integral form of the theorem in this course, but ts proper derivation and formulation requires both Maxwell's equations in differential form and some "real" multivariate calculus in the form of *differential vector identities*...

case we just solved, the energy of a ball of uniform charge. We write:

$$\begin{aligned}
dU &= \eta_e dV = \frac{1}{2}\epsilon_0 E(r)^2 4\pi r^2 dr \\
U = \int dU &= \int \eta_e dV = \frac{1}{2}\epsilon_0 \int_0^\infty E(r)^2 4\pi r^2 dr \\
&= \frac{1}{2}(4\pi\epsilon_0)\left\{\int_0^R \left(\frac{k_e Q}{R^3} r\right)^2 r^2 dr + \int_R^\infty \left(\frac{k_e Q}{r^2}\right)^2 r^2 dr\right\} \\
&= \frac{1}{2}\frac{1}{k_e} k_e^2 Q^2 \left\{\int_0^R \frac{r^4}{R^6} dr + \int_R^\infty \frac{1}{r^2} dr\right\} \\
&= \frac{1}{2} k_e Q^2 \left\{\frac{1}{5R} + \frac{1}{R}\right\} = \frac{1}{2} k_e Q^2 \frac{6}{5R} \\
&= \frac{3}{5}\frac{k_e Q^2}{R} \qquad (4.37)
\end{aligned}$$

exactly as we obtained at the end of Week/Chapter 3! This is a rather complicated variation in $\vec{E}$, and yet it gives us exactly the right answer. This is strong evidence that our form is general (although as noted this evidence is not proof and a proper derivation of this expression is beyond the scope of this course). You will obtain still more evidence by verifying this expression for some other arrangements of charge in your homework.

4.3: Adding Capacitors in Series and Parallel

At this point, we know how to compute the capacitance of our three "simple" geometries, and know *in principle* how to proceed for more complicated cases (although the integrals and so on may be very difficult in the general case, as always). Once we've either computed or, even better, *measured* the capacitance of a capacitor, we won't really care much what the geometry is. We can start to treat a capacitor as an "object" in its own right, and give it a *symbol* to use in designing e.g. electrical circuits. Our "standard symbol" for a capacitor will be a pair of stylized "plates" viewed edgewise, with a wire running into each plate.

Let's use this symbol (and our knowledge that $C = Q/V$) and compute the *total* capacitance of *series* and *parallel* arrangements of capacitors. We'll start with series.

In figure 4.8 we see two arrangements. The top arrangement consists of three capacitors, labelled C_1, C_2, C_3, in a *line*, so that the tail of each is connected to the head of the next one by a *conducting wire* (which appears as a simple straight line in the figure). This arrangement is called *series* as each capacitor "follows" the next. Underneath this is a single capacitor labelled $C_{\rm tot}$.

We need to find what $C_{\rm tot}$ has to be for these two arrangements to behave *identically* in an electrical circuit. That is, when our devil-dude moves a charge Q from one *end* to the other *end*, we want the potential difference *between the ends* to be exactly the same. Here's how you can understand what goes on.

Suppose you have a charge $+Q$ on the leftmost plate as shown (which came from the rightmost plate in either arrangment, leaving behind a charge of $-Q$). This pair of charges creates a *field* in between. However, there can *be no field* in the conducting plates and wires

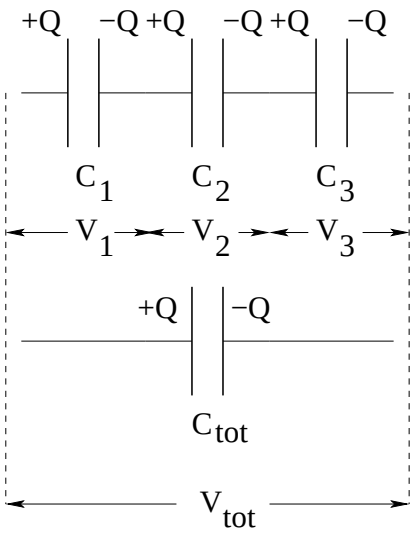


Figure 4.8: Find the total capacitance of a much of capacitors *in series*.

in the middle of the top row – they are in equilibrium! To cancel the field produced by the first plate, a charge $-Q$ is attracted to the plate facing it. But it cannot come from any part of the conducting plates or wires in between, it has to come from the surface of the next plate (leftmost of capacitor C_2) charging *it* up to $+Q$. This in turn attracts $-Q$ to the right plate of C_2, leaving a charge $+Q$ on the left plate of C_3. At this point (and you should check this) the capacitors should all be happy. Each one has a charge $\pm Q$ on it, with a field confined to live only between its plates. The field is zero inside the plates themselves and in the connecting wires. Note that all we really used in this reasoning is *charge conservation* – we couldn't create charges anywhere, only move charges around – and the idea that conductors in equilbrium can have no field inside.

Now consider the *potential differences* across each capacitor on top. Clearly the potential difference across C_1 is $V_1 = Q/C_1$, the potential difference across C_2 is $V_2 = Q/C_2$, across C_3 is $V_3 = Q/C_3$. Similarly the potential difference across our desired total capacitance is $V_{\rm tot} = Q/C_{\rm tot}$, since it has to have the *same* charge on its left plate as the arrangement on top.

Each wire between the capacitors is *equipotential*, because conductors in electrostatic equilibrium have no field inside and are thus equipotential. If we want to find the *total* potential difference across the top row of capacitors, we just have to *add up the potential difference across each capacitor*. You can think of this as doing a piecewise continuous integral across the wire at one end (get zero), the gap (pick up potential difference V_1), across the next wire (get zero), across the next capacitor's gap, (get V_2) etc. We end up with the *two* equations for the upper and lower arrangements:

$$V_{\rm tot} = V_1 + V_2 + V_3 + ... = \frac{Q}{C_2} + \frac{Q}{C_2} + \frac{Q}{C_3} + ... \tag{4.38}$$

$$V_{\rm tot} = \frac{Q}{C_{\rm tot}} \tag{4.39}$$

where the dots indicate that there was nothing special about *three* capacitors in a row – there could have been any number! We just add the potentials across as many as we have (with the same charge on each capacitor) to get the total potential difference for the series row.

These two forms must be *equal* for equal Q on the two arrangements. That's the *definition* of the total capacitance of the upper arrangement – the equivalent single capacitor one could replace the row with and get the same potential difference for the given Q. Equating them and cancelling the common Q, we get:

$$\frac{1}{C_{\rm tot}} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + ... = \sum_i \frac{1}{C_i} \tag{4.40}$$

where again the ... and final summation indicates that we just sum over as many capacitors as there are in the series row. For capacitors in series, the *reciprocal* of the total capacitance equals the sum of the *reciprocals* of the individual capacitors in series.

Why is this rule so odd? Because in series, we would get a more intuitive result by thinking of adding capacitors as if they were *volticitors*, and "volticitance" is the reciprocal of the capacitance!

Why is series addition of capacitors important and useful? Putting capacitors in series *reduces* the total capacitance (check this for yourself!) and isn't a big capacitor better than a small one? Well, yes and no. It turns out that most capacitors can only support a *finite voltage* across them before *dielectric breakdown* occurs across the intervening gap, shorting them out and burning them out. If you want to put more voltage than that maximum across a capacitor in a circuit (and don't have any rated at the desired voltage) you can put a bunch of capacitors rated at a lower voltage in series until you *can* put the desired voltage across them without exceeding the maximum for any single capacitor in the series leg. Or, you might have a bunch of big capacitors in your box and need a smaller one that wasn't in your box – adding several up in series can let you save a trip to radio shack!

So how about parallel? When several circuit elements are connected on both sides by a

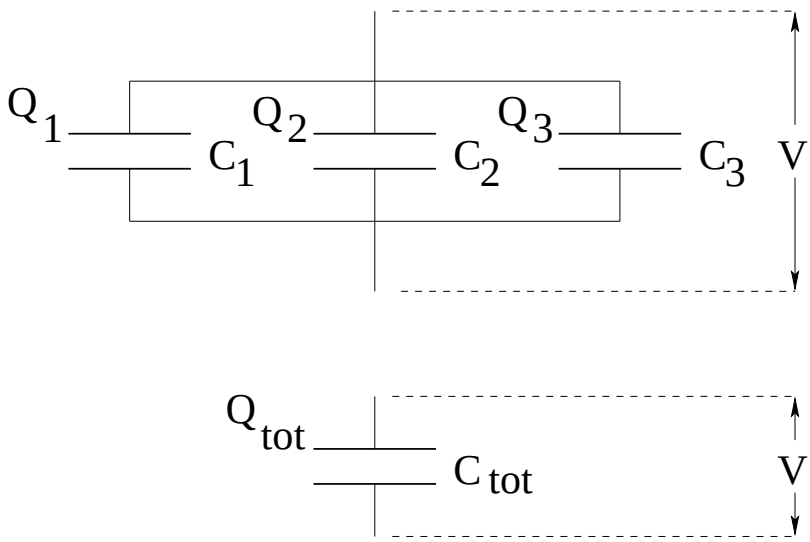


Figure 4.9: Find the total capacitance of a much of capacitors *in parallel*.

common conductor, the conductor on each side is *equipotential*. That means that all of the elements have the *same potential difference* across them. Note that this time I am not bothering to explicitly indicate the charge $-Q_1$ etc on the other plate of each capacitor. Recall, a capacitor is presumed to *always* have equal and opposite charges on its plates unless someone goes far out of their way to make up a problem with something different.

In figure 4.9 *each* capacitor in the top arrangement has a potential V across it. Therefore the first capacitor has a charge $Q_1 = C_1V$, the second has a charge $Q_2 = C_2V$, the third $Q_3 = C_3V$. The equivalent *total* capacitance $C_{\rm tot}$ with the *same* voltage V across it has a charge $Q_{\rm tot} = C_{\rm tot}V$ on it. For them to be the same, the total charge store on the top arrangement has to equal that on the bottom.

This makes the problem of finding the total capacitance really easy!

$$\begin{aligned} Q_{\rm tot} &= Q_1 + Q_2 + Q_3 + ... \\ C_{\rm tot}V &= C_1V + C_2V + C_3V + ... \\ C_{\rm tot} &= C_1 + C_2 + C_3 + ... = \sum_i C_i \end{aligned} \tag{4.41}$$

where we note that our rule works for *any* number of capacitors in series and write the final rule accordingly. Capacitors in parallel add!

We can understand these two rules intuitively in the following way. Capacitors in parallel increase the effective *area* where charge is stored, and hence just add. Capacitors in series increases the effective *separation* of the plates for a given area, and hence reduce the capacitance, adding reciprocally.

Before moving on, it is important to make one final observation. Capacitors (as we shall see) behave in electrical circuits the way *springs* behave in mechanical systems – they store energy and exert a restoring force on the charges that are stored that is *proportional to the charge*. Note well the analogy:

$$F_x = -k_s x \tag{4.42}$$

$$V = -\frac{1}{C}Q \tag{4.43}$$

where $1/C$ behaves like a "spring constant" and where the minus sign indicates that the potential created *opposes* the addition of more charge (we ignore this in the definition of C, but used it in the computation of U). If one computes the effective spring constant of *springs* in parallel or in series, one obtains very similar results. Springs in parallel add, with a total spring constant equal to the sum of the spring constants. Springs in series add as reciprocals, where the total spring constant is *less than the smallest* constant of the springs in the series.

Later we will learn that this analogy is nearly exact, after we discover the quantities which behave like "friction" or "drag forces" in circuits and even discover a quantity that behaves like a "mass". In the end we will find ourselves solving an equation that is identical in form to the damped, driven harmonic oscillator studied last semester, only this equation will yield the currents flowing in the circuit as a function of time. At that time it will be very fruitful to be thinking "the capacitor is like a spring" to help us understand what is going on.

4.4: Dielectrics

We have taken some care to study electric dipoles as the most common arrangement of matter that leads to an electric field, given the generally neutral character of matter. Indeed, all of the capacitors studied above can be thought of as stylized "dipoles" storing energy by separating

charge. We have also observed that conductors placed in an electric field polarize and create a (mostly dipolar) arrangement of surface charge that completely cancels the electric field inside. But what of insulators? They too are made up of neutral atoms and molecules, but lack the "free charges" that carry current, as the electrons associated with each molecule prefer to stay home instead of wandering off long distances under the influence of any vagrant electric field.

To understand what a neutral atom does in the presence of an electric field, it will be very useful to have a *model* of an atom. We know that an atom consists of a tiny, massive nucleus with a charge $+Ze$ where Z is the *atomic number* of the atom. Surrounding this nucleus is a "cloud" of Z electrons (for a total charge of $-Ze$ resulting in an electrically neutral atom), bound to the nucleus by the electrostatic force. We rather expect the neutral atom to be spherically symmetric in its distribution of charge so that there is little or no electric field outside of the charge cloud.

We still don't know *all* of Maxwell's equations, but when we do, we will be forced to confront the unpleasant truth that it is impossible for the electrons to be moving in "convenient" planetary-style classical orbits and for Maxwell's equations to be true. Of course we also don't know how to solve the associated quantum problem. Se we might as well construct the simplest possible model and hope that it provides us with some insight.

Example 4.4.1: The Lorentz Model for an Atom

The model we will build is a to imagine the atom to consist of a pointlike nucleus surrounded by a *uniform ball* of negative charge with a total charge of $-Ze$ and a radius a (where a is around one angstrom). This is called the *Lorentz model* for the atom, and works surprisingly well – so much so that physics graduate students still use a dynamical version to understand dielectric polarization and dispersion! See figure 4.10:

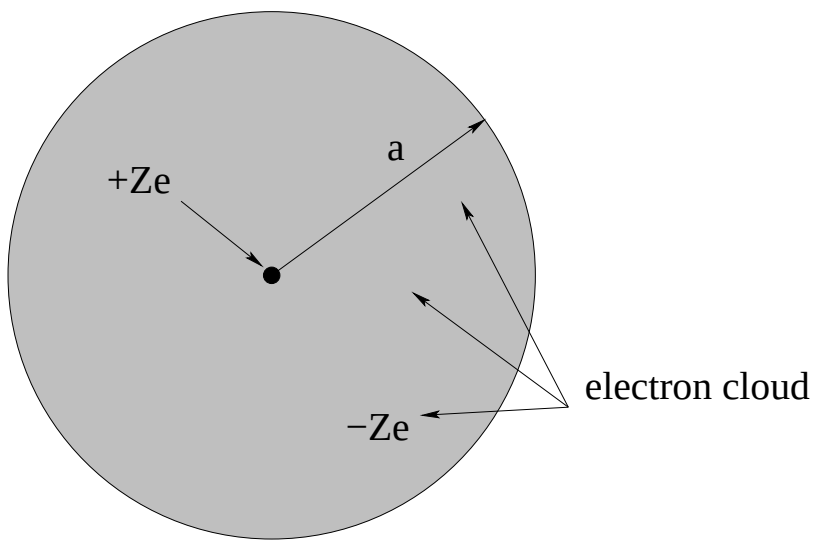


Figure 4.10: An "atom" consisting of a tiny massive nucleus surrounded by a *uniform* ball of negative charge modelling the "electron cloud".

Now we can easily *compute* what will happen when we place this atom into a "weak" electric field! We imagine that the field doesn't change the shape or size of the electron

cloud but simply diplaces the nucleus away from its equilibrium position in the center to a *new* equilibrium where the force exerted on it by the external electric field $\vec{E}_0$ balances the force on it due to the electron cloud:

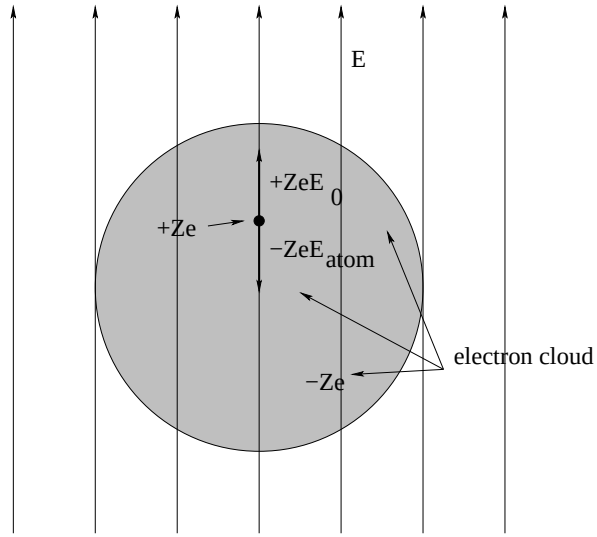


Figure 4.11: An "atom" polarized by an external electric field.

The upward field is E_0 in the $+z$ direction. The electric field of a uniform distribution of $-Ze$ in a ball of radius a is (see above or better yet, use Gauss's Law to derive it again for yourself):

$$E_{\text{atom}} = \frac{-k_e(Ze)z}{a^3} \tag{4.44}$$

(down). Thus the forces balance when:

$$+ZeE_0 - \frac{k_e(Ze)^2 z_0}{a^3} = 0 \tag{4.45}$$

We can then solve for the dipole moment of the polarized atom:

$$p_z = (Ze)z_0 = \frac{a^3}{k_e}E_0 = 4\pi\epsilon_0\ a^3\ E_0 \tag{4.46}$$

There are two very important things to note about this. One is that the polarization of the model atom is *directly proportional to the applied field*. Second, since *each* atom has a dipole moment of this magnitude, one can compute the *average* dipole moment per unit volume by dividing this estimate by the approximate volume occupied by each polarized atom in a solid or liquid or gas. We call this "dipole moment per unit volume the *polarization* of the material and give it the (vector) symbol $\vec{P}$. If (for example) we imagine a simple cubic lattice of spherical atoms, there is one atom per cube of side $2a$, with volume $8a^3$. Thus:

$$P = \frac{p_z}{8a^3} = \frac{\pi}{2}\epsilon_0 E_0 \tag{4.47}$$

where E_0 is the field in the immediate vicinity of the atom (which in general will be the field *inside* the material, not necessarily the applied external field).

There was nothing special about our guestimate of a volume of $8a^3$ per atom, and of course the actual field will probably not be exactly what we compute above in the model – we might

well expect it to depend on the kind of atom and its quantum structure, on the time dependence of the field (if any) and perhaps on still other things – but we nevertheless *expect* that the restoring force will be linear in the charge displacement for weak fields because of the usual argument, a Taylor series expansion of the energy about the equilibrium position gets a leading possible contribution from the quadratic piece, corresponding to a linear restoring force.

Overall, we expect quite generally that an insulating material will polarize, that the polarization for weak to moderate field strengths will be linear in the field, and that the order of the polarization density will be some pure number times $\epsilon_0 E$. We give that *dimensionless* number a special name and its own symbol – we call it the *electric susceptibility* χ_e such that:

$$\vec{\boldsymbol{P}} = \chi_e \epsilon_0 \vec{\boldsymbol{E}} \tag{4.48}$$

Note well that the units of polarization are *coulombs per square meter* – those of *surface charge density*. It remains to find a surface for which the polarization tells us a surface charge density.

To continue our observations above, χ_e will, in general, be characteristic of the material; it will depend on whether the material is solid or liquid or gas (gases usually have a very weak polarization response because of the large volume occupied per atom) and of course upon the neglected details of the material in our model – the quantum structure and/or molecular structure of the material. For solids and liquids it will generally be of the order of unity – in our example, $\chi_e = \pi/2 \approx= 1.5$ – where for gases it will usually be "small" as there simply aren't a lot of atoms or molecules per unit volume, so no matter how well they polarize individually you won't build up much of a polarization density.

We are only interested in the static limit of the susceptibility in this *intro* course, but it really depends on the *time dependent behavior of the electric field*, on temperature, and much more. It takes the charge in a real material *time* to respond to changes in the applied field and response times depend on the natural frequencies and damping times of the charges that are responding. Many physicists have spent their entire careers studying quantities that amount to general susceptibilities for various materials (which can have very odd properties indeed!)

4.4.1: Dielectric Response of an Insulator in an Electric Field

Now that we understand what *each* atom in an insulating material does when the material is placed in an external field, let's try to understand what the material *as a whole* does – in particular, what happens to the electric field inside, which is now the *sum* of the external field and the field produced by all of those dipoles!

In figure 4.12, we see an imaginary lattice of atoms, all polarized by an external field in the direction indicated. Note well that we've erased the *details* of even our simple model – we represent each atom as a neutral object with a small dipole moment where "some" charge is split by "some" distance by the general *process* derived and discussed in the previous section. We've drawn several possible Gaussian Surfaces inside the material.

Now let use Gauss's Law. On the *inside*, if we draw any Gaussian Surface S large enough to contain "many atoms", since the atoms are neutral the average charge inside will be *zero*[71].

[71] If it contained an integer number of whole atoms, it would be exactly zero. If the surface cuts through atoms

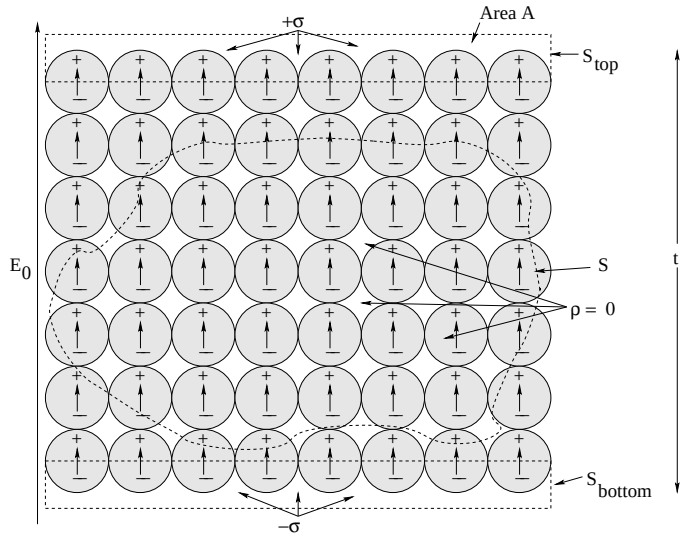


Figure 4.12: A lattice of atoms polarized by an external electric field.

Note that even where it contains an extra charge or two of either sign by splitting an atom, those charges are almost always paired with charges above or below on the neighboring atoms and the bulk remains neutral, with an average charge density $\rho \approx 0$. The interior atoms, then, do not directly modify the average field.

This is *not* true on the *surface*. If we draw a Gaussian surface $S_{\rm top}$ so that it just contains the upper half of the polarized atoms we see that it contains a nonzero positive charge; inside a similar surface $S_{\rm bottom}$ on the lower surface there is an equal and opposite negative charge. These charges make up a *surface charge layer* with a surface charge density $\pm\sigma_b$ that is directly proportional to E, the net field in the medium.

Note Well: I put a subscript "b" on σ to indicate that this kind of "surface charge" produced by the polarization of ***neutral insulator atoms or molecules*** where the plus and minus charge is "bound" together and not "free" to move as it is in a conductor is generally referred to as ***bound charge***. We will only consider bound *surface* charge σ_b in this course as the most common important case, but in principle one can generate bound bulk charge distributions ρ_b.

In contrast, the charge we have discussed up to this point is primarily is "bare", isolated, normal, unbalanced charge, the charge that is directly producing electric fields or potentials that we have evaluated various ways. In contexts where both are present, we will usually differentiate them by means of "f" and "b" subscripts: A net free charge might be referred to as Q_f and a net bound charge might be referred to as Q_b. Now, back to the thread of our discussion.

---

to include or exclude some of their charge, the surplus charge is limited to be some fraction of the charge on the atoms on the surface. But the number of atoms on the surface scales with the characteristic length scale of the volume D like D^2 where the volume inside the surface scales like D^3, so the *average* charge scales smoothly to zero as the volume gets larger.

Let us understand this in this particularly simple case, where the upper and lower surfaces are conveniently perpendicular to the field and the cross-section of the material is rectangular. The total dipole moment of the system is given by the total charge on the upper or lower surface, times that thickness (recall that all the charges in between sum to zero). That is:

$$p_{\text{system}} = Q_{\text{surface}} t = (\sigma_b A)t = PV = P(At) \tag{4.49}$$

(all in the direction of the field) or clearly:

$$\sigma_b = P \tag{4.50}$$

This argument is actually more general than one might suspect – if you think about it in terms of calculus you can see why it would be true for less conveniently shaped objects in a uniform field and how it might be changed to accomodate an angle between the polarization density direction at a surface and the normal to the surface there. In any event, the modifications of the field we deduce from this below are reasonably general and hold for arbitrary objects in nearly arbitrary fields[72].

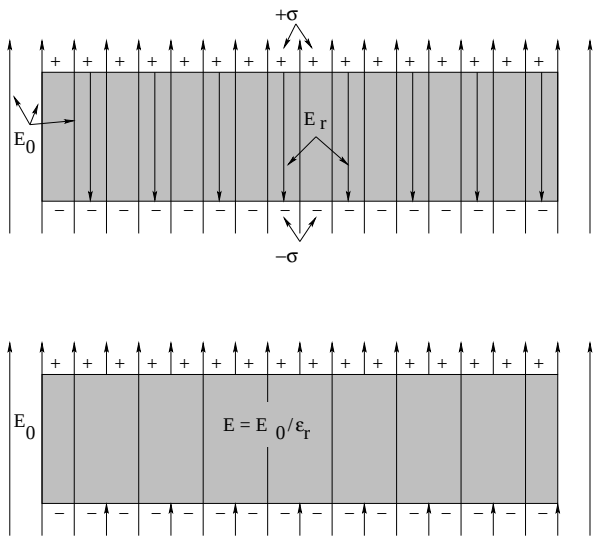


Figure 4.13: The polarized material generates a *reaction field* E_r that *opposes* the applied field and partially cancels it, making the total field in the material smaller. A dielectric material thus *reduces* the applied electric field inside the material.

Now let's imagine this figure redrawn on a length scale where atoms are *tiny* – too small to be seen in the figure (as they are in any macroscopic chunk of matter large enough to be seen with the naked eye). When we consider the field between the surface charge layers, the block of matter starts to look like, and behave like, a *capacitor* internally, with a reaction field E_r that flows from the positive to the negative charge layers in the *opposite direction to the applied external field*. This situation is portrayed in figure **??**.

Applying Gauss's Law to the induced surface charge layers in this simple rectangular geometry, we expect:

$$E_r = \frac{\sigma_b}{\epsilon_0} \tag{4.51}$$

[72]Truly advanced students might look ahead at a book on electrodynamics and learn about how this statement is not precisely true and how polarization density itself both satisfies certain partial differential equations and how our entire picture at this level relies on a ***linear*** response that is at best an (often quite good) approximation.

The total field is then:

$$E = E_0 - \frac{\sigma_b}{\epsilon_0} = E_0 - \frac{P}{\epsilon_0} = E_0 - \chi_e E \tag{4.52}$$

We can rearrange this into:

$$E(1 + \chi_e) = E_0 \tag{4.53}$$

and solve for E, the field inside the material, in terms of E_0, the applied external field:

$$E = \frac{E_0}{1 + \chi_e} = \frac{E_0}{\epsilon_r} \tag{4.54}$$

where we have introduced the *relative permittivity*

$$\epsilon_r = (1 + \chi_e) \tag{4.55}$$

as a dimensionless constant characteristic of the material. Note that $E \leq E_0$ because $\chi_e \geq 0$. This also means that $\epsilon_r \geq 1$! The electric field is *reduced* inside a dielectric – this is what the "di-" in "dielectric" means!

Note Well: Most introductory physics books written for college or high school physics courses omit any explicit mention of the susceptibility (leaving students with quite a chore later if they go on in physics and have never seen it the next time they take electricity and magnetism) and use the symbol $\boldsymbol{\kappa}$ to represent $1 + \chi_e$ and call it the *dielectric constant* for the material, as in:

$$\kappa = (1 + \chi_e) = \epsilon_r \tag{4.56}$$

This usage is deprecated even in introductory treatments because in general neither ϵ_r nor κ are ***constant*** (doh!) and because it encourages confusion with the sensible definition of the ***permittivity of the material***. We therefore use ϵ_r exclusively in this textbook.

This may seem very confusing to you, so let me review. ϵ_0 is functionally equivalent to k_e, a constant of nature that connects the units of charge and length to those of field and force at the microscopic scale of elementary particles (or in a vacuum), where of course $k_e = 1/(4\pi\epsilon_0)$. The presence of bulk neutral matter *modifies* the electric field $\vec{\boldsymbol{E}}_0$ produced by bare/isolated/free charges Q_f that *would* be there in a vacuum; the field *polarizes* the material, which creates a reaction field that strictly reduces the applied field inside the material. The polarization density (dipole moment per unit volume) of the medium is related to the *net* field in the medium $\vec{\boldsymbol{E}}$ by $\vec{\boldsymbol{P}} = \chi\epsilon_0\vec{\boldsymbol{E}}$. The net field itself is related to the applied field by $\vec{\boldsymbol{E}} = \vec{\boldsymbol{E}}_0/\epsilon_r$ where $\epsilon_r = 1 + \chi$.

There is one more thing we can do with the ***relative*** permittivity, the thing that gives it its name. We can use it to define the ***permittivity of any medium***:

$$\epsilon = \epsilon_r\epsilon_0 \tag{4.57}$$

This form proves to be most useful in the more advanced treatments of electrodynamics that e.g. physics majors will take that build on this course, but is beyond the scope of this

course. It is still worth reading about in passing for "culture", or to plant a seed or two that might flower later if you continue studying physics. If this does not describe you (and it well might not!) feel free to skip the material between the next two separator lines.

We see that the field produced by the usual free charge we considered in the first three chapters changes form "suddenly" – is ***displaced*** – at the surface of neutral dielectric materials. It is useful to define a new field, closely related to the electric field (and force) experienced by a bare test charge anywhere in space in a medium of some sort or not. We will think of this new field as being produced *only* by bare unbalanced charge, and *explicitly exclude* from consideration the "bound" neutral charge that we have been discussing above. We will call this non-bound charge *free* charge. This field *will not change form* as it propagates from one material to another!

The field in question is called the ***electric displacement***:

$$\vec{\boldsymbol{D}} = \epsilon \vec{\boldsymbol{E}} \tag{4.58}$$

Note well that this is a very odd name. One would be inclined to call the *reaction field* produced by the surface bound charge the "displacement" of the vacuum field inside a medium, but ***this is incorrect***. On the other hand, the electric displacement ***does not change*** at the surface of a dielectric medium, totally counterintuitively! This drove me batty for years of study as a physics major and even as a graduate student because it is some sense an abuse of the English language.

Don't fight it, accept it! The electric displacement "is what it is" according to this definition, and is the *un*-displaced version of the electric field. Sure, it might have been more useful and descriptive to call it the "charge field", but we are at this point all stuck with the name, so if you plan to go on in physics you might as well learn it.

The fundamental advantage of this electric displacement (field) is that we can write Gauss's Law *anywhere*, inside a dielectric, conductor, or vacuum, in a form that depends *only* on the free charge present, not on any dielectric response of the medium. Since we've cancelled out all dependence on permittivity, this form is just:

$$\oint_S \vec{\boldsymbol{D}} \cdot \hat{\boldsymbol{n}} dA = \int_{V/S} \rho_f dV \tag{4.59}$$

where ρ_f is the free charge density only. Note the *absence* of any form of the dielectric permittivity! If we solve this, we can find the resulting field inside any linear medium by just dividing $\vec{\boldsymbol{D}}$ by $\epsilon = \epsilon_r \epsilon_0$.

Following this reasoning, the electric displacement of a point charge is *even simpler* than the electric field of a point charge in charge centered coordinates:

$$\vec{\boldsymbol{D}} = \frac{1}{4\pi} \frac{Q}{r^2} \hat{\boldsymbol{r}} \tag{4.60}$$

Note well the absence of ϵ_0! The displacement itself has the units of charge per unit area and completely captures the *geometry* of Gauss's Law, but it is a *vector* that does not correspond

in any way to an actual surface charge density. In some sense it corresponds to the *imaginary* (as in pretend, not complex) surface charge density one would get if one took the central charge, ***displaced it uniformly*** by a distance r, producing the same charge smeared out uniformly over the spherical surface of radius r, and then made it a vector directed outward for positive charge an inwards for negative charge.

All clear now? Well, probably not so much. Possibly even as clear as mud! But if you think about it even a bit now, and pay attention to my warnings about the undisplaced displacement field that depends only on the free charge and never on the dielectrics present, it will make the more mathematically involved treatments of this in intermediate and advanced electrodynamics a whole lot easier later.

4.4.2: Dielectrics, Bound Charge, and Capacitance

At this point you hopefully understand how a dielectric insulator is polarized by a field, how the polarization appears as a surface charge layer, how the surface charge creates a reaction field that opposes the applied field and reduces it inside the dielectric so that we can wrap *all of that up* in the simple relation:

$$E_{\text{material}} = \frac{E_0}{\epsilon_r} \tag{4.61}$$

where ϵ_r is the relative dielectric permittivity of the material. It seems like a good time to list a few useful relative permittivities in a table:

Material	ϵ_r	Dielectric Strength (MV/m)
Vacuum	1	20 - 40
Air	1.00006	0.4 to 3.0
Paper	3.5	
Silicon Dioxide (Quartz)	3.9	
Glass	3.7 to 10	9.8 to 13.8
Water	80	30 (Ultra-pure)
Polyethylene	2.25	
Ethylene Glycol	37	
Strontium titanate	310	
Barium strontium titanate	500	
Barium titanate	1250	

Table 2: Table of relative dielectric permittivities at room temperature (20° C) and some associated dielectric strenths.

So fine, so what are dielectrics *good* for? Dielectric insulators are often inserted between the plates of capacitors! Dielectrics have *three purposes* in capacitor design:

a) They mechanically separate the plates.

b) They increase the capacitance.

c) They prevent dielectric breakdown (most dielectrics have a dielectric strength greater and more reliable than that of air, which is relatively small and varies with pressure and humidity).

You can easily experience all three benefits by *building your own capacitor.* Take a roll of aluminum foil, and cut two square pieces 10 cm by 10 cm. Use tape to fasten an unbent paper clip to each one. Cut a piece of white printer paper 12 cm by 12 cm.

For grins, try setting up the two pieces of foil so they are separated by a perfect 0.1 mm air gap. Don't worry, if you wreck the foil you can cut new pieces. Can't do it, right? And if you did, somehow, manage it, the first time you put an equal and opposite charge on the "plates" they would *attract*, and being as how they are made out of *foil*, they'd bend until they touched, pop, end of capacitor.

Now just lay down one sheet of foil on the table. Cover it (symmetrically) with the paper. Top it with the second piece of foil. Tape the foil to the paper on both sides. Congratulations! You've made a capacitor! When the foil is pressed tight to the paper, the gap d is roughly 0.1 mm (a ream of 500 sheets of printer paper is roughly 5 cm = 50 mm thick) and has an area $A = 0.1^2 = 0.01$ square meters. The paper prevents the paper from touching and is more resistant to arcing than 0.1 mm = 10^{-4} meters of air!

To compute the capacitance, we have to solve the parallel plate capacitor problem all over again. Suppose you put a charge $\pm Q$ on your capacitor (e.g. moving a net charge Q from one plate and putting it on the other). This charge is ***free charge***, unbalanced charge that distributes itself on the conducting plates of the capacitor, so perhaps we should refer to it as Q_f to cleanly differentiate it from bound charge on the surface of the dielectric paper.

The capacitor plates have an area A, so the magnitude $\sigma_= Q_f/A$ and Gauss's Law tells you that the magnitude of the field in between the plates if there were *no* paper there would be:

$$E_0 = 4\pi\epsilon_0\sigma_f = \frac{\sigma_f}{\epsilon_0} \tag{4.62}$$

However, now there *is* a dielectric in that space. The field is modified to become:

$$E = \frac{E_0}{\epsilon_r} = \frac{\sigma_f}{\epsilon_r\epsilon_0} = \frac{\sigma_f}{\epsilon} \tag{4.63}$$

Next, we compute as usual the potential difference:

$$V = -\int_d^0 \frac{Q_f}{A\epsilon_r\epsilon_0}dz = \frac{Q_f d}{A\epsilon_r\epsilon_0} \tag{4.64}$$

and the capacitance:

$$C = \frac{Q_f}{V} = \epsilon_r\frac{\epsilon_0 A}{d} = \epsilon_r C_0 \tag{4.65}$$

where (**note well!**) the definition of capacitance involves only the *free* charge on the plates, as that is the charge we actually moved around charging it and where C_0 is the capacitance of the same plate geometry *without* the dielectric!

Recall that $\epsilon_r > 1$. We see that the presence of a dielectric between the plates *increases the capacitance* compare to a vacuum, or air, between the plates, *in addition* to mechanically

separating the strongly attracting plates and prevenint dielectric breakdown. So what (approximately) is the capacitance of our homemade capacitor?

That's left as an exercise, a few seconds work with a calculator. To save you a bit of time for this *estimate*, you can assume that

$$\epsilon_0 = 8.85 \times 10^{-12} \approx 10^{-11} \frac{\text{farads}}{\text{meter}} \tag{4.66}$$

and now you can probably do the estimate *without* a calculator!

Before we move on, we need to do one final thing: relate the *free* surface charge that we put on the actual conducting plates of our parallel plate capacitor with a dielectric to the *bound* surface charge that appears on the polarized dielectric in the resulting field. We can easily do this with Gauss's Law or equivalently with our knowledge of the free field and the reaction field in terms of the surface charges.

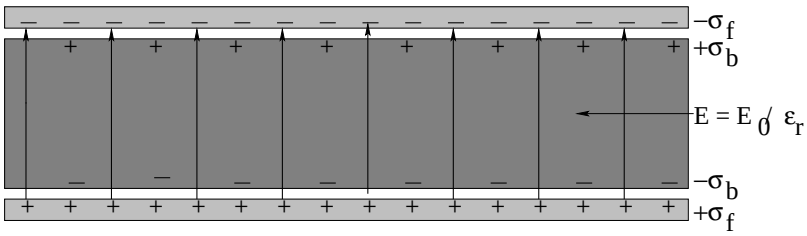


Figure 4.14: Bound and free charge in a capacitor filled with a dielectric.

In figure 4.14 we can write the field in the dielectric in two ways:

$$E = \frac{E_0}{\epsilon_r} = E_0 - E_r \tag{4.67}$$

where recall that E_r is the reaction field generated by the surface charge σ_b, which is also equal to the local polarization density at the surface. If we write out the fields E_0 and E_r in terms of the charges that produce them (basically using Gauss's law on the two surface charges), we get:

$$\frac{4\pi k_e \sigma_f}{\epsilon_r} = 4\pi k_e \sigma_f - 4\pi k_e \sigma_b \tag{4.68}$$

If we cancel out the common factor of $4\pi k_e = 1/\epsilon_0$, we get:

$$\frac{\sigma_f}{\epsilon_r} = \sigma_f - \sigma_b \tag{4.69}$$

or

$$\begin{aligned} \sigma_b &= \left(1 - \frac{1}{\epsilon_r}\right)\sigma_f \\ &= \left(\frac{\epsilon_r - 1}{\epsilon_r}\right)\sigma_f \\ &= \left(\frac{-\chi}{1+\chi}\right)\sigma_f \end{aligned} \tag{4.70}$$

where the last form is in terms of the material's susceptibility instead of the more commonly used ϵ_r.

Note that an alternate, perhaps simpler, route to this relation is through the observation that the *magnitude* of the bound surface charge density $\sigma_b = P = \epsilon_0 \chi_e E$ (from our previous discussion of polarization density and the definition of the susceptibility).

$$\begin{aligned} \sigma_b &= \epsilon_0 \chi_e E \\ &= \epsilon_0 \chi_e \frac{E_0}{\epsilon_r} \\ &= \epsilon_0 \chi_e \frac{\sigma_f}{\epsilon_0 \epsilon_r} \\ &= \sigma_f \frac{\chi_e}{1+\chi_e} \end{aligned} \tag{4.71}$$

where we once again used $\epsilon_r = 1 + \chi_e$ by definition. In this case one must put in the sign relation (the bound charge always has the opposite sign of the free charge that it faces) by hand.

We see that the bound surface charge on the dielectric σ_b is closely related to the free surface charge σ_f on the actual plate of the conductor. Note well that $Q_f = \sigma_f A$ is the actual charge *stored* on the conductor, but the presence of the bound charge layer reduces the field that charge produces across the dielectric and therefore reduces the potential difference between the plates of the capacitor for any given charge. This is, by definition, an increase in the capacitance of the arrangement – more charge stored per volt of potential difference.

Although we've done all of our derivation and examples in the cases above in the context of a parallel plate capacitor, they hold in the *general* case for fields in materials, even where the fields vary. The electric field in a medium is *always* given by $E = E_0/\epsilon_r$, even where the field is varying as a function of coordinates. This latter derivation has the advantage in that the first two lines hold for *any* source of the free-space field E_0, not just a presumed external parallel plate capacitor with its uniform field. For example, if we surround a bare point charge with a dielectric shell as portrayed in figure 4.15:

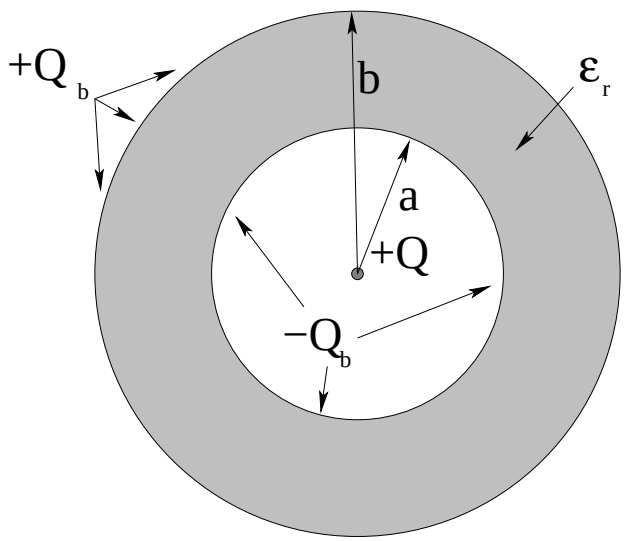


Figure 4.15: A bare charge $+Q$ surrounded by a dielectric shell with relative permittivity ϵ_r.

Hopefully we all know quite well at this point that the "bare" field of the free charge $+Q$ in

the center is just

$$E_r = \frac{Q}{4\pi\epsilon_0 r^2}$$

From the reasoning above:

$$\sigma_b = -\chi_e \frac{Q}{(4\pi a^2)\epsilon_r} = -\frac{Q(\epsilon_r - 1)}{(4\pi a^2)\epsilon_r} \tag{4.72}$$

$$\sigma_b = +\chi_e \frac{Q}{(4\pi b^2)\epsilon_r} = +\frac{Q(\epsilon_r - 1)}{(4\pi b^2)\epsilon_r} \tag{4.73}$$

Note well that the *total* bound charge on *either* surface has magnitude:

$$Q_b = Q\frac{\epsilon_r - 1}{\epsilon_r} \tag{4.74}$$

The charge on the inner surface reduces the field produced by Gauss's Law "just right" to produce a field of E/ϵ_r in the dielectric; the charge on the outer surface puts it back so that the usual field obtains outside of the dielectric sphere!

Advanced: This can safely be skipped to the next separator line if you are not a physics major.

Before we go on to energy density, we should at least put down the more advanced relations that you will derive and learn in a more advanced course in Electrodynamics and hint at how such a derivation would proceed. Suppose $\hat{\boldsymbol{n}}$ is a normal unit vector perpendicular to a dielectric surface, where the polarization density is e.g. $\vec{\boldsymbol{P}} = \epsilon_0\chi_e\vec{\boldsymbol{E}}$ for $\vec{\boldsymbol{E}}$ just **inside** the material. Then σ_b, which is a ***scalar***, is given by:

$$\sigma_b = \vec{\boldsymbol{P}} \cdot \hat{\boldsymbol{n}}$$

Our treatment above was valid for the special case that $\hat{\boldsymbol{n}}||\vec{\boldsymbol{P}}$, but note that the dot product gets the sign of σ_b right *and* corrects for the "tilt" of the surface relative to the field! We had to put the former in "by hand" above, and had no clue about the latter (although you can show it easily enough if you recapitulate the original argument connecting σ_b to P above for a tilted surface).

The last important relation involving bound charge is well beyond the scope of this course to discuss, but note well that one *can* in principle generate a dielectric material with nonzero ***bulk*** bound charge, that is, with a bound charge *density* ρ_b distributed throughout the material itself and not just confined to the surface. In this case, the polarization density becomes a function of this bound charge that is given by solving:

$$\vec{\boldsymbol{\nabla}} \cdot \vec{\boldsymbol{P}} = -\rho_b$$

or equivalently:

$$\oint_S \vec{\boldsymbol{P}} \cdot \hat{\boldsymbol{n}} dA = -\int_{V/S} \rho_b dV$$

(the two are equivalent due to the divergence theorem).

This expression looks a lot like Gauss's Law, but for the polarization density, which in turn is related to the local field, which is in turn related to the total charge inside Gaussian surfaces,

and in fact one derives this expression in the next course up from this one by considering all of these things and working out how the total field is modified by the presence of a (e.g. linear response) dielectric material *and* extra bound charge distributed through the dielectric.

That is:

$$\vec{\boldsymbol{E}} = \vec{\boldsymbol{E}}_0 - \frac{\vec{\boldsymbol{P}}}{\epsilon_0}$$

(see above) and if we take the divergence of both sides:

$$\vec{\boldsymbol{\nabla}} \cdot \vec{\boldsymbol{E}} = \frac{\rho_{\rm tot}}{\epsilon_0} = \vec{\boldsymbol{\nabla}} \cdot \vec{\boldsymbol{E}}_0 - \frac{\vec{\boldsymbol{\nabla}} \cdot \vec{\boldsymbol{P}}}{\epsilon_0} = \frac{\rho_f + \rho_b}{\epsilon_0}$$

where we used $\vec{\boldsymbol{\nabla}} \cdot \vec{\boldsymbol{E}}_0 = \rho_f/\epsilon_0$ from Gauss's Law for the free charge only. It all works out just as it should!

So much to look forward to, if you are going on in physics!

As a last remark, consider field energy density inside a dielectric. If we recapitulate the argument for field energy density for a parallel plate capacitor filled with a dielectric, we get:

$$U = \frac{1}{2}CV^2 = \frac{1}{2}\frac{\epsilon_r\epsilon_0 A}{d}(Ed)^2 \tag{4.75}$$

where E is still the field between the plates, in this case the field inside the dielectric. Hence

$$\eta_e = \frac{dU}{dV} = \frac{1}{2}\epsilon E^2 \tag{4.76}$$

where $\epsilon = \epsilon_r\epsilon_0$ is the dielectric permittivity of the material. This is the correct form of the energy density to use inside a linear dielectric material.

This is all we need to know about dielectrics, although the problems below will challenge you with half-filled capacitors and the like to make sure you understand it well enough to be able to use it.

Homework for Week 4

Problem 1.

Physics Concepts

Make this week's physics concepts summary as you work all of the problems in this week's assignment. Be sure to cross-reference each concept in the summary to the problem(s) they were key to. Do the work carefully enough that you can (after it has been handed in and graded) punch it and add it to a three ring binder for review and study come finals!

Problem 2.

Derive the capacitance for:

a) A spherical capacitor with inner conductor radius a and outer conductor radius b.

b) A cylindrical capacitor with inner conductor radius a, outer conductor radius b, and length L (where $L \gg b - a$);

c) A parallel plate capacitor with cross-sectional area A and plate separation d;

d) Show that in the first two cases that the capacitance is approximately:

$$C \approx \frac{\epsilon_0 A}{d}$$

(the answer to the third case) where A is the area of the cylinder/sphere and $d = b - a \ll a$ ("small" separation). You will probably need to use the power series expansion

$$\ln(1 + x) \approx x - \frac{x^2}{2} + \frac{x^3}{3} + ...$$

– a result worth remembering – to do the cylinder.

Problem 3.

Prove that the energy stored on a capacitor with a charge Q at a potential difference $V = Q/C$ can be evaluated:

a) In terms of the *work required to charge it* using e.g. $dU = V\,dQ$;

b) Using the the *energy density*: $U = \int_{\mathcal{V}} \eta_e d\mathcal{V}$ where $\eta_e = \frac{1}{2}\epsilon_0 E^2 = \frac{dU}{d\mathcal{V}}$.

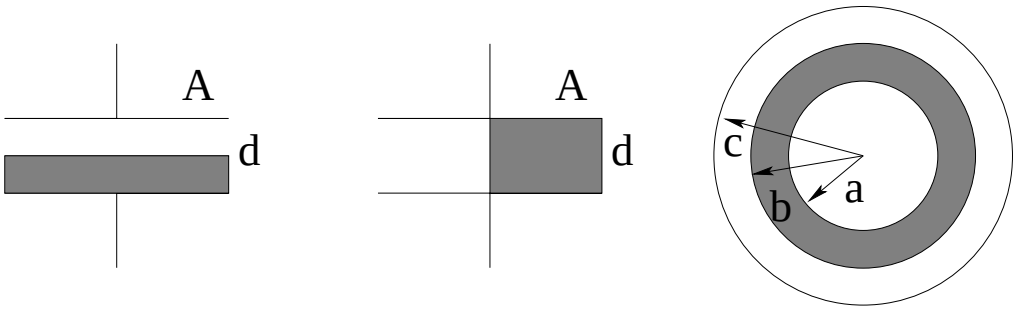
to show that

$$U = \frac{1}{2}QV = \int_{\mathcal{V}} \frac{1}{2}\epsilon_0 E^2 d\mathcal{V}$$

for ***all three geometries*** (where the energy density integral is over the volume $\mathcal{V}$ between the plates).

Problem 4.

Find the capacitance of the following arrangements:



where the first two are parallel plate capacitors *half-filled* with a dielectric material with relative dielectric permittivity ϵ_r as shown, and the third is a spherical capacitor *partially* filled with the same dielectric as shown.

Problem 5.

Derive the rules for adding parallel and series capacitance:

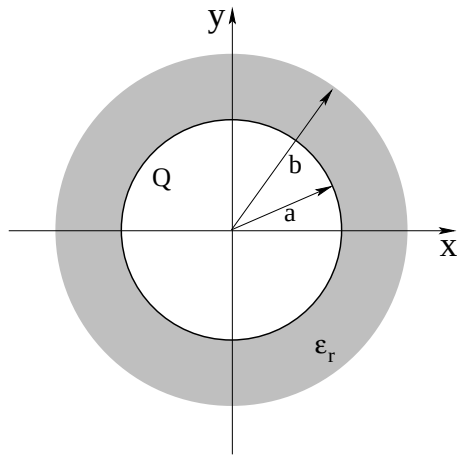
$$\frac{1}{C_{\text{tot}}} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + ... \qquad \text{(series)}$$

and

$$C_{\text{tot}} = C_1 + C_2 + C_3 + ... \qquad \text{(parallel)}$$

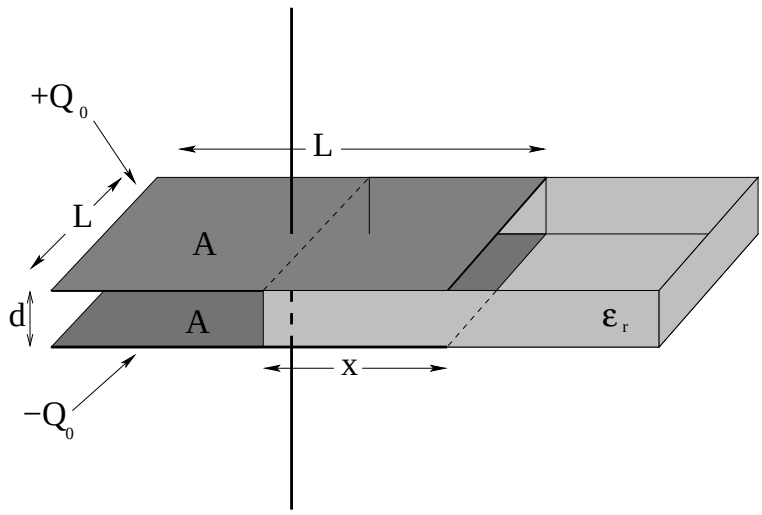
Try not to look at/copy the text as you do so! The idea is to learn both the rules and how to obtain them at the same time by *doing it!*

Problem 6.



A conducting sphere of radius a has a charge Q on it. It is surrounded by a spherical insulating dielectric shell of inner radius a, outer radius b and relative dielectric permittivity ϵ_r. Find the electrostatic field in all space, the potential in all space, and the bound surface charge on both surfaces of the dielectric in terms of the givens.

Problem 7.



You are given a square parallel plate capacitor of side L and plate separation d and a slab of dielectric material with relative dielectric permittivity ϵ_r that exactly fills the volume between the plates if fully inserted. At the moment, however, the slab is inserted only a distance x. The capacitor has a *constant* free charge Q_0 on it.

a) Find the potential energy of the partially filled capacitor, as a function of $Q_0, L, d, \epsilon_0, \epsilon_r$ and x.

b) Is the potential energy minimal when the dielectric slab is fully inserted or fully removed? Explain why.

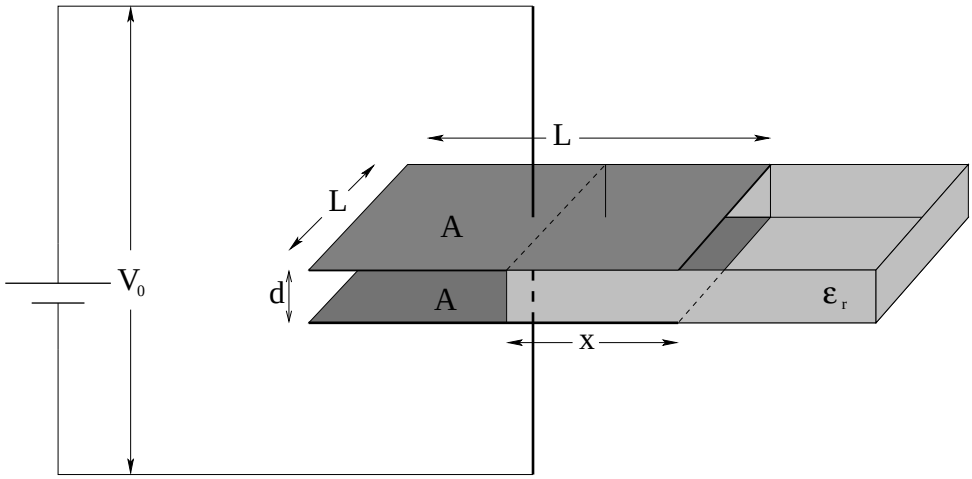
c) By using

$$F_x = -\frac{dU}{dx}$$

find the force on the partially inserted dielectric slab. Does the force pull the dielectric slab in (to fill the plate volume) or does it push it out from between the plates?

d) Draw a simple picture involving the probable bound charge distribution on the partially inserted dielectric slab that physically explains this force.

Advanced Problem 8.



You are given a square parallel plate capacitor of side L and plate separation d and a slab of dielectric material with relative dielectric permittivity ϵ_r that exactly fills the volume between the plates if fully inserted. At the moment, however, the slab is inserted only a distance x. The capacitor has a *constant voltage* V_0 connected across it that can *do work* adding charge to or taking charge away from the capacitor as the slab is inserted or removed (!).

a) Find the potential energy of the partially filled capacitor, as a function of $V_0, L, d, \epsilon_0, \epsilon_r$ and x.

b) Draw a simple picture involving the probable bound charge distribution on the partially inserted slab and the plates. Do you, based on this distribution, expect the slab to be pulled into or expelled from between the plates?

c) Find the amount the potential energy of the capacitor changes when one inserts the slab an additional amount Δx. Does the energy increase or decrease? Is this result surprising given your gut level physical expectation from the picture? (Don't worry, your gut is correct...)

d) To find the "missing energy", determine the amount of work done by the voltage source as one inserts the capacitor an additional amount Δx. Note that this is related to the additional charge that flows onto the capacitor at constant voltage and represents the *decrease* in the potential energy of the *voltage source*.

e) By using

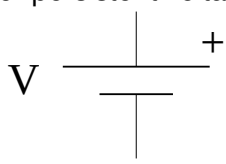
$$F_x = -\frac{dU}{dx}$$

where U is the *total* potential energy of the voltage source and capacitor, find the force on the partially inserted place. Does the force pull the dielectric slab in (to fill the plate volume) or does it push it out from between the plates *after all*?

Week 5: Resistance

- A *battery* is a chemical device that functions as a "persistent capacitor" that can deliver charge at a given voltage for a very long time. In a sense, it is made up of a vast number of tiny molecular-scale capacitors in parallel, each one of which is "neutralized" as charge is transferred. Batteries store and deliver energy as they function as a source of electric *current*.

  The symbol for a battery (or other persistent voltage ***source***) in an electric circuit is:

  Technically, this symbol is for an electrical *cell*, and a battery is a collection of cells in series (with their voltages adding to create a higher voltage than we could otherwise create with the chemical process) but the terms will be used interchangeably in this introductory work.

- Current is defined as:

$$I = \frac{\Delta Q}{\Delta t} = \frac{dQ}{dt} \tag{5.1}$$

  This is the charge per unit time flowing (for example) from one terminal of a battery to the other or from one plate of a capacitor to the other through a conducting pathway.

- Ohm's Law is:

$$\Delta V = IR \tag{5.2}$$

  which can be modelled from:

$$R = \frac{\rho L}{A} = \frac{L}{\sigma A} \tag{5.3}$$

  where L is the length of the material, A is its cross-sectional area, $\rho = 1/\sigma$ is its *resistivity* where σ is its *conductivity*. Since $\Delta V = EL$ (the potential difference across it is the uniform field inside times the length) we can also write Ohm's Law as:

$$\vec{\boldsymbol{J}} = \frac{\Delta Q}{A\Delta t}\hat{\boldsymbol{n}} = \sigma\vec{\boldsymbol{E}} \tag{5.4}$$

  where $\vec{\boldsymbol{J}}$ is the vector *current density*. From this we can see that *electric fields are not zero in a conductor carrying a current!*

- The power dissipated by a resistance carrying a current is:

$$P = VI = \frac{V^2}{R} = I^2 R \tag{5.5}$$

where the first form is the easiest to understand.

- Adding resistors in series:

$$R_{\text{tot}} = R_1 + R_2 + ... \tag{5.6}$$

- Adding resistors in parallel:

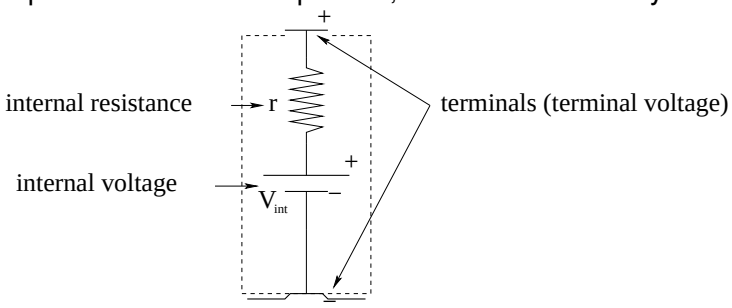
$$\frac{1}{R_{\text{tot}}} = \frac{1}{R_1} + \frac{1}{R_2} + ... \tag{5.7}$$

- Kirchhoff's Rules:

  a) **Loop Rule:** The sum of the voltage changes around a circuit *loop* must be zero (conservation of energy).

  b) **Junction Rule:** The sum of the currents flowing into a circuit *junction* must be zero (conservation of charge).

- The battery described above is an "ideal" battery that can in principle deliver any amount of power. A *real* battery (or other power supply) can never deliver an arbitrarily large electrical power to a circuit. One model that (quite accurately) describes the limiting of power delivered from a battery is that of ***internal resistance***. In this model, a "real world" battery consists of *two* components integrated inside the battery housing – a source of electrical energy (usually chemical energy for traditional batteries) and an effective ***internal resistance*** of the chemical medium and the rate limiting aspects of the chemistry itself. When power limitation is important, batteries will usually be represented as:



(where I have added a small representation of the terminals of a typical commercial battery such as a D or AA cell).

- When a real battery is delivering *no* current, the voltage drop across the internal resistance is zero, and if the chemical "fuel" of the battery is not totally depleted, you will usually measure an internal voltage determined by the chemical potential of the reaction itself. In this case, the terminal voltage will be equal to the internal voltage, which is generally the nominal/rated voltage of the battery or cell. When the cell is delivering current I, however, the terminal voltage (between the physical terminals on the ends of the battery) is:

$$V_{\text{terminal}} = V_{\text{int}} - Ir \tag{5.8}$$

The internal resistance determines the *maximum current and power deliverable by the battery* when the battery is *short circuited* – its terminals connected by a presumed perfect conductor. They are:

$$I_{\max} = \frac{V_{\text{int}}}{r} \tag{5.9}$$

and:

$$P_{\max} = V_{\text{int}} I_{\max} = \frac{V_{\text{int}}^2}{r} \tag{5.10}$$

- RC circuits are simple loops where a capacitor is charged or discharged through a resistance. You should be able to derive the time-dependent discharge of a capacitor through a resistor as the following ***exponential decay***:

$$V_C(t) = V_0 e^{-t/RC} \tag{5.11}$$

or:

$$Q(t) = Q_0 e^{-t/RC} \tag{5.12}$$

where Q_0 is the initial charge on the capacitor and $V_0 = Q_0/C$ is the initial potential across the capacitor. This result follows from applying Kirchhoff's voltage law around a loop and converting it into a first order, linear, ordinary differential equation of motion that can be directly integrated.

- The "exponential time constant" of this decay is $\tau = RC$. Recall that the time constant τ is the fixed time interval in which the initial charge/potential decays to $1/e$ of its value at the start of the interval. Exponential processes always gain/lose the same *fraction* of their initial value in any given interval of time.

- A charging capacitor (initially uncharged) can similarly be shown to exponentially approach an asymptotic charge or potential:

$$V_C(t) = V_0 \left((1 - e^{-t/RC} \right) \tag{5.13}$$

or:

$$Q(t) = Q_0 \left(1 - e^{-t/RC} \right) \tag{5.14}$$

where V_0 is the magnitude of the charging potential and $Q_0 = CV_0$, in both cases the *final* values found on the capacitor after a very *long* time, specifically many exponential time constant intervals.

Note on notation: At one time the voltage produced by e.g. a battery or mechanical power supply was called (by Allesandro Volta, one of the original discoverers of the chemical electrical cell) an *electromotive force*, and this usage was continued by later researchers such as Faraday. This was a horrible misnomer – Volta's model for the cause of the voltage (that "motivated" the choice) was incorrect, and of course the ***units of force, Newtons, are completely different from the units of voltage, Joules per Coulomb***. The SI unit of potential and potential difference, the Volt, is named after Volta.

Unfortunately many physics textbooks perpetuate the tradition of referring to the voltage produced by *any* means as an electromotive force or use the acronym "EMF" to describe this voltage without actually using the word force. In addition, the symbol $\mathcal{E}$ is often used in place of the symbol V to label the voltage of a cell or induced voltage (discussed in a few chapters) as an $\mathcal{E}$-MF. Although this is a calligraphic/script font version of E, it is still remarkably easy to confuse with the electric field and of course a voltage isn't conceptually or dimensionally an electric field, either!

This book will (hopefully consistently) use the symbol V to describe the voltage sources or sinks of a circuit element or the circuit itself, including electrical cells or induced voltages, and will eschew the use of the symbol $\mathcal{E}$ or the descriptors EMF or (worse) "electromotive force" used to describe a potential or potential difference no matter what it results from. This should do no conceptual harm to the general topic of electricity and magnetism; indeed it should simplify the treatment of potential differences. Students should be aware of the more common usage, however, to the extent that they use additional textbooks or references to supplement this one as they study.

5.1: Batteries and Voltage Sources

Up to now, we haven't really considered *how* the capacitors in the sections above got charged up. Our model of matter is electrically neutral atoms and molecules, and while conductors have lots of mobile charge we don't know how to *grab* that charge and push it around yet. Or rather, we do – one way to push it around is to use *the electric field itself* to do the pushing!

This is how one charges things like amber and glass or clouds by rubbing them. The fields of the atoms rub together and knock off charges and transfer them preferentially in one direction or the other. But another way of grabbing things with fields is to exploit the electrostatic field that holds atoms and molecules together in *chemistry* – a *battery*[73].

5.1.1: Chemical Batteries

It is probably instructive to look at the actual chemical reaction associated with at least one *specific* kind of battery, even though one can make a cell out two different kinds of almost *any* metal stuck into an electrolyte solution (e.g. an acid). So let's look at the two reactions associated with a lead-acid battery, the kind you probably have in your car.

A lead-acid battery consists of two plates. The anode (positive pole) is made out of ordinary lead. The cathode (negative pole) is made of lead coated with lead oxide. Both are immersed in a solution of water and sulphuric acid. At the anode[74] :

$$\mathrm{Pb} + \mathrm{HSO_4^-} \rightarrow \mathrm{PbSO_4} + \mathrm{H^+} + 2\mathrm{e^-}$$

while at the cathode:

$$\mathrm{PbO_2} + \mathrm{HSO_4^-} + 3\mathrm{H^+} + 2\mathrm{e^-} \rightarrow \mathrm{PbSO_4} + 2\mathrm{H_2O}$$

or overall:

$$\mathrm{Pb} + \mathrm{PbO_2} + 2\mathrm{H_2SO_4} \rightarrow 2\mathrm{PbSO_4} + 2\mathrm{H_2O}$$

[73]Technically, a single device that generates a voltage in this way is called a *cell* – a *battery* is composed of several cells – but we'll just call anything that generates electricity a battery because nobody speaks of "flashlight cells" when they go to the store to get a pack of D's, they say "I'm going to get some batteries for the flashlight".

[74]Wikipedia: http://www.wikipedia.org/wiki/Lead-acid battery. There are more complete ways of writing out the chemical reaction that show more of what is going on with the water in all of this, but this is sufficient. Either way, you are of course encouraged to visit the link and read more about it.

plus the transfer of two electrons, driven by the chemical energy of the reaction, between the cathode and the anode.

The electrolyte provides both the (ionized) sulphuric acid required at both ends and a conducting pathway for the electrons to be transported from the anode to the cathode. Energy is released by this reaction; the end products are *more* stable than the original ones so the reaction is *favored*.

However, once a few atoms in the anode have given up their electrons and they've been pulled over to the cathode, the reaction stops! The poles are then *charged up* and it costs too much work to remove any more electrons, more than one *gains* in the chemical reaction. The anode is then charged up *positively* (as an electron *donor* to the reaction in the battery itself) while the cathode is charged up *negatively* (having received the electrons). The top and bottom plates behave *just like the plates of a capacitor* and maintain an electrical potential difference of around 2 volts (per *cell* in a *battery* of six cells, in a typical twelve volt battery in a car) between them that just balances the chemical potential of the arrangement.

There is, however, an important difference. If one provides a *conducting pathway* between the anode and the cathode *outside* of the solution, then the negative charge surplus on the cathode can flow *back* over to the anode and participate in another reaction, then another, then another. Charge continues to be driven in this way until all of the lead and lead oxide is converted into lead sulphate and water. For every mole of lead converted into lead sulphate, two moles of electrons have to move from cathode to anode. That is $1.2 \times 10^{24}/1.6 \times 10^{19} = 0.75 \times 10^5$ *Coulombs* of charge, enough to drive an Ampere of current (one Coulomb/second) for around a day. A mole of lead is around 207 grams, which weighs around a half a pound. Allowing for the electrolyte and sulphuric acid, roughly a pound of battery will drive a load of two watts (one ampere at two volts) for just under a day (where we'll work out energy relations below to justify this in a moment).

A second advantage of this particular battery is that it is *rechargable*. If one simply places a voltage across the cell that exceeds its terminal voltage, charge flows *the other way*, reversing the reaction and turning lead sulphate back into lead or lead oxide. By careful design, one can charge and discharge the battery many times before too much lead sulphate falls off of the electrodes or crystalizes out across the space in between the terminals and shorts out the batter, at which time the battery must be remanufactured (to avoid dumping toxic lead into the environment).

Vehicle batteries, of course, weight many pounds – as many as fifty or sixty – and have six cells, and therefore can drive bigger currents at higher voltages, currents that can easily be large enough to be dangerous. In fact, a car battery[75], and can easily kill you if you handle it carelessly by the poles with e.g. wet hands or cuts on your fingers! I've gotten "hit" this way myself handling a car battery by the poles in a rainstorm, and it hurts! This kind of battery can (multiplying out the coulombs, volts, and seconds) do around 150,000 joules of work per pound in the ideal case, probably less than half this in the real world case.

However, all batteries have a *finite rate* at which they can do *work*, determined by the physical limitations on the rate at which the chemical reaction can proceed. So even if one

[75]http://www.darwinawards.com/darwin/darwin1999-50.html Not just a car battery. You can kill yourself with a nine volt transistor radio battery, and one of my favorite Darwin awards went to a Navy officer who demonstrated this the hard way after being warned about the danger.

shorts out a battery with a *perfect* conductor, one won't get an infinite current at a constant voltage. As the current goes up, the voltage goes down, until at some point all of the energy is released as the heat of reaction in the electrolyte and none to the battery load. Some batteries are designed to provide a fixed voltage and low current for a long time; others are designed to produce a fixed voltage and a *large* current for a *short* time. Car batteries in particular are usually pretty good at both.

5.1.2: The Symbol for a Battery

All of this is too complicated for intro physics, of course. We want to start by idealizing a battery and replacing it in all circuits we consider with a single simple symbol. The symbol we will use is the nominal potential difference maintained by the battery between its terminals (its "terminal voltage") and where the $+$ sign (and longer plate) indicate the *anode*, the side of the battery *from* which positive current flows (where we are suffering from Franklin's Mistake, because the actual motion of charge in the chemical reaction above is negative electrons flowing the other way). Again, the battery behaves like an "inexhaustible capacitor" in an electrical circuit, *increasing* the potential by V as one moves from the cathode (small plate) to the anode (large plate) in any circuit diagram containing this symbol.

Our *ideal* battery never runs out of power, has no limitations on the amount of current it can provide at its rated voltage, and its voltage is rigorously constant. None of these is going to be true in practice for real batteries, and after we define resistance and work out Ohm's Law, series resistance addition rules, and Kirchoff's rules below, we'll revisit the battery and see how we can *compensate* for these features by assigning an *internal resistance* r to the battery itself. This internal resistance is not entirely a fiction – batteries and other power supplies *do* have some actual internal resistance – but it often also represents the practical effect of other rate limiting physics, such as the maximum rate that some given force can do work on a piece of generating apparatus.

This internal resistance will quite naturally cap the power and current the battery can provide as one cranks up the load on it. It still doesn't indicate the way voltage and current depend on things like temperature, the degree to which the battery is discharged already, and how old the battery is – *all* of these things and more affect *real* batteries, dynamos (electric generators), solar cells, and any other method we have of turning (potential) energy into electrical power. But we will do quite well with our idealized battery, and even better with our idealized battery with an internal resistance – the rest is a mix of more advanced physics and associated engineering and doesn't change the idea, only the details.

5.1.3: Batteries and Renewable Energy

Before we move on to resistance, it is worth pointing out that battery physics and engineering are *important* in our society, and becoming *more important* as we move in the direction of renewable energy sources, hybrid or flat-out electric cars, rechargable electronic devices galore and more.

One of the biggest obstacles to the widespread adoption of solar or wind generated power is the difficulty of storing power that is generated when the sun is high and bright or when the

wind blows strongly for use at night or on calm days. With fuel-generated energy, as long as one provides the fuel one can produce the energy! This is not possible with sunlight, and parts of the Earth get no sunshine at all for months at a time (as well as sunshine 24 hours a day other months at a time). Similarly, even "windy" locations can have calm weather for days or even weeks at a time.

It requires hundreds of pounds of lead-acid batteries *per person* just to store the average power needed for a single day (say) generated from solar energy or wind energy collected in intervals during that same day. Lithium batteries that store the same amount of energy are much smaller and lighter, but lithium is an alkali metal and burns when exposed to air, making it more difficult to safely engineer high-capacity batteries. Alternative battery technologies (say, zinc-oxide batteries, lithium batteries, and more with very different chemistry, both wet and dry) are constantly being explored, driven by the need to store at least a few days' worth of power from intermittant sources to bridge those times when the source is not available, as well as to make it possible for our laptops, tablets, and phones to run for days on a single charge and for electrical cars to travel long distances on a charge and recharge quickly.

The inventor(s) of a really, really compact and efficient way of storing energy would both make a well-deserved fortune from the idea and would enable any number of beneficial changes to our energy hungry society. In the meantime, rechargable batteries have and are likely to continue to have many problems: They are (so far) bulky and massive, they get hot while operating at high power levels (due to their internal resistance!), they are often made with toxic or comparatively scarce materials, they are consequently difficult to safely dispose of, they (so far) wear out and can store much less energy after a few hundred or at most a few thousand charges, they can explode or catch on fire if overdriven (making them very nearly a munition in the hands of the unscrupulous or violent). Put all of this together and so far, batteries are *very expensive*, both in direct dollar cost per unit of energy stored and in terms of environmental cost and risk! Yet there is little doubt that within the decade, batteries will be running many if not most of our homes and cars in addition to all of the things that they are used for now.

If this topic interests you you can learn a great deal about rechargable/secondary battery technology (which is very much a moving target, where the costs per unit of energy stored by a rechargable battery have decreased by some 60 to 80% over the last ten or fifteen years) by visiting:

Wikipedia: http://www.wikipedia.org/wiki/Rechargable Battery

To summarize: at this point (with this paragraph being written in 2017) large capacity, high density, long lifetime rechargable batteries for capable of running a "typical" U.S. household (that uses, say, around 30 kilowatt-hours of energy a day) costs less than $5000 at full retail to an individual consumer. I personally expect that by 2020 battery technology will advance so that the full retail cost crosses the $0.10/watt-hour threshold (where now it is more like $0.15), so that a full-day battery will cost roughly $3000 (with a two day supply or supply for a larger household still quite affordable). There is no good reason to think that retail costs will not continue to fall beyond this point as technology improves and manufacturing capacity increases and enables various economies of scale. Well within the decade, individual houses will be easily and cheaply equippable with "backup batteries" that can store days' worth of energy for the entire house that will last for a decade or more with most of their charge storage capability intact.

Similarly, as of this writing an array of rooftop solar cells capable of recharging these batteries with the energy received in a single "typical" sunny day in most of the United States costs around $5000, and this number will *also* continue to decrease to 2020 and beyond as new technologies emerge and maufacturing capacity increases.

Electrical energy purchased from a utility company in the United States currently costs an *average* of 12 cents per kilowatt-hour, so a year's worth of electrical energy for a "typical household" is around $1300 in 2017. If one invests approximately $10,000 (plus $3000 for installation), one can *already* go "off-grid" in most U.S. locations (ones with adequate insolation) and generate very close to 100% of the electrical energy needed to run a typical household, and *break even* on the investment in roughly a decade, for about what a top of the line high efficiency air-conditioner/heat pump for that same household would cost.

The amortization time required to recover the investment will very likely drop to seven years or even less within the next few years, making this a no-brain decision for most households – one can borrow the money required to convert over and pay off the loan in a matter of years for about what one would pay for a new car and entirely funded by reduced electrical utility bills and enjoy "free" electrical power for the rest of the useful lifetime of the hardware, estimated at this time to be in excess of twenty years.

5.2: Resistance and Ohm's Law

Fine, so now we have a battery. We place a chunk of conducting matter between the poles/terminals of the battery, and what happens? Well, *current flows*, that's what happens! We have created a situation where a conductor is *not* in electrostatic equilibrium, and *charge moves in time* through the conductor in response to the force created by the battery, with *energy released* in the process. This is actuall fine, and we might even say, it's about *time* that we got out of statics (which are kind of boring, as not much happens, right?) and into *dynamics*, where things happen. All we need, then, is to come up with a model for what goes on inside the conductor as the current flows, and we can start to analyze dynamical electrical systems once again, which has to be more interesting than just thinking about a charged capacitor sitting around all do doing nothing much but just storing charge.

A microscopic picture, of course, begins with atoms, each with a heavy nucleus and surrounded by electrons, arranged in some sort of solid lattice, with some of the electrons "free" to move within the lattice. Free to move, however, is not the same thing as non-interacting. Electrons that move through the lattice interact with the lattice and transfer their momentum to the lattice so that (in equilibrium) their average velocity is zero. The lattice therefore exerts a kind of *drag force* on the electrons that brings them back to equilibrium.

The *simplest* model for conduction of electrons through a material that "resists" their motion via a drag force caused by the collision of the moving electrons with each other and the underlying atoms in the lattice is one with a *linear drag force* – one that is proportional to the average velocity of transport of the electrons through the resistive lattice. If the electrons are being pushed through the conductor by some constant force, then, they'll arrive quickly at a *terminal velocity* that is proportional to that force, where the forces balance.

5.2.1: A Simple Linear Conduction Model

We now use this model for "linear drag" to build a working description of voltage changes, electric fields, and electrical currents and current densities inside conductors that are ***not*** in electro***static*** equilibrium. They are not really static because charge is being pushed by electric fields and is moving, but they are still in a kind of *dynamic* equilibrium where forces on the charges balance. This model will also work for *"slowly varying" currents* – currents we can treat as being *approximately* constant on small time intervals Δt – but ultimately it will ***fail*** when we take into account the possibility of the conductor *radiating energy and momentum* into space for rapidly varying currents (a consequence of the Maxwell Equations we haven't learned yet in *electrodynamics*). It is thus a "quasi-static" theory and should not be taken too seriously or considered to be completely general or correct.

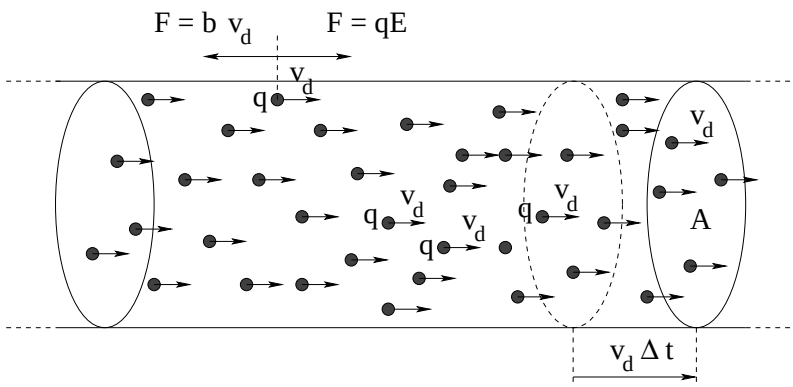


Figure 5.1: The simple linear model for conduction in a resistive lattice.

In figure (5.1) we see a model for a conducting wire. This wire has a cross-sectional area of A and contains an "inexhaustible" supply of free charged particles (recall, order of one free charge per atom) each with charge q. An electric field is created within the wire by a *battery* (not shown) that exerts a force on any given charge carrier to the right of $F = qE$. The wire resists the flow of that charge carrier with a "drag force" bv to the left, where b is a phenomenological "drag coefficient" characteristic of the imperfect conductor. Microscopically, we can initially mentally picture this drag force as being the result of an ongoing average loss of momentum as each free charged particle speeds up in the direction of the electric field for a time but then is suddenly slowed down enough to "start again" as it collides with the atoms or molecules of the material (incidentally heating the material).

In "dynamic equilibrium" (steady, or nearly steady currents) we require these two forces to balance:

$$v_d = \frac{qE}{b} \tag{5.15}$$

where we introduce the ***drift velocity*** v_d, defined to be the average "terminal velocity" of charges in the conductor[76]. It is important to keep in mind that in a typical normal metal our charge carriers are negatively charged *electrons* (recall "Franklin's mistake") and all of the

[76]We will give a particular, simple, classical model called the ***Drude model*** for the drift velocity that will give us an actual functional form for b in a more advanced section below that can safely be omitted by students uninterested in majoring in physics or more advanced studies in e.g. engineering (although it is not terribly difficult and is a worthwhile exercise in mechanics).

vectors are reversed for a current and field that still go from left to right, but this makes no difference in anything we care about (yet!); the argument given below works for either sign of the charge carrier.

Let's carefully examine the picture and see what we can deduce. We are interested in computing the ***electric current***, defined to be the ***charge per unit time*** that is being carried by the conductor. Ordinarily, we'd think of this as the charge per unit time travelling in some chosen direction that passes some point in the conducting wire under the influence of the force created by the battery (or other source of potential difference across the wire). However, what does "passing a point" mean? How can we manage our choice of direction? All of the charge may not be travelling in the same direction! The conductor may not be a simple cylinder like that pictured above but instead be some contorted shape cast in metal with many branches! We need a better, less ambiguous definition.

We *can* unambiguously estimate how much charge passes through a given *surface* cutting across the metal. Since a surface in three dimensions has one dimension perpendicular to the surface, we can always assign our direction unambiguously in this perpendicular direction. The wise student should already be saying to themselves "But that sounds a lot like our reasoning when we talked about the flux of the electric field a few chapters back.." and that wise student would be quite right!

But first things first. For the time being, let's confine ourselves to the simple case of the cylinder above with the surface in question being one that cuts across it *at right angles*, the surface A pictured above. From the picture we can see that *all of the charge* ΔQ in the volume between the plane surface bounded by the dashed circle and the plane surface A bounded by the circle at the far right of the conductor passes through the cross-sectional area A perpendicular to the direction of motion of the charges in a time Δt. So how much is that?

To answer this, we need to define a few quantities. One is:

$$n = \frac{\#\text{ of charge carriers}}{\text{unit volume}} \tag{5.16}$$

the number of (free!) charge carriers q per unit volume. We can then turn this into the *free charge density*:

$$\rho_{\text{free}} = nq \tag{5.17}$$

Using these quantities, we see that:

$$\Delta Q = nqv_d A\Delta t = \rho_{\text{free}} v_d A\Delta t \tag{5.18}$$

which we read as "the number of charge carriers per unit volume times the charge per carrier times the volume $v_d A\Delta t$". This means (dividing out the Δt) that the total charge per unit time that goes through A is:

$$I = \frac{\Delta Q}{\Delta t} \approx \frac{dQ}{dt} = nqv_d A = \rho_{\text{free}} v_d A \tag{5.19}$$

In passing we note that the SI units of current are *Amperes* (or Amps for short) where

$$1\text{ Ampere} = \frac{1\text{ Coulomb}}{1\text{ Second}} \tag{5.20}$$

The result $I = nqv_d A$ will occur again and again when we pass from a microscopic description of e.g. magnetic forces on charges to macroscopic forces on current carrying wires,

so keep it in mind! It isn't just a transient "use once" result; it is the key to understanding many things.

5.2.2: Current Density and Charge Conservation

Note well that in the picture above, we determine the current that passes "a point in the wire" by evaluating how much charge passes through some *surface* that contains the point! The particular surface we chose in our simple derivation is one perpendicular to the direction of the motion of the charge, but we cannot possibly guarantee that all conductors carrying a current will have some simple known surface where this is true. Also, as we noted above, this picture should remind you of something – it is very similar to the pictures we used to talk about ***electric flux*** in the context of Gauss's Law!

The problem we face is that there are *many* surfaces that pass through any given point, so talking about how much charge passes a point on the wire isn't very well defined. We would do better talking about how much charge passes through a closed curve drawn around the wire (or other, arbitrarily shaped) conductor, but even so, there are an infinite number of surfaces bounded by any closed curve. We need the electric current through such a loop not to depend on the surface chosen, at least in the (quasi) stead-state dynamical equilibrium currents we are talking about here.

We can achieve this by recapitulating the reasoning of electric flux for (again) a single, simple cylindrical wire where we can count on help from geometry. We want the current through our surface A perpendicular to the direction of motion of the charge to be the same as the current through a second surface A' that is cut through the wire at more or less the same place but is tipped at an angle θ relative to the direction of the current. As before, the tipped surface area area $A' = A/\cos(\theta)$ is larger than A. In order to get the *same* current I from these two surfaces, we need to compensate for the cosine on the bottom with one on the top:

$$I = nqAv_d = nq\frac{A}{\cos(\theta)}\cos(\theta)v_d = nqA'v_d\cos(\theta) \tag{5.21}$$

We can get the cosine out of a dot product between the *local direction* of the *vector* drift velocity $\vec{v}_d$ (assumed to be parallel to the actual current at any point in the wire) and $\hat{\boldsymbol{n}}$, the *directed normal unit vector* to the surface A or A':

$$I = nqA\vec{\boldsymbol{v}}_d \cdot \hat{\boldsymbol{n}} = nqA'\vec{\boldsymbol{v}}_d \cdot \hat{\boldsymbol{n}}' \tag{5.22}$$

We have a single choice to make in this expression – there are two possible directions perpendicular to the surface and we have to choose (for example) either left to right or right to left as being positive $\hat{\boldsymbol{n}}$.

Again as before in our discussions of electric flux, we can take an arbitrary *curved* surface and break it up into tiny differential chunks dA, each with its own normal vector $\hat{\boldsymbol{n}}$ selected with the same left-to-right or vice-versa sense. The chunks are small enough that we can treat all the charges that pass through them as *locally* all going in the same, unambiguous direction $\vec{\boldsymbol{v}}_d$. For each of these, the differential current through the chunk is:

$$dI = nq\vec{\boldsymbol{v}}_d \cdot \hat{\boldsymbol{n}}dA \tag{5.23}$$

and we can now unambiguously sum up all of the current through an arbitrary *curved* surface or through plane surfaces where the flow of charge is *not* all parallel and perpendicular to the surface.

If we chose as our surface *any* open surface S that cuts completely across a branch of our conductor, we will find that it is always bounded by a closed curve C on the surface of the branch. We can then write the following, completely general and correct definition for the "current in the branch" in the steady state:

$$I_C = \int_{S/C} nq\vec{\boldsymbol{v}}_d \cdot \hat{\boldsymbol{n}} dA = \int_{S/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA \tag{5.24}$$

where S/C is read "through the surface S bounded by the closed curve C and where:

$$\vec{\boldsymbol{J}} = nq\vec{\boldsymbol{v}}_d = \rho_{\text{free}}\vec{\boldsymbol{v}}_d \tag{5.25}$$

is called the ***current density***. In other words, ***the current through an open surface*** S ***bounded by a closed curve*** C ***is the flux of the current density through that surface.*** Note well that this is still just $I = nqv_dA = \rho_{\text{free}}v_dA = JA$ for the simple cylindrical wire and perpendicular surface A we began with, but it can now handle far more general flows of current.

We are now in a position to be able to derive a beautiful form for the ***Law of Charge Conservation***. Consider a simple closed surface S (like the ones we considered for Gauss's Law) located anywhere in space. We already know that the closed surface S encloses some volume V/S, and we already know how to compute the total charge inside:

$$Q_{\text{in S}} = \int_{V/S} \rho dV \tag{5.26}$$

or, the total charge inside S is the integral of the charge density inside.

If charge can never be created nor destroyed, the only way the total charge in V can change is if *charge moves across the surface* S*!* Charge can flow *in* to the volume through S or *out* of the volume through S, or both at the same time, but if S is impervious to charge (say it is a "perfect insulator"), the charge inside can never change.

Quantitatively, then, the total current through S has to equal the rate of change of the total charge inside. All we have to do is assign a choice for the direction of $\hat{\boldsymbol{n}}$ – into or out of the volume – and write this in differential/integral form. Let's choose "out" because it then is consistent with Gauss's Law (which will prove strangely useful to us later on!):

$$I_{\text{out}} = \oint_S \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA = -\frac{d}{dt}\int_{V/S} \rho dV = -\frac{dQ_{\text{in S}}}{dt} \tag{5.27}$$

which we rearrange as:

$$\oint_S \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA + \frac{d}{dt}\int_{V/S} \rho_e dV = 0 \tag{5.28}$$

This equation is *very important!* It is, in fact, a *law of nature*, based on substantial empirical evidence. It is the *law of charge conservation* written in mathematical form. Basically, it says that the amount of charge inside any volume bounded by a closed surface can only decrease

(increase) if charge flows *out (or in) through the surface!* The net charge inside cannot just poof into or out of existence, it has to get there by coming in from outside[77].

5.2.3: Advanced: Differential Form and Maxwell's Equations

If/when you take a more advanced course in electromagnetism, one of the very first things you will do is apply the divergence theorem to the Law of Charge Conservation, Gauss's Law, and expressions containing flux integrals in general and convert them to vector differential form. Treating the divergence theorem and doing this algebra is beyond the scope of this course (although advanced students may have done it in the starred homework problem in the Gauss's Law chapter earlier and can get the same result with the same procedure here) but we put down the result (only) here for completeness and to make it easier to make the connection in a future course.

The law of charge conservation in differential form is:

$$\vec{\boldsymbol{\nabla}} \cdot \vec{\boldsymbol{J}} + \frac{\partial \rho_e}{\partial t} = 0. \tag{5.29}$$

This ends up being much more convenient for doing the math associated with solving serious electrodynamics problems. It is also has a critical invariance property when one learns about the *four-dimensional geometry* associated with the theory of special relativity – basically charge is conserved in all inertial reference frames even when relativity is taken into account.

We can also look ahead a bit at this point. Soon we will discover that Maxwell's equations are called Maxwell's equations because Maxwell more or less discovered an *inconsistency* in the treatment of current in the original form of one of the laws that could only be made consistent by adding a term to it to *account* for the implications of charge conservation and the arbitrariness of the infinity of surfaces "through" which charge can flow that are all bounded by a single closed curve C.

To help the interested or advanced student out, consider figure 5.2. In this figure, we split the closed surface S bounding V into two pieces, $S = S_1 + S_2$ by drawing single closed curve C all the way around it. S_1 on the left and S_2 on the right are ballooned out so that they resemble two "fishing nets" placed face to face, through which charge can flow.

If current is flowing in a "steady state" way and ***charge is conserved***, the current from left to right through the two surfaces S_1 and S_2 must be equal – the current through the first must equal the current through the second because ***in the steady state, no charge is building up in between***.

However, if we put e.g. a capacitor plate in between the two surfaces (or charge is accumulating in some other way), current may not be flowing in a steady state way – current may be *building up* inside the *closed* surface S. In that case the ***difference*** between the current through S_1 and the current throught S_2 is the rate at which charge builds up inside V:

$$\int_{S_1} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA - \int_{S_2} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA = I_{\text{into V thru } S_1} - I_{\text{out of V thru } S_2} = \frac{d}{dt} \int_{V/S} \rho_e dV = \frac{dQ_{S/V}}{dt} \tag{5.30}$$

[77] There is another way charges can appear inside the box that doesn't violate this law – they *can* be created or destroyed a *pair at a time* in such a way that the *net* charge of the pairs remains zero. This actually happens in high energy quantum mechanical collisions – making it beyond the scope of this course – but the creation of a positron-electron pair does not violate *net* charge conservation.

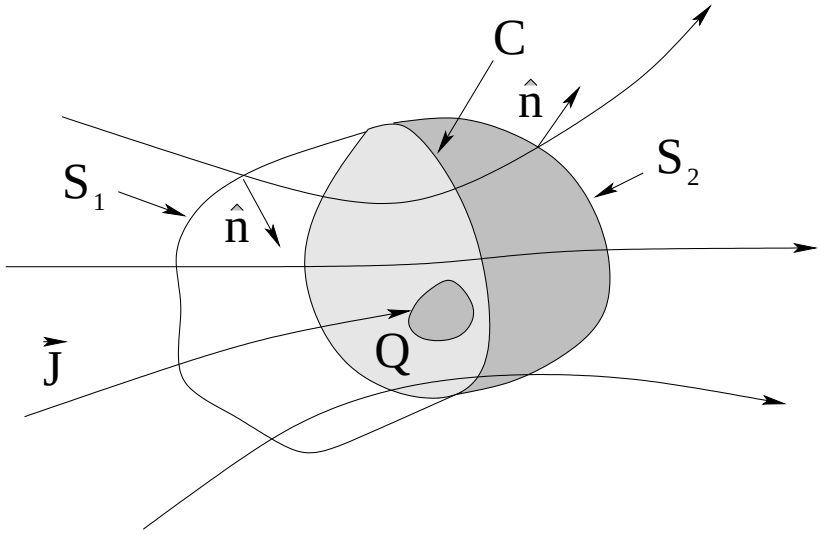


Figure 5.2: Charge Conservation.

(for Q in S/V). Note that all we really did to get this result is split the integral over the closed surface S in our previous discussion into two pieces, and change the sign/direction of $\hat{n}$ for the first surface so that they both go "left to right" instead of "out". You should verify that this makes sense on your own.

Armed with this result, students are encouraged to "play Maxwell" as they go along, and see if they can discover and fix this inconsistency *all by themselves* without looking ahead to see how it is done when Ampere's Law is introduced. You now have all the information you need to do so except for, of course, the actual equation that needs to be repaired which is covered in a later chapter. When you cover it, your instructor may refer you back to this section and suggest again that you give it a try.

5.2.4: The Drude Model

The earliest forms of the *passive pinball game* [78] worked by dropping a small metal sphere (usually a ball bearing of some sort) into a vertical box fronted by a piece of glass and studded on the inside with "pins" as pictured in figure 5.3. The ball would bounce down through the pins in not-quite-random ways and end up in one of several slots at the bottom. One could then gamble on just which slota ball would end up in, or try to use skill in the way the ball was dropped to determine the outcome. However, because the (essentially classical) motion is effectively chaotic, it is nearly impossible to drop a ball into the array of such a way that the final outcome could be controlled or predicted in anything but a statistical way after the first two or three collisions with pins.

Note well that the pinballs cannot escape through the sides, and to avoid complications such as a ball striking a side and falling *straight down to the bottom* along a side, we will assume that the sides are perfectly elastic bumpers that effectly *reflect* a ball back into the lattice of pins in the horizontal direction without affecting its vertical motion.

[78] Wikipedia: http://www.wikipedia.org/wiki/pinball. By "passive" we mean that the bumpers are not electrical and don't add energy to the bouncing ball.

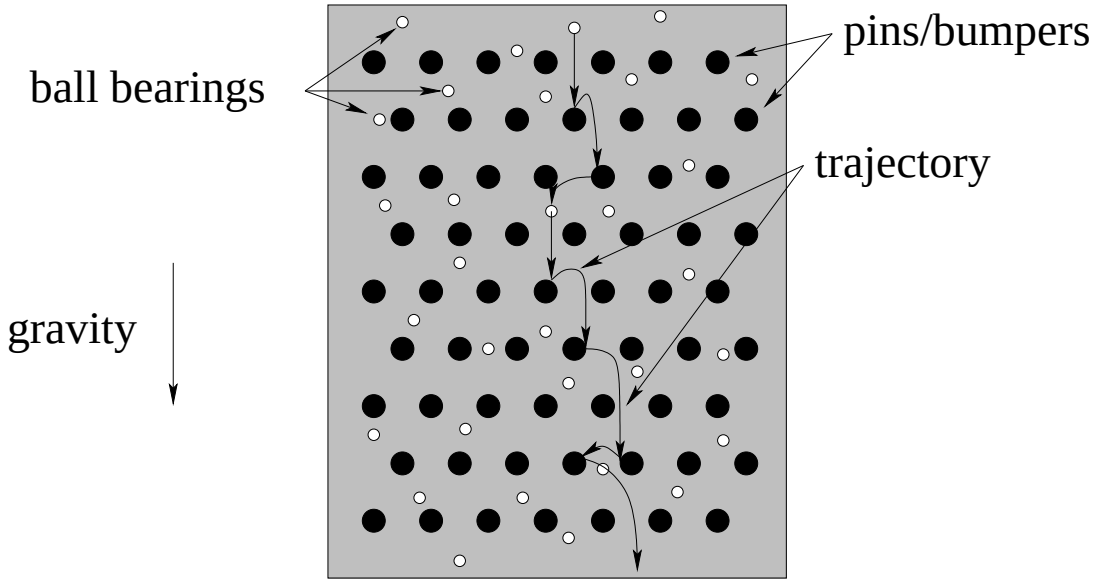


Figure 5.3: An early pinball machine. Balls (typically small ball bearings) dropped in at the top fall into an array of "pins" that function as bumpers but are vertically"stopped" by the pins after falling a short time τ so they only build up a finite downward average speed.

Physicists, mathematicians and statisticians got involved in the game at a very early point – for example the ***Bean Machine*** [79] was built specifically to demonstrate the *central limit theorem*, an important result in the theory of probability and statistics. This sort of machine is equally useful in the context of understanding classical resistance. Let us build a very simple "pinball" model for conduction where the electric field that pushes charge through a lattice of atoms is replaced by gravity pulling down ball bearings and where the atoms in a lattice are replaced by the pins. One can still sometimes find simple pinball machines of this sort (sometimes called Pachinko machines) sold as toys.

Let's use this pinball model to make a simple conduction model, replacing the balls with free charges and the pins with the lattice of atoms through which the charges move. There is just one catch – in the ***passive pinball model*** illustrated above, the balls fall between pins only due to the force of gravity and the pins themselves effectively *stop their downward motion* on each collision so they have to build up speed again, until the *next* collision stops it – again.

In an *actual* lattice of atoms, the atoms are at a *finite temperature* and the extremely light electrons are in *thermal equilibrium* (more or less) with the lattice. In a nutshell, as we explore in detail below, this means that the average thermal kinetic energy of the conduction electrons is much, much larger than the energy they might gain from the field between collisions in the passive pinball model!

To put it another way, suppose it takes an electron a time τ_E to "fall" (say) some average distance between atoms/collisions and a time $\tau_{\rm therm}$ for the atom to travel that same distance due to their average speed due to their temperature. If their average thermal speed is much larger than the average speed built up between passive collisions, then:

$$\tau_E \gg \tau_{\rm therm}. \tag{5.31}$$

[79] Wikipedia: http://www.wikipedia.org/wiki/Bean Machine. Be sure to play with the dynamic graphics!

and a conduction charge will undergo *many, many* thermal collisions in the time it would have taken to experience *one* passive collision with the lattice!

These thermal collisions are not biased in any particular direction – they are equally likely to make the charge go up, down, right, left, forward, backward in the lattice, so they do not *directly* contribute to the flow of current! However, during the shorter time $\tau_{\rm therm}$, the component of the velocity of an electron in the direction of the force due to the electric field is *slightly* increased so that – on average – the electron "drifts" in the direction of this force as it otherwise bounces around randomly and rapidly in all directions.

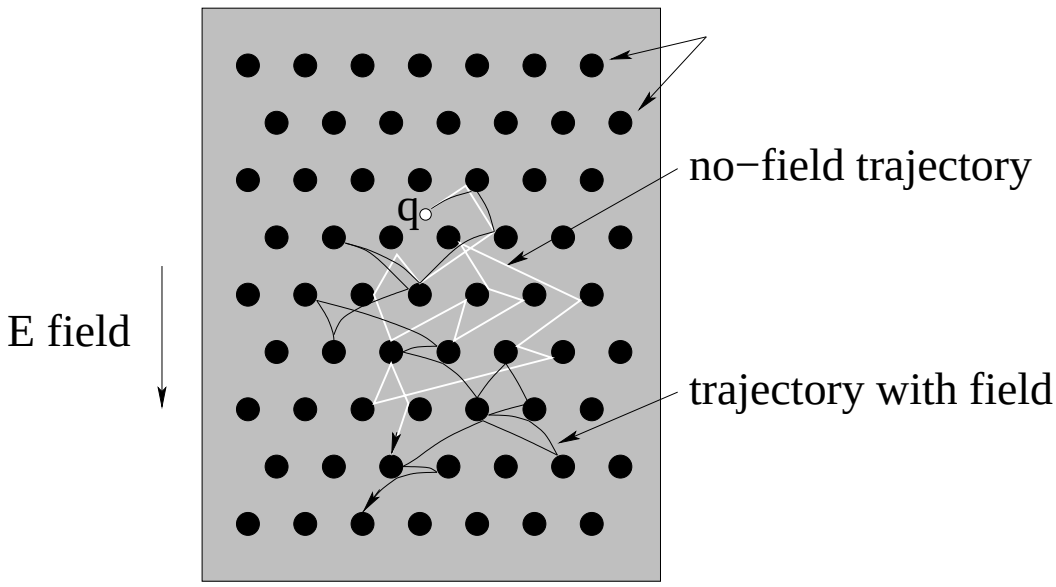


Figure 5.4: The Drude model: With no field, the charges q bounce very rapidly between atomic "bumpers" that maintain a roughly thermal distribution of charge speeds. On average, these collisions form a "random walk" with no direction and zero average displacement. With a field, in the very short time τ between collisions these random free trajectories are *very slightly curved in the direction of the field* and the random walk is now *biased*, with a net displacement that slowly accrues in the direction of the field.

This "rapid thermal collision with weak electrical force biasing electron drift" pictured in figure 5.4, is called the ***Drude model***. With no voltage/field, the lattice of atoms and charges behaves like a horizontal "active" pinball table filled with pinballs (charges) and bumpers (atoms) that are firing/vibrating at an extremely rapid rate so that charges constantly bounce between the atoms in an unbiased random walk, leading to ***zero average displacement***. The charges themselves are (recall) strongly repulsive and there is an average of roughly one charge per atom.

The application of a *voltage* across the conductor is equivalent to *tipping the active pinball table* up through a small angle relative to gravity. In such a tipped, active table gravity (acting only during the brief time between collisions) will *still* make the balls gradually drift lower in the direction of the component of the gravitational field parallel to the table by slightly *biasing* the direction of their otherwise random motion.

In a conductor as well there is a small net acceleration in the direction determined by the electrostatic force on the charges during this *short* time between bounces. Quantitatively (as

we shall see below), we expect each charge to "bounce"' roughly thousands of times between the atoms in the time that it would take a it to cover the distance from one collision to the next due only to the applied field in the *passive* pinball model, but this small, asymmetric force is enough to ***bias the random motion of the charges so that they slowly drift in the direction of net force***.

This is why v_d is called the "drift velocity" – it is a velocity that is quite distinct from the *actual speed* of the particles as they bounce around vigorously between the bumpers! To make a proper conduction model out of this, we also need to (metaphorically) constantly take "pinballs" (charges) off of the lower end of the table and elevate them back up to the top of the table with a conveyor belt of some sort or another so that charge doesn't accumulate at the bottom and generate a backwards-directed field of its own that stops the process. This constant lifting of charges back to "start over again" is an excellent mental model of a *battery*! We finally have a microscopic picture of sorts of a *simple electrical circuit* consisting of a battery and a resistance!

Non-physics majors can probably skip the details in most of the next subtopic (depending on the wishes of your instructor) although it will still be helpful even for non-majors to skim through the the section to get a feel for what is going on and how this depends on that, etc. Physics majors almost certainly *should* work through it in detail even if your instructor *doesn't* plan on testing you on any of those details – you will encounter the Drude model *again* in future courses where you *are* likely to be held responsible for doing the math on a quiz or exam, and it will greatly simplify matters for you then if you make some effort to understand the model in some detail now without the pressure of testing.

Note that *all* students are responsible for the key result – getting Ohm's Law (in both forms) out of the general argument, most of which is intended to justify the linear relationship between v_d and E that is key to the relationship.

5.2.5: Advanced: Details of the Drude Model

Let's now generate the full algebraic description of the Drude model, using the insight that conduction in a resistor can be quantitatively modelled "like the motion of pinballs in a very slightly tipped pinball table with a lattice of highly active bumpers".

We start by *estimating* the mean speed of the conduction charges from thermodynamics. If the lattice is at temperature T, and the free charges are in thermal equilibrium with the lattice (a reasonable assumption), then from the ***equipartition theorem*** we expect the average kinetic energy of the free charges in a three dimensional space when the "table" is not tipped by an applied electric field to be:

$$K = \frac{3}{2}k_bT = \frac{1}{2}m(\langle v_x\rangle^2 + \langle v_y\rangle^2 + \langle v_z\rangle^2) = \frac{1}{2}m\,\langle v\rangle^2_{\rm thermal} \tag{5.32}$$

We solve algebraically for $\langle v\rangle_{\rm thermal}$ to get:

$$\langle v\rangle_{\rm thermal} = \sqrt{\frac{3k_bT}{m}} \tag{5.33}$$

We can now evaluate/estimate this easily enough at (say) 300 degrees kelvin for electrons

from $k_b = 1.38 \times 10^{-23}$ J/K and $m_e = 9.1 \times 10^{-31}$ kg as:

$$\langle v \rangle_{\text{thermal}} = 1.17 \times 10^5 \approx 10^5 \text{ m/sec.} \sim \sqrt{T} \tag{5.34}$$

to get the order of magnitude for the thermal speed expected in metallic conductors "around" room temperature. This *scales like the square root of the absolute temperature!* Even the model is inexact in some detail or another, we can (eventually) test the predictions of the model against this scaling with temperature!

Note that this is *very, very fast* (an order of magnitude greater than escape speed from the Earth!), and at that turns out to be ***not fast enough*** in purely classical quantitative models because the electrons are *not* a classical gas of non-interacting particles, they are quantum mechanical strongly interacting fermions whose effective "speed" is determined by something called the ***fermi energy*** that is only weakly dependent on the termperature: A quantum mechanical treatment leads to average speeds roughly an order of magnitude larger:

$$\langle v \rangle = \langle v \rangle_{\text{QM}} \sim 10^6 \text{ m/sec} \tag{5.35}$$

Note that this is around a *hundred* times escape speed, an appreciable fraction of the speed of light!

Next, we define the ***mean free path*** d as the average distance a free charge travels in some unbiased random direction between atomic "bumpers" at this average speed. We'll assume (for our purpose of numerical estimation) that $d = \langle v \rangle \tau_{\text{therm}} \approx 10^{-10}$ meters or one angstrom, the typical order of the distance between atoms in a metal. Then we can estimate the average time between "bumper events" when the charges interact violently with the atoms in the lattice as:

$$\tau_{\text{therm}} = \frac{d}{\langle v \rangle} \sim 10^{-16} \text{ seconds} \tag{5.36}$$

Before we go any further, we need to compare this time to the time τ_E it ought to take for a charge to start at rest and to move a distance d due to the electric field only in the *passive* pinball model, where the only thing giving the ball speed is the electric field itself and the collisions completely cancel the vertical speed, on average, at each collision. To get a *quantitative* estimate, we have to pick an electrical field strength. Let's assume a (fairly strong) field, one we might expect to find in the filament of an incandescent light bulb with a 100 volt potential difference across a 1 cm filament:

$$E = \frac{\Delta V}{\ell} = \frac{100 \text{ volts}}{0.01 \text{ meters}} \approx 10^4 \text{ volts/meter.} \tag{5.37}$$

This *large* number is fairly representative of the field strength in significant resistive loads, and is orders of magnitude larger than the field strength one would expect in a halfway decent conductor such as household copper wiring.

To estimate τ_E, we use ordinary kinematics:

$$d \approx \frac{1}{2} a \tau_E^2 \qquad \Rightarrow \qquad \tau_E = \sqrt{\frac{2d}{a}}. \tag{5.38}$$

The acceleration follows from $F = qE = ma$:

$$a = qE/m = 1.6 \times 10^{-19} * 10^4 / 9.1 \times 10^{-31} = 1.76 \times 10^{15} \text{ m/sec}^2 \tag{5.39}$$

This lets us estimate:

$$\tau_E = \sqrt{\frac{2md}{qE}} \sim 10^{-13} \gg 10^{-16} \sim \frac{d}{\langle v \rangle_{QM}} = \tau_{\rm therm} \quad \text{(seconds)} \tag{5.40}$$

which is orders of magnitude greater than $\tau_{\rm therm}$ as expected even for quite strong fields. It also lets us estimate what *would* have been v_d in a passive pinball model, using the fact that the average velocity of a particle starting at rest with a constant acceleration is half the acceleration times the time (because the velocity is linear in the time):

$$v_d = \langle v \rangle_{pp} = \frac{d}{\tau_E} = \frac{1}{2} a \tau_E \sim 10^3 \text{ m/sec} \tag{5.41}$$

During the short time $\tau_{\rm therm}$ between thermal collisions, we expect the force exerted by any reasonable electric field inside the material to be "small" compared to the force exerted by the thermally vibrating atomic "bumpers" so that *biased* accumulation of momentum in the direction of the field during the time τ is approximately differential. Also, the strong interaction between lattice and charge maintains near thermal equilibrium with the much more massive lattice; any kinetic energy gained from the field during the time $\tau_{\rm therm}$ is (on average) *lost again* (transferred to the lattice) in each lattice collision so that any given charge doesn't systematically accumulate kinetic energy as it moves in the direction of the field but rather *heats the entire material* while remaining in thermal equilibrium with it. This heating is called *Joule* heating and we will discuss it further later.

We are finally ready to build the Drude model. From Newton's second law it is easy to find the accelertion of a charge q during the time between collisions:

$$\vec{\boldsymbol{F}} = q\vec{\boldsymbol{E}} = m\vec{\boldsymbol{a}} \qquad \Rightarrow \qquad \vec{\boldsymbol{a}} = \frac{q\vec{\boldsymbol{E}}}{m} \tag{5.42}$$

This acceleration only applies for the average time $\tau_{\rm therm}$ before the charge is redirected in a random direction with the original unchanged thermal distribution of speeds. During this time its average velocity is:

$$\langle \vec{\boldsymbol{v}} \rangle = \frac{1}{2}\vec{\boldsymbol{a}}\tau_{\rm therm} = \frac{1}{2}\frac{q\vec{\boldsymbol{E}}}{m}\tau_{\rm therm} = \vec{\boldsymbol{v}}_d \tag{5.43}$$

where the term "drift velocity" is now formally justified as it is the differential *bias* of a much more rapid and violent random process of the charged particles bouncing around between the atoms.

You will note that this is *exactly the same as the expression we obtained in the pinball model* except that instead of using τ_E from $d = \frac{1}{2}\frac{qE}{m}\tau_E^2$ (where this force is *all* that makes the charge move the distance d starting each time from rest), we instead use the average time $\tau_{\rm therm}$ deterined by the strong interaction of the charged particles with the surrounding material while remaining in thermal equilibrium with it! This seems like a small change, but it is a very important one, as the drift velocity one estimates is a thousand times smaller in the second (more correct) case!

In this "active pinball" Drude model, then, we expect from equation 5.43 that:

$$\boxed{\vec{\boldsymbol{J}} = nq\vec{\boldsymbol{v}}_d = \frac{nq^2\tau_{\rm therm}}{m}\vec{\boldsymbol{E}} = \sigma_c\vec{\boldsymbol{E}}} \tag{5.44}$$

which scales ***linearly with the electric field strength***. In this expression, we know or can estimate all of the parameters, and can easily combine the estimates into σ_c, a quantity we will call the ***conductivity*** of the material. Note that $\tau_{\rm therm}$ is *independent of $\vec{E}$ in this model!* This is a crucial point, as we shall see next.

Suppose we use the time $\tau_E = \sqrt{\frac{2md}{qE}}$ from the naive *passive* pinball model above in exactly the same argument. We then obtain an average speed (not really a "drift velocity" any more) of:

$$\langle v \rangle_{pp} = \frac{1}{2} a \tau_E = \frac{1}{2}\frac{qE}{m}\tau_E = \frac{1}{2}\frac{qE}{m}\sqrt{\frac{2md}{qE}} = \sqrt{\frac{dqE}{2m}} \tag{5.45}$$

or

$$J = nq\langle v \rangle_{pp} = nqA\sqrt{\frac{dqE}{2m}} \tag{5.46}$$

(all vectors in the direction of the applied field). The current density would then scale with the ***square root of the field***, which does not empirically agree with Ohm's Law! This is a ***critical failure*** of the passive pinball model, one that cannot be explained away by mere "slop" in our estimation process!

If we compare the Drude model result equation 5.44 to the equilibrium condition $qE = bv_d$ from our elementary discussion assuming linear drag, we see that the "linear drag coefficient" be introduced phenomenologically is given by:

$$\boxed{b = \frac{m}{\tau_{\rm therm}}} \tag{5.47}$$

or:

$$\vec{F}_d = \frac{m\vec{v}_d}{\tau_{\rm therm}} = \frac{\Delta\vec{p}}{\Delta t} \tag{5.48}$$

This is conceptually perhaps the easiest way to see what's going on. $m\vec{v}_d$ is the average momentum of each charge carrier in the conductor. In each interval $\tau_{\rm therm}$ between "active bumper" collisions, the carrier starts from (on average) rest and gains this much momentum from the field, and then is *brought suddenly back to rest* by the lattice. The "drag force" thus equals the ***average momentum change per unit time*** of the free charges as they move through the lattice of atoms, and depends on two easily understood parameters.

5.2.6: Ohm's Law

If you skipped over the last subtopical section, take a moment to look back at the boxed equations 5.44 and 5.47. These two results are the most important things to take from the omitted section if you are just skimming it without working through its arguments in any detail, as they contain the *testable* components of the Drude model. Equation 5.44 shows that it gives a current density proportional to $\vec{E}$ and hence to the potential difference across a piece of conductor of some given length. It is also proportional to the time between thermal collisions $\tau_{\rm therm}$ between the charge carriers and the atomic/molecular lattice, which in turn scales inversely with *the square root of the absolute temperature*. The implicit *scaling* of these relationships can easily be compared to observation even if some of the details of the development of the model were just estimated and perhaps were even a bit sketchy.

With these results in hand we can easily establish the connection between Ohm's Law (a well-known empirical result) and the Drude model. We just showed above that the current density $\vec{J}$ is proportional to the applied electric field $\vec{E}$. In so doing, we wrapped up all of the complexity – all the unknown stuff about a conductor, including, b, n, q, m, $\tau_{\rm therm}$– into a single parameter called the ***conductivity***:

$$\sigma_c = \frac{nq^2\tau_{\rm therm}}{m} = \frac{nq^2}{b} = \frac{q\rho_{\rm free}}{b} = \frac{1}{\rho_r} \tag{5.49}$$

In this equation we note the terrible collision of symbols that is (sadly) just the way it is when discussing the conductivity σ_c and its reciprocal, the ***resistity*** ρ_r of the material (also defined in this equation). As you no doubt have observed, physicists reuse the symbol for volume charge density ρ for resistivity, and worse, reuse the symbol for *surface* charge density σ for *conductivity*! Who invented this stuff! The equation for ρ_r even contains $\rho_{\rm free} = nq$, the density of *free* (conduction) charge, just to maximally confuse you! To do my best to help you out, I'm going to use ρ_r with the little 'r' subscript for resistivity, and will similarly label σ_c with a 'c' for conductivity, but you'll still need to be careful at first!

Fortunately, with a little practice you will rapidly learn to identify which symbol goes where from its units/dimensions and from context, so you eventually you won't be any more confused than you are by sentences like "The two hippopotami, each wearing a tu-tu that was too, too much, went over to the bar to order two beers." We can even *hear* this sentence and effortlessly track two (number), tu-tu (ballarina dress), too (comparator), to (short for 'toward'), and to (infinitive form of the verb 'order') without much thinking about it. So shall it eventually with you and symbol overloading in physics.

The resistivity and/or conductivity[80] is a characteristic of the material of the conductor in question and as the Drude model (and experience) suggests, depends on many things, such as absolute temperature and details in the structure of the conductor – almost certainly more than are included in the model or our crude estimates so far!

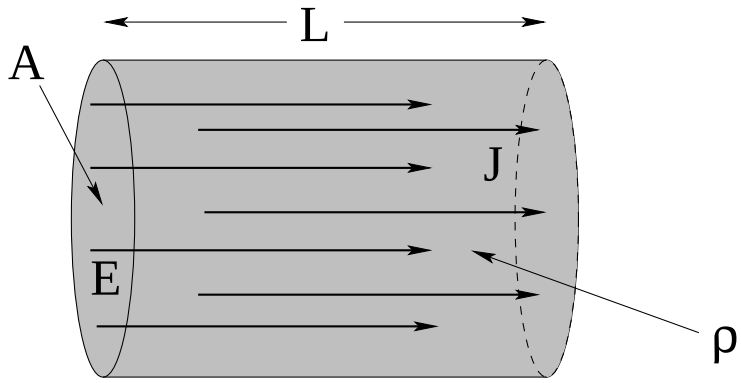


Figure 5.5: A simple resistor with resistivity ρ, length L, and cross sectional area A.

For now, let us consider an archetypical "resistor": a uniform conductor with resistivity ρ_r, length L, and cross-sectional area A (where the ends are at right angles to the sides), as pictured in figure 5.5.

[80]Wikipedia: http://www.wikipedia.org/wiki/Electrical resistivity and conductivity. As usual, follow this wikipedia link to learn more about resistivity and conductivity than this short treatment allows, as well as to access tables of resistivities and temperature coefficients of resistivity.

We can rearrange the current density equation 5.44 as:

$$\vec{\boldsymbol{E}} = \rho_r \vec{\boldsymbol{J}} \tag{5.50}$$

The electric field and current density inside of this volume are both uniform (in steady state, all of the charges must move through the volume at the same speed or charge would build up somewhere in the volume). The electrical current is the flux of the current through either end, so:

$$\int \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA = EA = \rho_r \int \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA = \rho_r I \text{ through the resistor} \tag{5.51}$$

which we can rearrange as:

$$E = I\frac{\rho_r}{A} \tag{5.52}$$

If we integrate both sides a second time in the direction $d\vec{\boldsymbol{l}}$ from one end of the conductor to the other in the direction of the current, we get the potential difference:

$$V_R = -\int \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{l}} = -EL = -I\frac{\rho_r L}{A} = -IR \tag{5.53}$$

V_R is thus the amount the electric potential *decreases* going from one side of the resistor R to the other *in the direction of the field/current*. We will often write the potential without the R subscript to simplify the algebra a tiny bit when there is no ambiguity introduced by so doing, and will similarly usually omit the sign and just remember that the potential *drops* going across a resistor in the direction of the current.

We've introduced a new quantity R, called the *resistance* of of the conducting material in this particular geometry:

$$R = \rho_r \frac{L}{A} \tag{5.54}$$

(known as ***Pouillet's law***, if you care) so that in terms of it:

$$\boxed{V \quad (\text{or } V_R) = IR} \tag{5.55}$$

where as noted we have left off the sign. This equation is known as *Ohm's Law* and we will use it extensively in the weeks to come. Note that we could equally well have called equation 5.44 Ohm's Law, as the two basically assert exactly the same thing, one in terms of current density and field, the other in terms of current and potential, and you may well see it referred to by this name in more advanced electrodynamics textbooks.

The SI units of the resistance are known as ***ohms*** (volts per ampere, obviously) and given the symbol Ω in most literature. Since a volt is a joule per coulomb, and an ampere is a coulomb per second,

$$1 \text{ ohm} = \frac{\text{joule} - \text{second}}{\text{coulomb}^2} = \frac{1 \text{ second}}{\text{farad}} \tag{5.56}$$

Note well that we used the fact that tthe SI units of capacitance, farads, are coulombs squared per joule, so the SI units of R times C are *seconds*, a pure time. This will be important to us by the end of this chapter.

Just from the simple relation $R = \rho_r L/A$ we can tell many things about the ways resistances will add in various configurations. If we put two identical resistances one right after another in a circuit, that's the same as one resistance twice as long, so we expect resistances in series

to *add*, increasing the total resistance. If we put two identical resistances in parallel, that's the same as one resistance with twice the area, which will *decrease* the resistance by a factor of two. We therefore expect that parallel resistance will obey a *reciprocal* addition rule. We will derive these two results more carefully below.

Before going on, it is worthwhile to point out the *analogy* between current flowing in a wire with finite resistance and water flowing in a pipe packed with something e.g. sand that similarly resists the flow of water. The flow of water through a sand-filled pipe is proportional to the *pressure* difference across the pipe, so pressure difference is analogous to voltage difference. The current of water is analagous to the current of charge. The resistance of the pipe is analogous to the resistance of the sand-filled pipe. A pipe twice as long will let half the water through at the same pressure difference. A pipe twice as wide will let twice the water through at the same pressure difference. There is even a "current density" for the water in motion that is the analogue of the current density of the charge. Even pipes that are *not* filled with sand have an "Ohm's Law" of the form $\Delta P = IR$ where R is the "resistance" of the pipe and I is the volumetric current in the pipe, as we discussed in the chapter on fluids in the first semester textbook.

This is really a rather compelling analogy, and since students are sometimes more comfortable visualizing the flow of water in pipes than they are imagining electrons flowing in wires, it is offered up to help you build up your conceptual understanding of the latter using your prior knowledge and experience of the former, where a day doesn't pass where you don't "switch on and off" the flow of water by means of increasing or decreasing the area of a pipe using a tap and where the flow of water out against the resistance of all of the plumbing isn't increased or decreased by the water pressure entering your house from the main.

In this analogy, a *capacitor* can also be visualized as a wide section of pipe containing a *piston on a spring*. The piston blocks water flow, but if one applies a pressure difference then water flows *into* the pipe section, compressing the spring, until the back-force of the spring balances the force on the piston due to the pressure difference. At that point this "capacitor" has stored some *water* on one side and has had an equivalent amount pushed *off* the other side, just like a regular capacitor. Note well that this suggests *correctly* that capacitors will dynamically behave like *springs* in an electrical circuit, storing potential energy and charge and releasing it back to the circuit, causing current and charge to *oscillate*. Later we'll discover a quantity and associated electrical device that behaves just like *mass* in such an analogous arrangement, and our analogical reasoning will be complete!

5.2.7: Dependence of Resistivity on Temperature

Before we leave the topic of resistivity/conductivity, we need to address an important question. As we've seen, the fundamental assumption of the Drude model, that charges are pushed through a resistance like pinballs drift lower in a very active pinball table, *correctly* leads to a linear dependence of $\vec{J}$ on $\vec{E}$, where a slightly simpler passive pinball model ends up proportional to $\sqrt{E}$, which does not agree with the *empirically verified* Ohm's Law in either of its forms.

However, I've also mentioned that even though it gets the E-dependence right, it turns out to be *wrong* in certain very important ways. Let's look at one of them – the expected

dependence of ρ_r (and hence R for any given chunk of material) on *temperature*.

Recall that in the Drude model, $\langle v \rangle = \sqrt{3kT/m}$, hence $\tau_{\rm therm} = d/\langle v \rangle \propto 1/\sqrt{T}$. From equation 5.44:

$$\frac{nq^2\tau_{\rm therm}}{m} = \frac{1}{\rho_r} \qquad \Rightarrow \qquad \rho_r = \frac{m}{nq^2\tau_{\rm therm}} \propto \frac{1}{\tau_{\rm therm}} \tag{5.57}$$

Ignoring most of the parameters of this as they are independent of temperature, we therefore we expect the resistivity to vary with absolute temperatre like:

$$\rho_r \propto \sqrt{T} \tag{5.58}$$

at least for T close to room temperature (at very low temperatures quantum phenomena come into play and at very high temperatures our simple model breaks down in other ways).

This is not what is observed. In fact, the resistivities of most common resistive metals increase approximately *linearly* with temperature, at least in the range of temperatures near room temperature, although they deviate significantly from this at low temperatures. Some metals aren't even linear *or* square root in temperature - they are best fit by power laws with exponents *not* equal to 1 or 1/2. Worse, insulators tend to have their (initially large) resistivity *decrease* with *increasing* temperature – going the *exact opposite* to the way we expect from anything *like* the classical Drude model.

This is why our classical discussion of resistance has been more to give you a *plausible* picture of conduction and resistance and explain *some* features of the result without ever claiming to be a *good* model or a *correct* model. To correctly understand resistance in materials, one simply has to use quantum statistical mechanics and electronic band theory from the beginning, making it a remarkably difficult subject to study or make quantitative predictions about. The classical predictions get *some* features right for *some* materials, but get some wrong (like temperature dependences) for nearly all materials.

Fortunately, none of this detail matters too much to people who want a *practical* description of temperature dependence that will work well enough for most materials near room temperature. In this case one can take the correct thermal dependence – whatever it might be – and develop a linearized *Taylor series expansion* and express the result with tabulated coefficients ρ_0 , α, and T_0 (a reference temperature, e.g. 20°C = 293°K corresponding to ρ_0 and α):

$$\rho_r(T) = \rho_0 \left\{1 + \alpha(T - T_0)\right\} \tag{5.59}$$

In this expression, ρ_0 is the resistivity at temperature T_0 and $\alpha = \frac{1}{\rho_0}\frac{\partial \rho}{\partial T}$ evaluated at $T = T_0$. α is called the *temperature coefficient of resistivity*. This equation allows resistivity to be accurately computed across a moderate, relevant, range of temperatures by means of three mutually tabulated quantities[81] .

However, this linearized expression, even, is too complicated for *most* of our purposes here[82]. In this introductory *classical* textbook we we will generally assume that $\alpha \approx 0$ (or that we are *at* $T = T_0$) so that $\rho = \rho_0$ for any given material and concentrate instead below on

[81] Wikipedia: http://www.wikipedia.org/wiki/Electrical resistivity and conductivity. This is a good article for you to look over to get a hint of the quantum theory as well as a useful table for many materials of these parameters in the linearized expression for $\rho_r(T)$.

[82] Unless, of course, you are a physics major or are interested in electrical engineering, in which case you would do well to at the very least earmark this discussion for future reference in more advanced courses.

the simple scaling of resistance with length and area of the resistor, Pouillet's law. Obviously, one cannot do this if one is designing circuits that heat up significantly as they operate or that have to function correctly across a wide range of temperatures, and this whole approach fails for things like semiconductors or superconductors that can be understood only with a correct treatment in quantum theory or very *different* functional equations.

5.3: Resistances in Series and Parallel

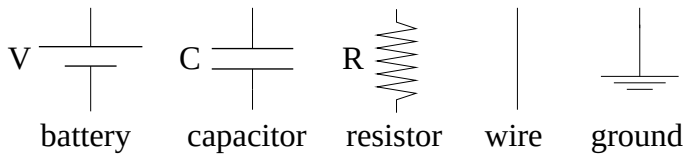


Figure 5.6: Symbols for batteries, capacitors, resistances, wires, and ground.

Before proceding any further, we need to add a symbol to our collection of symbols for circuit elements. We already have a symbol for capacitance, for a voltage source or battery and for a "wire", but now that conducting wires have this new property of resistance, we need to be a bit more specific. From now on, wires will be assumed to have ***zero resistance*** in all circuit diagrams. This specifically means, since $V_R = IR$, that the voltage drop across any ideal wire is *zero* independent of the current carried by that wire. Obviously, this is not physical, but if the resistance of the wire is important, it will (and should) be indicated as an explicit "resistor" in series with the wire in question that represents the resistance of that particular segment of wire. Resistance itself has the new symbol indicated above, typically labelled with its resistance value in Ohms or a suitably indexed R. Batteries and capacitances are unchanged (although both may have internal, non-ideal resistance that will similarly be represented by in-line series or parallel resistance symbols when appropriate). Finally, the ground symbol, indicating a specific potential of *zero* for all wires connected directly to it, is recapitulated.

We are now ready to draw collections of individual resistors connected in series or in parallel, and to derive the effective total resistance of these arrangements. These are pictured in figure 5.7.

5.3.1: Series

Suppose we apply a fixed voltage V_{ab} across the contacts in the upper (a) diagram. This produces some current $I_{\rm tot}$ in the *single* (serial) line of resistors. Since charge is conserved and there is nowhere for it to go but through the resistors, this same current passes through each resistor in turn. We can thus use Ohm's Law to determine the voltage drop across *each* resistor in terms of this total current:

$$V_1 = I_{\rm tot} R_1 \tag{5.60}$$

$$V_2 = I_{\rm tot} R_2 \tag{5.61}$$

$$V_3 = I_{\rm tot} R_3 \tag{5.62}$$

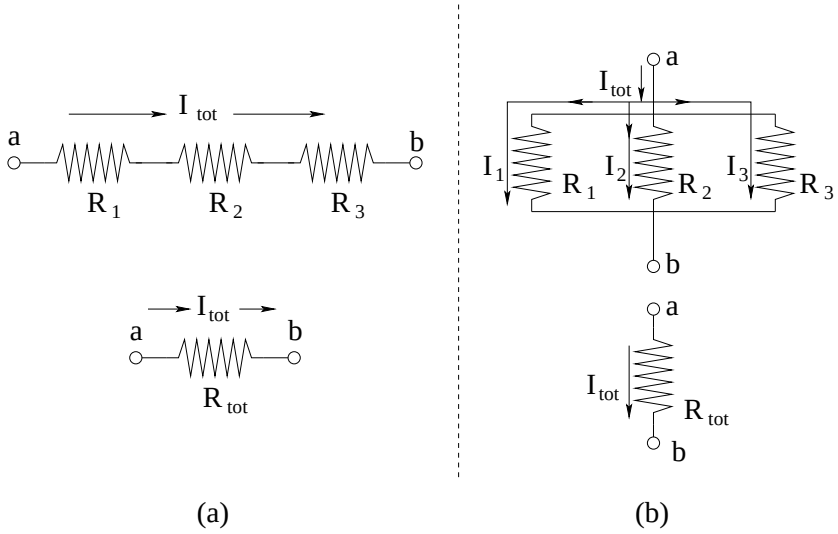


Figure 5.7: Three resistors R_1, R_2, R_3 arranged in *series* (left, (a)) and *parallel* (right, (b)), along with the equivalent/total resistances of each one portrayed below. In both cases the total resistance is "equivalent" when applying a voltage V_{ab} across the a and b contacts produces the *same total current* $I_{\rm tot}$ *in the top and bottom figure.*

Obviously the total voltage V_{ab} is given by:

$$V_{ab} = V_1 + V_2 + V_3 = I_{\rm tot}(R_1 + R_2 + R_3) \tag{5.63}$$

If we look at the lower (a) diagram, Ohm's Law yields:

$$V_{ab} = I_{\rm tot} R_{\rm tot} \tag{5.64}$$

Equating and cancelling the common $I_{\rm tot}$, we get:

$$R_{\rm tot} = R_1 + R_2 + R_3 \tag{5.65}$$

There was nothing "special" about having only three resistors. We could have had, four, five, or N resistors in series and we'd simply have more terms in a general equation:

$$V_{ab} = \sum_{i=1}^{N} I_{\rm tot} R_i = I_{\rm tot} \sum_{i=1}^{N} R_i = I_{\rm tot} R_{\rm tot} \tag{5.66}$$

so that *in general* the rule for the addition of N resistors in series is:

$$R_{\rm tot} = R_1 + R_2 + ... + R_N = \sum_{i=1}^{N} R_i \tag{5.67}$$

5.3.2: Parallel

In the case of resistances in parallel, we have the *same* voltage V_{ab} applied across all of the resistors in parallel. If we look at the upper (b) figure, we can use Ohm's Law to evaluate the

current through each resistor, given a common voltage V_{ab} across them:

$$I_1 = \frac{V_{ab}}{R_1} \tag{5.68}$$

$$I_2 = \frac{V_{ab}}{R_2} \tag{5.69}$$

$$I_3 = \frac{V_{ab}}{R_3} \tag{5.70}$$

Now, consider the total current $I_{\rm tot}$ flowing into the arrangement from point a. Charge is conserved, so that all of the charge that flows into the first junction connecting the three independent conducting pathways through the resistors *must flow out of it* and into the three resistors. From this we conclude that:

$$I_{\rm tot} = I_1 + I_2 + I_3 = \frac{V_{ab}}{R_1} + \frac{V_{ab}}{R_2} + \frac{V_{ab}}{R_3} = V_{ab}\left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}\right) \tag{5.71}$$

As before in the lower (b) figure we have:

$$I_{\rm tot} = \frac{V_{ab}}{R_{\rm tot}} \tag{5.72}$$

and when we equate these two forms and cancel the common V_{ab} we get:

$$\frac{1}{R_{\rm tot}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} \tag{5.73}$$

There is nothing special about three resistors, and once again we can easily generalize this argument to N resistors as:

$$\frac{1}{R_{\rm tot}} = \frac{1}{R_1} + \frac{1}{R_2} + ... + \frac{1}{R_N} = \sum_{i=1}^{N} \frac{1}{R_i} \tag{5.74}$$

We conclude that the total resistance of several resistors in series is the simple sum of the individual resistances, while the *reciprocal* of the total resistance of serveral resistors in parallel is the sum of the *reciprocals* of the individual resistances. This is the exact opposite of the rules for summing capacitances in seris and parallel.

5.4: Kirchhoff's Rules and Multiloop Circuits

In the previous sections we used two rules implicitly that we should make explicit so that we can use them in the more complicated circuits we will study over the next few weeks. In studying series capacitors and series resistors, we used the idea that we could *add* the changes in voltage across objects in a common wire carrying a steady state current (including no current at all) to find the voltage changes between any two points in the wire. This is an idea related to *energy conservation*. In studying parallel capacitors and and parallel resistors, we used the idea that the total charge moving around in these circuits must be conserved to track its distribution over time whether or not it is actually moving.

These two rules (which we will derive and discuss below) are known as Kirchhoff's Rules[83]
.

[83]Wikipedia: http://www.wikipedia.org/wiki/Kirchhoff's Circuit Laws.

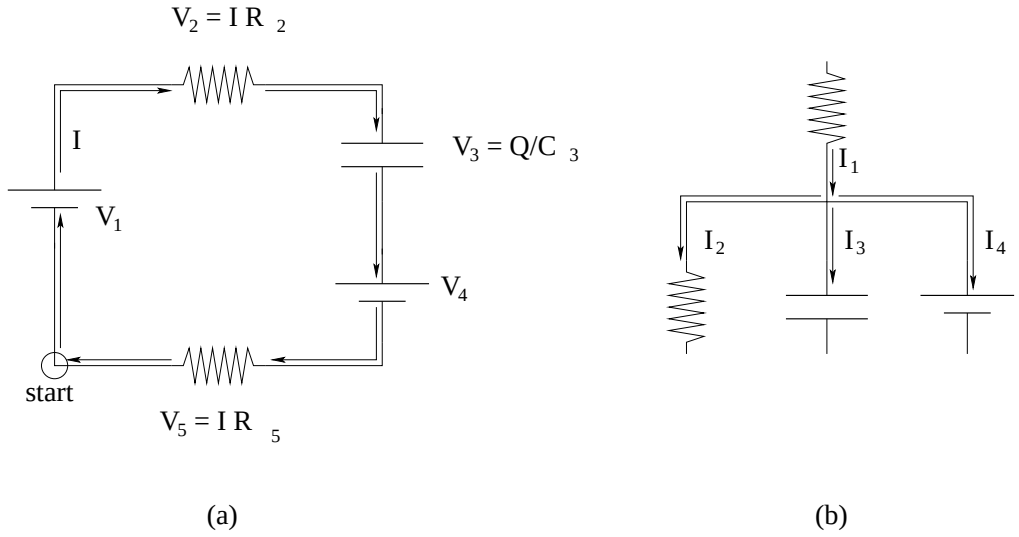


Figure 5.8: (a) A single "generic" circuit loop; (b) A single "generic" circuit junction.

5.4.1: Kirchhoff's Loop Rule

Consider the generic *circuit loop* in figure 5.8 (a) above. The particular devices in this loop are not too important – I drew a fairly arbitrary mix of the three devices we are aware of so far, but later we will learn about still more devices we might want to put into a circuit to do some startlingly useful things.

Let us imagine that we watch a charge $+q$ moving around this circuit loop in the direction of the current beginning at the (arbtrary) point "start". As it goes across each potential $V_1, V_2, ...$ the energy of the charge goes up, goes down, goes up, goes down. By the time it gets back to the start position, its potential energy has changed by:

$$\Delta U = qV_1 + qV_2 + qV_3 + qV_4 + qV_5 = q\sum_i V_i \tag{5.75}$$

If $\Delta U \neq 0$, then the charge gets back to its starting point with a *different energy* than the one it started with! Its kinetic energy will have changed!

However this is *almost* impossible. Electrons in particular, as fermions, are nearly *completely incompressible* in a wire. This means that the current in any line segment is the same at all points in the segment. Changes in the electric field that *produces* the current at all points in the conductor propagate nearly *instantaneously* throughout the entire loop, because the speed of light is very large compared to the size of the loop. As potentials across the elements in the circuit vary, the current adjusts almost instantaneously. Consequently within a *very* tiny margin associated with this propagation time, the net energy gain or loss of a charge in a pass around the circuit loop must be *zero!*

This means that:

$$\sum_i^{\text{loop}} V_i = 0 \tag{5.76}$$

is a simple statement of *energy conservation* for the charges as they progress around the loop. This equation is known as *Kirchhoff's Loop Rule*, and we will use it repeatedly to write down

equations that lead to equations of motion for dynamical circuit loops or conditions that must be satisfied for loops that carry steady state currents.

5.4.2: Kirchhoff's Junction Rule

Consider the generic *circuit junction* in figure 5.8 (b) above. Again it doesn't matter much what devices are on any of the legs. Charge is conserved – it is neither created nor destroyed. The junction itself cannot act as a reservoir for charge – it has negligible capacitance because it is part of a continuous volume of (presumed perfect) conductor that can conduct any charge surplus of the (incompressible) charge away as rapidly as it develops.

This means that all the charge going *into* the junction has to go *out* of the junction along the various wires that join together at the junction. This rule can be written, and thought of, in two different ways:

$$I_{1,\text{in}} - I_{2,\text{out}} - I_{3,\text{out}} - I_{4,\text{out}} = 0 \tag{5.77}$$

with the *convention* that current going *into* the junction is positive and current coming *out* of the junction is negative. Alternatively, you can sort out the currents coming in and the currents going out and equate them:

$$I_{1,\text{in}} = I_{2,\text{out}} + I_{3,\text{out}} + I_{4,\text{out}} \tag{5.78}$$

Note that these two equations are the same.

Let us generalize the first form to:

$$\sum_i^{\text{junction}} I_i = 0 \tag{5.79}$$

and call it *Kirchhoff's Junction Rule* (using the $\pm$ convention). Remember, the junction rule is just the symbolic expression of charge conservation, just as the loop rule is the symbolic expression of energy conservation.

Just for grins, let's put both rules down side by side, the way you should probably remember them:

$$\sum_i^{\text{loop}} V_i = 0 \qquad \text{Loop Rule} \tag{5.80}$$

$$\sum_i^{\text{junction}} I_i = 0 \qquad \text{Junction Rule} \tag{5.81}$$

Example 5.4.1: The Internal Resistance of a Battery

Previously, we indicated that any real battery (or electrical power supply, not necessarily a battery) is incapable of doing an *infinite* amount of work or deliving an *infinite* amount of power. If you put a load of any sort on a real battery, you can increase the load (the power draw) up to a point, but if you try to draw *more power than the battery can deliver*, the net power delivered to the load will actually *decrease*.

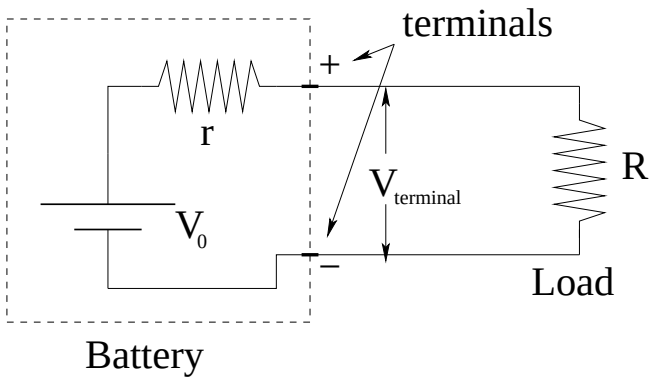


Figure 5.9: A non-ideal battery in a circuit with a resistive load.

This is actually a key principle of electrical design, where one often wishes to deliver the *maximum* power to some load – a speaker, a radio antenna, even a light bulb – that the power supply will support.

If one ***short-circuits*** a battery – connects a very low/zero resistance across its terminals – then the battery will usually deliver its maximum power, and its maximum possible current. A very simple, but quite accurate, model for this limiting is indicated in the figure 5.9 above. In it, a hypothetical chemical battery is represented as the two circuit elements inside the dotted box. One is the actual chemical potential generated by the chemical reaction. This is called the ***internal voltage*** of the battery[84]. As we shall see, this is also the voltage between the terminals of the battery when there is *no load*, if the chemical process has not exhausted the reactants (if you like, the "fuel" of the battery). In addition to this, the battery is considered to have an ***internal resistance*** r that limits the current the batter can deliver even when completely short circuited.

We are now (with both Kirchoff's rules and/or series resistances in hand) well capable of understanding how all of this works. Kirchoff's rule for the circuit loop is:

$$V_0 - Ir - IR = V_0 - I(r + R) = 0 \tag{5.82}$$

or:

$$I = \frac{V_0}{r + R} \tag{5.83}$$

The ***terminal voltage*** is defined to be $V_t = V_0 - Ir$, the voltage between the *physical terminals* of the battery when it is delivering any given current I. If $R = \infty$, $I = 0$ and $V_t = V_0$ as indicated above. If $R = 0$ (the battery is "short circuited" when a zero resistance is connected across the terminals) we find that:

$$I_{\max} = \frac{V_0}{r} \tag{5.84}$$

[84]This was historically called the "electromotive force", or "EMF" of the battery, and it is still often represented as $\mathcal{E}$ in physics textbooks and called the EMF. I find it difficult to call or label something that is clearly a *voltage* a *force*, even by obscure inheritance. This is doubly so for chemistry, where the actual motivation is caused by the discrete *quantum* energy changes between the reactants and the products and where it is a lot of work to even define a good quantum analog of "force" at all. I therefore rebel in my own small way and just call a voltage a voltage and differentiate only with modifiers.

In practical terms, the internal voltage is usually known, fixed by the chemistry of the battery, and one can *measure* the internal resistance indirectly by short-circuiting the battery while measuring the delivered current. As batteries are discharged (or as rechargable batteries age) this internal resistance increases until their terminal voltage effectively drops to zero if a load of any sort is connected across the terminals.

Note well that we can easily compute the power delivered to the internal resistance (the battery itself, generally heating up the battery with its internal Joule heating) versus its load resistance R:

$$P_r = I^2 r = V_0^2 \frac{r}{(r+R)^2} \tag{5.85}$$

and

$$P_R = I^2 R = V_0^2 \frac{R}{(r+R)^2} \tag{5.86}$$

The sum of theses add up to the total power provided to the circuit by the internal voltage/energy source, as it must.

It is an instructive exercise to demonstrate that the power delivered to the load is a ***maximum*** when $r = R$, when the load resistance matches the internal resistance of the power supply. This is called *impedance matching* – impedance is a sort of generalized resistance that we will study in more detail in the chapter on AC circuits, but in the case of DC circuits it is equal to ordinary resistance. Impedance matching is an essential part of the engineering of things like earphones or speakers, where one limits the power deliverable to the load by any given amplifier.

Example 5.4.2: A Multiloop Resistance Problem

Although we will have many opportunities to use Kirchhoff's Rules in the chapters to come, it is worthwhile to apply it to an archetypical problem where it is necessary to use both rules to determine the currents in a multiloop circuit with resistors and batteries. The problem doesn't have to be particularly difficult, but it does need to illustrate all of the steps required to solve problems of this time, as well as some of the caveats – places where things one might try don't advance you towards the solution.

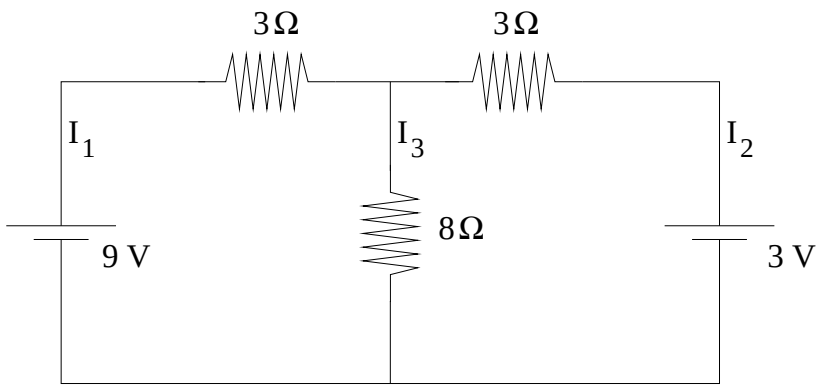


Figure 5.10: Use Kirchhoff's Rules to find the three unknown currents: I_1, I_2, I_3.

In figure 5.10, we see a typical arrangement of batteries and resistors in a multiloop problem. There are *three loops* and *three currents* visible in the problem (can you see the three

loops?). Our job is to *find the three unknown currents* given the information on the figure. We have to do this by writing Kirchhoff's loop and current rules *algebraically*, using the unknown currents, and then solve the resulting system of simultaneous equations to find the currents.

The *first* step, however, is to identify the loops and choose *tentative directions* for the currents. We don't need to worry yet about whether or not they are correct – our eventual answers will tell us the correct directions by means of their *signs* relative to our initial assumptions. Let's redraw the figure, appropriately decorated with current directions and loops identified:

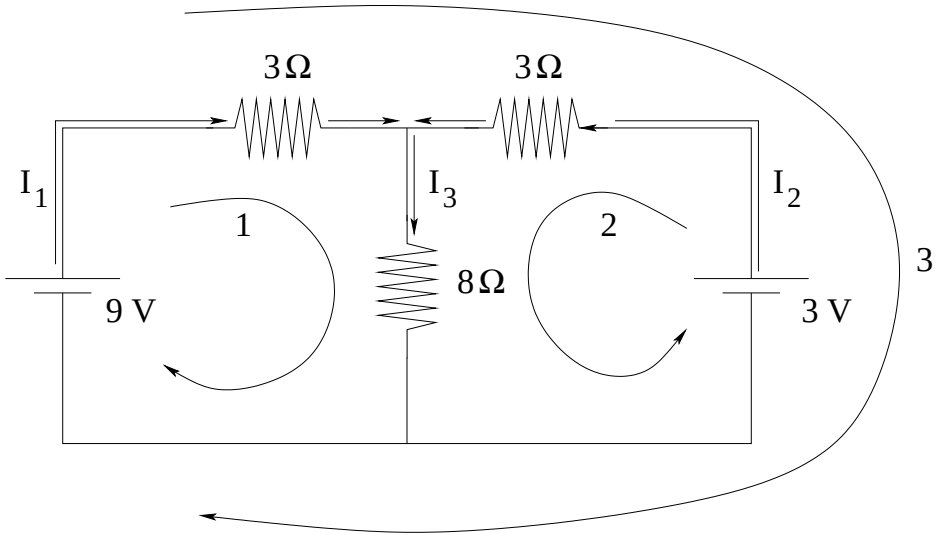


Figure 5.11: Note the loops and current directions identified on the figure.

Now let us write (and identify) all four equations that we can obtain from Kirchhoff's Rules in this problem:

$$9 - 3I_1 - 8I_3 = 0 \quad \text{(loop 1)} \tag{5.87}$$

$$3 - 3I_2 - 8I_3 = 0 \quad \text{(loop 2)} \tag{5.88}$$

$$9 - 3I_1 + 3I_2 - 3 = 0 \quad \text{(loop 3)} \tag{5.89}$$

$$I_1 + I_2 - I_3 = 0 \quad \text{(junction)} \tag{5.90}$$

Recall that the potential *decreases* when we go across a resistor *in the direction of the current*. We do not write the equation for the bottom junction because it is just -1 times the top junction equation and hence not independent.

We immediately notice that there is a wee problem – we have *four equations* and only *three unknowns!* This means that our equations cannot all be independent. If you examine the first three equations, a moment of reflection should convince you that the third equation (for loop 3) is the equation for loop 2 minus the equation for loop 1. This is *characteristic* of multiloop problems – the sum or difference of interior loops always adds up to exterior loops as the inner/shared voltages *cancel*.

This is very important to remember when we solve the simultaneous equations – *adding loop equations to eliminate variables does not make progress towards the solution!* It just gives you another loop equation. In order to make progress, you *must* use the junction equation(s) and a *subset* of the loop equations. Let's dump the equation for loop 3 and keep only the three

we need to solve the problem. With a bit of rearrangement, we get:

$$3I_1 + 8I_3 = 9 \tag{5.91}$$
$$3I_2 + 8I_3 = 3 \tag{5.92}$$
$$I_1 + I_2 = I_3 \tag{5.93}$$

There are many ways to proceed to find a solution to this linear system. One can line up the I's, form a matrix equation, and invert the matrix using more or less standard determinants and linear algebra. One can line up the I's and do Gauss elimination (being careful to use the junction rule before the loop rules) followed by back substitution. Or in the case of systems as simple as this one, one can just use substitution to eliminate one of the currents using the junction equation, then eliminate one of the two remaining currents (followed by back substitution), a sort of sloppy Gauss elimination. Being a sloppy kind of guy (and not wanting to teach a course in linear algebra on top of everything else) I'm going to illustrate the solution of this problem with this latter approach, but if you are down with using Cramer's Rule (the fancy name for the first approach) so am I.

So we substitute $I_3 = I_1 + I_2$ into the two voltage equations:

$$3I_1 + 8I_1 + 8I_2 = 9 \tag{5.94}$$
$$3I_2 + 8I_1 + 8I_2 = 3 \tag{5.95}$$

or

$$11I_1 + 8I_2 = 9 \tag{5.96}$$
$$8I_1 + 11I_2 = 3 \tag{5.97}$$

If we multiply the top equation by 11 and the bottom equation by 8, we get:

$$121I_1 + 88I_2 = 99 \tag{5.98}$$
$$64I_1 + 88I_2 = 24 \tag{5.99}$$

If we subtract the second equation from the first, we get:

$$57I_1 = 75 \tag{5.100}$$

or

$$I_1 = \frac{75}{57} = 1.316 \tag{5.101}$$

(in Amperes). We substitute this back into:

$$11\frac{75}{57} + 8I_2 = 9 \tag{5.102}$$

so

$$I_2 = (9 - 11\frac{75}{57})/8 = -0.785 \tag{5.103}$$

Finally

$$I_3 = I_1 + I_2 = 1.316 - 0.785 = 0.531 \tag{5.104}$$

Note well that I_2 comes out *negative* – this simply means that we guessed its direction *incorrectly* in our original decoration of the figure. The second battery is actually being charged

as the first one discharges. This is (as you can see from the numbers) about as nasty a problem of this sort as you are likely to see. Usually problems like this on a quiz or exam will have voltages and resistances that are chosen to give rational answers that one can work out without needing a calculator.

5.5: RC Circuits

So far everything we have done with charges and currents has been *static*. True, we have studied flowing currents but those currents have been *constant* in time, as have all potential differences. We now illustrate the use of Kirchhoff's Loop Rule to obtain an *equation of motion* for the charging or discharging of a capacitor through a resistance. We begin with a discharging capacitor, as the slightly easier problem.

Example 5.5.1: Discharging Capacitor

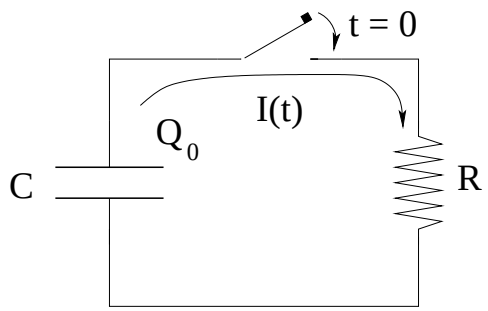


Figure 5.12: The capacitor C is initially charged to Q_0. At $t = 0$ the switch is closed and it discharges through the resistor, building up a current $I(t)$.

The capacitor in figure 5.12 is initially charged to Q_0. At $t = 0$, the switch is closed and charge begins to flow off of the capacitor and is driven through the resistor, so that at time t there is a charge $Q(t)$ left on the capacitor and a current $I(t)$ in the circuit. Our goal is to basically understand *everything* about this problem. We want to know $I(t)$, $Q(t)$, $V_C(t)$, $V_R(t)$, the power $P_C(t)$ delivered by the capacitor, the power $P_R(t)$ consumed by the resistor, and a full understanding of energy as a function of time in the circuit.

To find all of this, we begin by writing *Kirchhoff's Loop Rule* for the loop above (going clockwise around the circuit in the direction of the current), at some time t after the switch is closed:

$$\frac{Q}{C} - IR = 0 \tag{5.105}$$

The current and charge are not independent. The current is, in fact, the rate at which the charge on the capacitor decreases:

$$I = -\frac{dQ}{dt} \tag{5.106}$$

If we substitute this relation into Kirchhoff's loop rule, divide by R, and rearrange, we get

the following equation of motion for Q:

$$\frac{dQ}{dt} + \frac{Q}{RC} = 0 \tag{5.107}$$

This is a *first order, linear, homogeneous, ordinary differential equation*, in fact the equation for *exponential decay*. It can easily by solved by direct integration. The solution proceeds as follows. Rearrange the equation as follows:

$$\frac{dQ}{dt} = -\frac{Q}{RC} \tag{5.108}$$

Multiply through by dt, divide through by Q, to get:

$$\frac{dQ}{Q} = -\frac{dt}{RC} \tag{5.109}$$

Integrate both sides (indefinite integral on the right):

$$\ln(Q) = -\frac{t}{RC} + A \tag{5.110}$$

(where A is the constant of integration). To get Q, we exponentiate both sides:

$$Q(t) = e^{\ln(Q)} = e^{-\frac{t}{RC}+A} = e^{A}e^{-t/RC} \tag{5.111}$$

Finally, we set the constant of integration from the initial conditions, so that $Q(0) = Q_0$:

$$Q(t) = Q_0 e^{-t/RC} \tag{5.112}$$

From this we can easily find the other quantities mentioned above:

$$\begin{aligned}
I(t) &= -\frac{dQ}{dt} = \frac{Q_0}{RC}e^{-t/RC} && (5.113)\\
V_C(t) &= \frac{Q}{C} = \frac{Q_0}{C}e^{-t/RC} = V_0 e^{-t/RC} && (5.114)\\
V_R(t) &= -I(t)R = -\frac{Q_0}{C}e^{-t/RC} && (5.115)\\
P_C(t) &= V_C(t)I(t) = \frac{Q_0}{C}e^{-t/RC}\frac{Q_0}{RC}e^{-t/RC} \\
&= \frac{Q_0^2}{RC^2}e^{-2t/RC} && (5.116)\\
P_R(t) &= V_R(t)I(t) = -\frac{Q_0}{C}e^{-t/RC}\frac{Q_0}{RC}e^{-t/RC} \\
&= -\frac{Q_0^2}{RC^2}e^{-2t/RC} && (5.117)\\
& && (5.118)
\end{aligned}$$

Note well that the power *delivered to* (+) the circuit by the capacitor equals the power *used by* (-) the resistor!

The final little piece of magic we can look for is energy balance. Suppose we wait a very long ("infinite") time – we expect the charge on the capacitor to go to zero in that time. How

much energy appears in the resistor during that entire period?

$$\begin{aligned} U_R &= |/int_0^\infty P_R(t)dt| \\ &= \frac{Q_0^2}{RC^2}\int_0^\infty e^{-2t/RC}dt \\ &= -\frac{Q_0^2}{2C}\int_0^\infty e^{-2t/RC}\frac{-2\,dt}{RC} \\ &= -\frac{Q_0^2}{2C}\,e^{-2t/RC}\Big|_0^\infty \\ &= \frac{Q_0^2}{2C} \qquad (5.119) \\ & \qquad (5.120) \end{aligned}$$

which just *happens* to be the total energy initially on the capacitor:

$$U_C = \frac{1}{2}\frac{Q_0^2}{C} \tag{5.121}$$

The argument of an exponential (or any transcendental function) has to be dimensionless, so the units of RC must be a *time*, the so-called *exponential decay time* for the circuit:

$$\tau = RC \tag{5.122}$$

This is an important quantity to keep in mind when working with RC circuits, as it provides an instant estimate for how long it will take for the charge on the capacitor to decay.

Example 5.5.2: Charging Capacitor

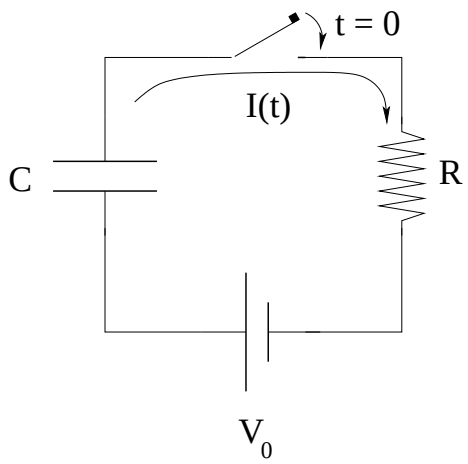


Figure 5.13: An initially uncharged capacitor being charged through a resistor by a battery with a fixed voltage V_0.

In figure 5.13 we have added a battery and changed the initial condition to $Q(0) = 0$, an initially uncharged capacitor. The solution to the problem proceeds *almost identically* to the charging case. From Kirchhoff's loop rule:

$$V_0 - \frac{Q}{C} - IR = 0 \tag{5.123}$$

The current is now the rate at which the charge on the capacitor *increases*:

$$I = +\frac{dQ}{dt} \tag{5.124}$$

Substituting as before and rearranging, we get:

$$\frac{dQ}{dt} + \frac{Q}{RC} = \frac{V_0}{R} \tag{5.125}$$

This is a *first order, linear,* in*homogeneous, ordinary differential equation*, in fact the equation for *exponential growth*. It, too, can easily by solved by direct integration:

$$\frac{dQ}{dt} = -\frac{Q}{RC} + \frac{V_0}{R} \tag{5.126}$$

Now, *pay attention* for a second, as it took me years of solving this inefficiently before I finally figured out how to do the algebra *efficiently*, and I'm going to share a little trick with you that will help you get the right answer for this equation (which occurs over and over again in physics, both last semester and this): *Before* multiplying out and trying to integrate *factor the coefficient of Q out of the entire left hand side!*:

$$\frac{dQ}{dt} = -\frac{1}{RC}\left(Q - CV_0\right) \tag{5.127}$$

Now multiply through by dt, divide through by $Q - CV_0$:

$$\frac{dQ}{Q - CV_0} = -\frac{dt}{RC} \tag{5.128}$$

and integrate both sides (indefinite integral on the right) to get:

$$\ln(Q - CV_0) = -\frac{t}{RC} + A \tag{5.129}$$

(where A is the constant of integration). As before, to get Q, we exponentiate both sides:

$$Q(t) - CV_0 = e^{\ln(Q - CV_0)} = e^{-\frac{t}{RC} + A} = e^A e^{-t/RC} \tag{5.130}$$

Finally, we solve for $Q(t)$:

$$Q(t) = CV_0 + e^A e^{-t/RC} CV_0 + Be^{-t/RC} \tag{5.131}$$

and set the constant of integration $B = e^A$ (the exponential of an unknown constant is still an unknown constant[85]) from the initial conditions, so that $Q(0) = 0$. Our final answer is:

$$Q(t) = CV_0\left(1 - e^{-t/RC}\right) \tag{5.132}$$

It is left as an exercise to evaluate the same list of quantities that we did for the discharging capacitor: $I(t), V_C(t), V_R(t), P_C(t), P_R(t)$. To this we add $P_V(t)$, the total power provided to the circuit by the voltage, and suggest that you demonstrate that as $t \to \infty$ the total energy provided to the circuit by the voltage equals the total energy stored in the capacitor in the end

[85] At your convenience, meditate upon the *units* implicit in this constant and figure out how they make it through the process above, where certain things have to be dimensionless and others do not...

plus the total energy burned in the resistor. Note well that because our solution was based on Kirchhoff's loop rule, which *is* the constraint that work-energy be satisfied, it should come as no surprise that in the end energy conservation is precisely embodied in the full integrated solution we obtain.

Yet to me, it always does. There is something amazing, almost magical, in the way that energy conservation works out in the equations of electromagnetism, given the complexity, the structure, the *detail* we see in the many different problems we work throughout the semester and beyond (as electromagnetism is a major foundation of our understanding of *everything*, in both classical and quantum physics). But it does.

We live in an enormously conservative Universe, where there are, quite rigorously, *no free lunches*, where mass-energy *never* whimsically appears or disappears, where one can, with sufficient care, trace out and balance every conserved quantity in any problem no matter how many bodies are involved or how complex the dynamics of the system.

This concludes our examination of RC circuits and our return to the world of dynamical equations of motion with nontrivial solutions, in this case exponential solutions (although we have done our best to keep our hand in with the occasional "discovered" oscillator or constant acceleration problem on the homework so far). RC circuits are quite important and occur in nature as well as in most electronic devices, where they are often used for timing purposes or where RC exponential charging or discharging behavior is an artifact of the circuit design that "softens" the edges of sudden square-wave-like transitions in voltage as they propagate into a circuit leg with nonzero resistance and capacitance.

The most important place that they occur in nature is probably inside the brain. The nervous system is decently modelled by neurons as tiny bioelectrical batteries that charge up capacitance across a membrane with variable resistance, a resistance that goes from very high to very low "suddenly" as the membrane depolarizes and channels open that permit the transport of e.g. sodium ions. As such there is a "rise time" required to charge up a neuron to where it can fire, followed by a sudden exponential drop in charge across the membrane when it does fire to create an electrical pulse capable of triggering the next neuron(s) down the network. From nothing but this we can deduce a number of important properties of biological neural networks: They have a maximum firing rate (consider the charging/discharging curves, where one has to exceed some threshold in order to be able to trigger downstream neurons upon depolarization). They consume energy, as all of the teensy biological batteries that charge them up deliver power to the circuit – the human brain, for example, consumes around 1/4 of the metabolic energy used by the entire human body, some 25 watts (out of 100 watts total). Neurotoxins such as *tetrodotoxin*[86] which block the sodium channel effectively freeze the otherwise variable resistance of the capacitative membrane, locking each neuron in the "charged" state and preventing the triggered discharge that is required for normal operation. Various nervous system disorders are related to "short circuiting" this network (by e.g. altering the resistance of the myelin sheaths that protect the axons of the neurons as they transport the current pulse downstream to the next neural synapse. Other disorders or neurotoxins are associated with the neurotransmitter-mediated transport across the synaptic gaps themselves.

[86]Wikipedia: http://www.wikipedia.org/wiki/tetrodotoxin. Found in pufferfish and blue-ringed octupi, for the marine biology crowd.

Basically, one cannot even *begin* to understand the biology of the nervous system of any organism without at least a conceptual understanding of batteries, resistances, and capacitances, and a *sound* conceptual understanding is always based on having really gone through the whole thing and worked it all out, in detail, at least one time in your life. So even if you don't plan to become a physicist and work on all of this (very cool) stuff for the rest of your life, *pay attention and work hard* on it now, because if you do you will reap the rewards in your work in *other* disciplines, where you will discover it lurking, time and again, to confound your understanding if you never worked hard enough to master it now.

This concludes our treatment of electrostatics with our first electro*dynamic* model. It is time to move on from the electrostatic field to the next major piece of the electromagnetic puzzle: The magnetic field.

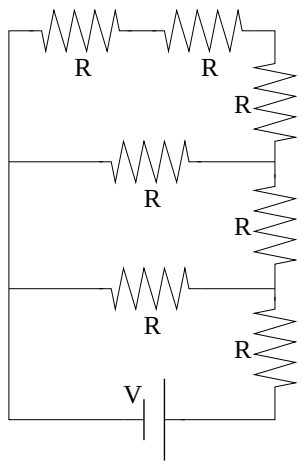
Homework for Week 5

Problem 1.

Physics Concepts

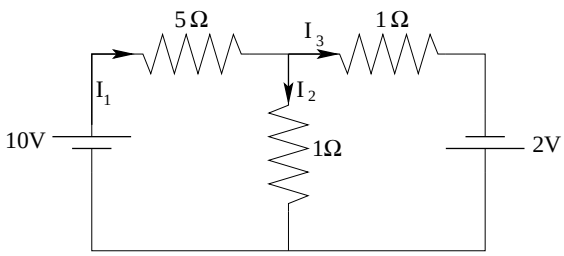
Make this week's physics concepts summary as you work all of the problems in this week's assignment. Be sure to cross-reference each concept in the summary to the problem(s) they were key to. Do the work carefully enough that you can (after it has been handed in and graded) punch it and add it to a three ring binder for review and study come finals!

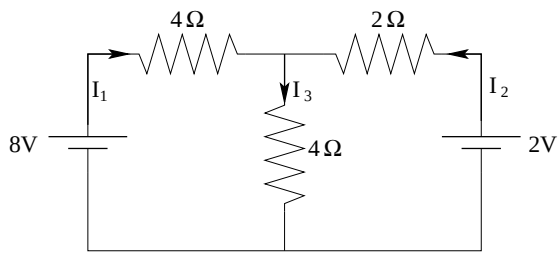
Problem 2.



Find the current through each resistor with a voltage V is placed across the resistance network as shown to the left. Note that all of the resistances R are equal. You'll basically need to use the series and parallel rules for adding resistances several times, as well as Ohm's Law and Kirchhoff's junction rule. **Hint:** You may find it useful to imagine $V = 18$ volts and $R = 1$ ohm. This makes the *numbers* easy, although it isn't that difficult to do this just with algebra.

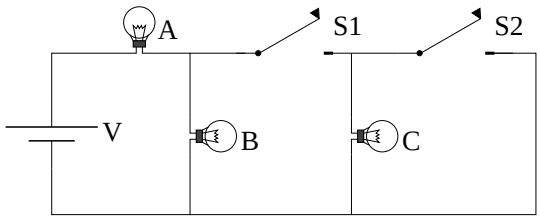
Problem 3.





Find the currents I_1, I_2, and I_3 in the two two loop, two voltage circuit above. Use the current directions ***as given*** in the figures.

Problem 4.



The circuit shown begins with switches S1 and S2 open as shown. Assume that the brightness of the identical bulbs shown increases monotonically with the current through the bulb, and otherwise imagine them to be identical resistances R. Initially, A and B are both on and equally bright. Identify the true statements in each line below:

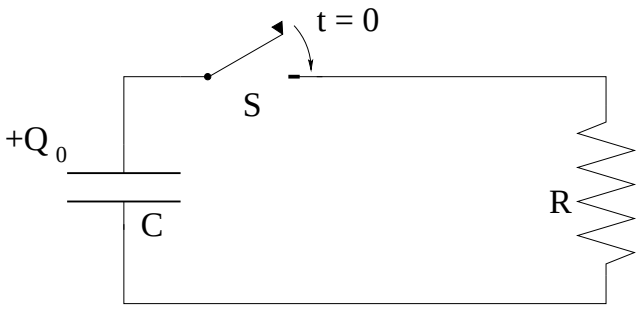
When Switch S1 is closed:

☐ A gets brighter	☐ A gets dimmer	☐ A goes off
☐ B gets brighter	☐ B gets dimmer	☐ B goes off
☐ C is as bright as A	☐ C is as bright as B	☐ C is off

When Switch S1 *and* S2 are closed:

☐ A gets brighter	☐ A gets dimmer	☐ A goes off
☐ B gets brighter	☐ B gets dimmer	☐ B goes off
☐ C is as bright as A	☐ C is as bright as B	☐ C is off

Problem 5.



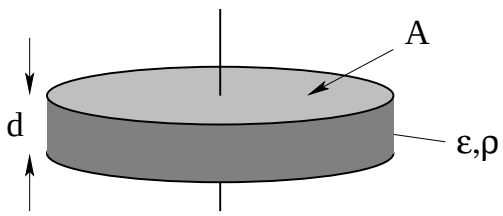
Suppose switch S is closed at time $t = 0$ when the charge on the capacitor is $Q_0 = Q_0$.

a) Write Kirchoff's Loop Rule for the circuit at an arbitrary time t after the switch is closed. Convert this into a (first order) equation of motion for Q (the charge on the capacitor).

b) Integrate the equation of motion to find $Q(t)$, $I(t)$, $V_C(t)$ and $V_R(t)$.

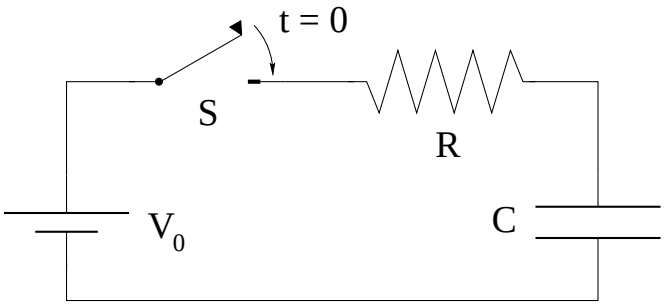
c) Using your results to b), show that the *net* power delivered to the circuit is zero, that is that the rate that the capacitor loses energy equals the rate that the energy appears (as heat!) in the resistor. This basically verifies that energy is conserved in this circuit (as it must be, given KLR).

Problem 6.



In the diagram to the left, a cylindrical "leaky capacitor" with area A and plate separation d is drawn. It is filled with a material that is ***both*** a dielectric with permittivity ϵ ***and*** a resistance with resistivity ρ so that any charge placed on the isolated capacitor decays with an exponential time constant τ. Find τ in terms of the givens. Does it depend on A and/or d?

Problem 7.



Suppose switch S is closed at time $t = 0$ when the charge on the capacitor is $Q_0 = 0$.

a) Write Kirchoff's Loop Rule for the circuit at an arbitrary time t after the switch is closed. Convert this into a (first order) equation of motion.

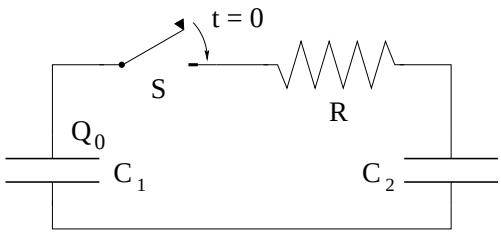
b) Integrate the equation of motion to solve for $Q(t)$ and use this result to find $I(t)$.

c) When the switch has been closed for a very long time, the capacitor is fully charged and the current I has gone to zero. Find the total work done by the voltage W_V, the total energy turned into heat in the resistor W_R, and the total energy stored on the capacitor W_C, and use the result to answer the following question: If one wishes to *waste* the least energy as heat charging the capacitor, should one make $R > 0$:

☐ As small as possible ☐ As large as possible

☐ The energy wasted doesn't depend on R

Problem 8.



A pair of capacitors C_1 and C_2 is connected as shown, with a resistance R in between them. Initially, C_1 carries a charge Q_0 and C_2 is uncharged. At $t = 0$ the switch is closed. Find:

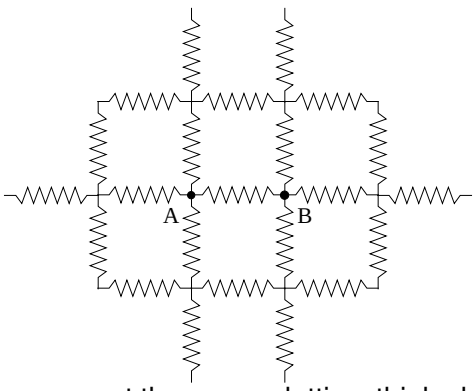
a) The equilibrium/final ($t = \infty$) charges on the two capacitors, Q_{1f} and Q_{2f}.

b) Using Kirchoff's laws for this arrangement, find the ordinary differential equation (ODE) for (say) $Q_1(t)$ and thereby the time constant for the equilibration process. Note that you do NOT have to solve the ODE, just formulate it with dt and some arrangement of R, C_1, and C_2 on the other side.

c) **All Students:** GUESS what the solution to the ODE looks like, based on your answers to a) and b). To do the latter, try visualizing what $Q_1(t)$ and $Q_2(t)$ will formally look like – it is just a matter of setting the various constants so that the asymptotic (final) and initial conditions are correctly represented and the approach to those conditions has the right time dependence.

d) **Advanced Students Only:** Solve the ODE (it is integrable, although a bit messy) for $Q_1(t)$ and $Q_2(t)$. It's probably best to solve for ***just one of the two***, and then use conservation of charge to find the other, right?

Advanced Problem 9.



Suppose you have an infinite network of identical resistors R, arranged in a square 2d lattice. Find the total resistance between two adjacent nodes as shown. Note well that ***there is a trick to this one***. Think about *current flowing into and out of this network through probes placed at junctions A and B shown* and superposition and symmetry.

Once you get the square lattice, think about *other* infinite lattices – for example infinite *triangular* lattices in 2d, or infinite *cubic* lattices in 3d. Can you just "write down" the total resistance of these two lattices? Are they the same or different? Why?

III: Magnetostatics

Week 6: Moving Charges and Magnetic Force

- A charge moving through space is observed to deflect according to the rule:

$$\vec{\boldsymbol{F}} = q(\vec{\boldsymbol{v}} \times \vec{\boldsymbol{B}}) \tag{6.1}$$

  which we use to *define* the magnetic field $\vec{\boldsymbol{B}}$ much as we defined the electric field in terms of the force observed and described by Coulomb's Law.

  For the moment we will ignore just how vB got there, as we live in a locally uniform magnetic field due to the Earth all the time and can discover magnetic materials in nature so natural sources of magnetism are ubiquitous.

- This translates into:

$$d\vec{\boldsymbol{F}} = I(d\vec{\boldsymbol{\ell}} \times \vec{\boldsymbol{B}}) \tag{6.2}$$

  for a small (differential) segment of wire carrying a current I in a magnetic field vB. Magnetic fields exert forces on current carrying wires.

- **Magnetic Forces Do No Work (on isolated charged non-spinning particles)!**

$$\vec{\boldsymbol{F}} = q(\vec{\boldsymbol{v}} \times \vec{\boldsymbol{B}}) \Longrightarrow P = \frac{dW}{dt} = \vec{\boldsymbol{F}} \cdot \vec{\boldsymbol{v}} = q(\vec{\boldsymbol{v}} \times \vec{\boldsymbol{B}}) \cdot \vec{\boldsymbol{v}} = 0$$

  is an *identity* of the cross product, so magnetic forces do no work on non-spinning charged particles.

- Motion of a point charge in the plane perpendicular to a uniform magnetic field is therefore *circular*:

$$|\vec{\boldsymbol{F}}| = qvB = \frac{mv^2}{r} \tag{6.3}$$

  (Newton's second law plus definition of centripetal acceleration). It has an angular velocity given by:

$$\omega_{\text{cyclotron}} = \frac{qB}{m} \tag{6.4}$$

  independent of its *speed*. This is called the *cyclotron frequency*.

- You should be able to derive/explain:
  - A cyclotron.
  - A velocity selector (region of crossed fields).

  - Thomson's apparatus for measuring $\frac{e}{m}$.
  - A mass spectrometer
  - The Hall effect (region of crossed fields in a conductor).

- The magnetic dipole moment of a plane current loop is:

$$\vec{\boldsymbol{m}} = NIA\hat{\boldsymbol{n}} \tag{6.5}$$

  where N is the number of turns, I is the current, A is the area, and $\hat{\boldsymbol{n}}$ is the right-handed normal to the plane of the loop.

- The *torque* on a magnetic dipole in a uniform magnetic field is:

$$\vec{\boldsymbol{\tau}} = \vec{\boldsymbol{m}} \times \vec{\boldsymbol{B}} \tag{6.6}$$

  Associated with this are its potential energy:

$$U = -\vec{\boldsymbol{m}} \cdot \vec{\boldsymbol{B}} \tag{6.7}$$

  and its force in a *non*-uniform magnetic field:

$$\vec{\boldsymbol{F}} = -\vec{\boldsymbol{\nabla}} U = \vec{\boldsymbol{\nabla}}(\vec{\boldsymbol{m}} \cdot \vec{\boldsymbol{B}}) \tag{6.8}$$

  Magnetic dipoles align with the field due to the torque, and then follow the field back to where it is stronger, just as do electric dipoles. Students have experienced this with toy magnets and refrigerator magnets from when they were very small – this is why bar magnets attract one another.

  You should be able to compute the magnetic moment of simple current loops, although we'll get more practice at this in the next chapter/week.

6.1: Magnetic Force versus Magnetic Field

In our discussions of the electrostatic force, we were able to start with a fundamental experimental result – Coulomb's Law – and proceed to systematically deduce nearly all of electrostatics including the more fundamental *expression* of Coulomb's Law: Gauss's Law for the Electric Field. Coulomb's Law *alone* told us *both* how to create an electric field *and* what the force was in terms of the field.

Life is not quite so simple for the magnetostatic field (where the "static" aspect refers to the field itself, not to the charges moving in or acting as sources of the field). In this and the next chapter we will learn that moving charges in a magnetic field experience a force according to a basic experimental rule (given a field) and moving charges in turn act as sources for a magnetic field (as one can experimentally verify by measuring forces). However, the *original* experiments, conducted by Ampere, that demonstrated both together involved *currents* and not *moving elementary charges*.

We, on the other hand, are interested in developing a "microscopic" description of fields that works for elementary point charges like electrons and quarks and that can be suitably

coarse-grain averaged into continuous distributions of charge and current (using the methods explored in the first part of the course). This suggests that we start with either force acting on *or* field produced by moving point charges and work our way *up* to Ampere's experimental results with current balances, instead of trying to work our way backwards.

For better or worse we will therefore begin with the force exerted by a magnetic field that we can think of as being *defined* by this force law, without (yet) worrying about where the field comes from. In the next chapter (next week), we will explore in great detail the sources of that field. *Do not hestitate*, however, to skip forward and backward between the two chapters as you study, as knowing at least the *summary* of the next chapter will help you with this one, just as you will certainly need to not instantly forget this chapter to move on and learn the next one. Together they ultimately produce a *single* view of the magnetic force between two moving charges and how it becomes the magnetic force between two currents.

6.2: Magnetic Force on a Moving Point Charge

With that said, let us proceed directly to the basic relation that *experimentally* describes the force exerted by a magnetic field on a charged particle. Note well that this force law can be more or less *directly* observed in a *Cloud Chamber*[87] placed in a magnetic field. Observations of many tracks (plus doing various current-based experiments) leads one to conclude that the force acting on a charged particle with charge q travelling at velocity $\vec{v}$ in a uniform magnetic field $\vec{B}$ is:

$$\vec{F} = q(\vec{v} \times \vec{B}) \tag{6.9}$$

Ooo! That pesky *cross product* rears its ugly[88] head! Sorry about that, but if you *don't* feel completely comfortable with a cross product yet, it is time to start really working on it. See the associated mathematical physics documentation linked to this course and start reviewing the good old right hand rule and the two or three ways available to compute them.

This law is (as you can see) *quite* different from the electrostatic rule, and the force depends on both the *magnitude* and *direction* of the *velocity* of the charge in the magnetic field, and doesn't point in the direction of the magnetic field at all! In fact, it points in the direction *perpendicular* to the plane determined by the magnetic field and the velocity vectors. Cross products are "twisty" beasts, always pointing off at right angles compared to any of the directions one might expect.

This twistiness, however, doesn't represent insoluble complexity, and you shouldn't throw your hands up in disgust or tremble in fear. As we will see, the motion produced by the

[87] Wikipedia: http://www.wikipedia.org/wiki/Cloud Chamber. Cloud chambers are actually quite easy to build, and I have had the building of an operational cloud chamber used for the extra credit/honors project my students often undertake. They are very cool – literally, as they are often cooled with e.g. dry ice or liquid nitrogen – and they *directly reveal to the eye* the tracks of otherwise invisible charged microscopic/elementary particles from the environment, from radioactive sources, from cosmic rays.

Just something to bear in mind if you are using this text in one of my classes with this third-of-a-letter-grade option!

[88] To introductory level students, at least. Actually, the cross product is amazingly *beautiful*, an essential part of a *geometric algebra* that generalizes the idea of complex variables to higher "grade" (number of complex dimensions). But to a student, "ugly" in this context is code for *more complicated* than the ordinary arithmetical multiplicative product or the scalar inner product between two vectors, and yet *essential to learn* in order to do well in the course!

magnetic force acting on a point charge is often quite *simple* and easy to understand and compute. To see this, we will begin at the beginning and solve for the motion in the simplest case, motion when the velocity is perpendicular to the (uniform) magnetic field.

One critical consequence of this form for the magnetic force law is that ***magnetic forces do no work on classical moving charged particles!*** We can easily see this by looking at the power delivered to a single charged particle by a magnetic field:

$$\vec{\boldsymbol{F}} = q(\vec{\boldsymbol{v}} \times \vec{\boldsymbol{B}}) \Longrightarrow P = \frac{dW}{dt} = \vec{\boldsymbol{F}} \cdot \vec{\boldsymbol{v}} = q(\vec{\boldsymbol{v}} \times \vec{\boldsymbol{B}}) \cdot \vec{\boldsymbol{v}} = 0$$

because (recall):

$$(\vec{\boldsymbol{A}} \times \vec{\boldsymbol{B}}) \cdot \vec{\boldsymbol{A}} = 0$$

(for any vectors $\vec{\boldsymbol{A}}$ and $\vec{\boldsymbol{B}}$ is an *identity* of the cross product. From the superposition principle, this must hold even in the coarse-grained limit where electric currents are made up of many moving charged particles.

Note well that when we say never, we *mean* ***never***, but that the statement is *qualified* by that "classical moving charged particles" bit. We will show later that if work is done on ordinary classical charged particles or currents in an electrodynamics problem, it is being done by the electric field, not the magnetic field. It may *look* like the magnetic field is doing work (and amazingly, that's how it works out algebraically) but for any arrangement of classical moving charges with no intrinsic magnetic moments the work is really done by electric fields instead.

Very shortly I will prove this statement for electric currents made up of the coarse-grained motion of ordinary charged particles through a conductor in a magnetic field. However, to avoid misleading you with a statement – however true it is in this *classical* physics course – that will get you into trouble when you hit quantum mechanics, I also need to qualify this statement with a "semi-classical" correction and tell you when magnetic fields *can* do (non-classical) work.

Elementary charged particles such as quarks and electrons (and very small composite particles like protons and neutrons or even atoms and molecules) often have an *intrinsic quantum mechanical magnetic moment* that does *not* arise from the motion of classical charged particles constrained by e.g. electrostatic or nuclear forces that hold matter together to remain within the medium the way electric currents are constrained to remain inside a conductor. Inhomogeneous magnetic fields can *indeed* do work on these non-classical point-like "intrinsic" magnetic dipoles! There is direct evidence of this – slow neutrons passed through an inhomogeneous magnetic field split into two beams in the neutron version of the *Stern-Gerlach Experiment*[89] . Since the neutron is electrically *neutral* and no electric fields are present, it is difficult to ascribe the work associated with this splitting to an electric field. However, the magnetic field associated with the intrinsic moment of electrons is also the source of work that pulls a refrigerator magnet towards a refrigerator, so you don't have to go far afield to see it proven!

Now let's move on to look at a variety of examples of using the Lorentz force law above (or just the magnetic part of it) to describe the motion of classical charged particles. As we will see in these examples, the magnetic field will never do any work on the particles, but sometimes *electric* fields will.

[89]Wikipedia: http://www.wikipedia.org/wiki/Stern-Gerlach experiment. This is a famous experiment that first demonstrated the existence of a magnetic moment due to intrinsic *quantum spin* for first electrons, then protons and neutrons.

Example 6.2.1: A Charged Particle Moving in a Uniform Magnetic Field

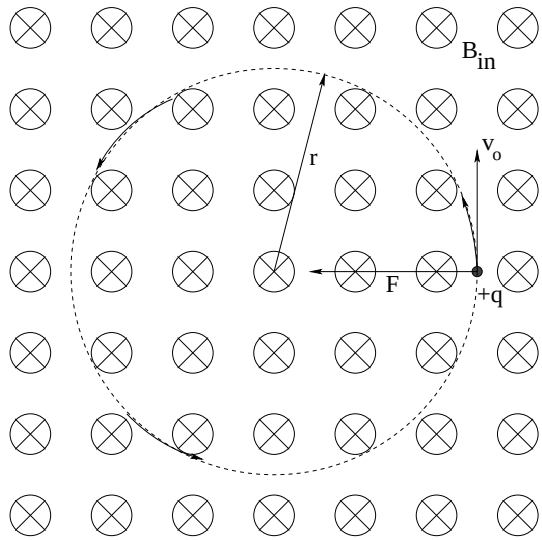


Figure 6.1: A charge particle with velocity perpendicular to a uniform magnetic field moves in a circle.

In figure(6.1 above, we see a charged particle $+q$ moving with initial velocity $\vec{v}_0$ perpendicular to a uniform magnetic field $\vec{B}_0$. The little crosses in this figure should be thought of as the "feather" ends of vector arrows and stand for a vector that points *into* the page – a circle with a dot will stand for the "tip" of the arrow and a vector pointing *out* of the page should we ever need it.

The force $\vec{F}$ acting on this charge is:

$$\vec{F} = q(\vec{v}_0 \times \vec{B}_0) \tag{6.10}$$

which has magnitude

$$F = qv_0B_0 \tag{6.11}$$

and which acts so that it is *always perpendicular* to the velocity of the particle! If you think back to your studies of *circular motion*, you should be able to easily see that this sort of force:

- Does no work. This in turn means that the *speed* of the particle is unchanged by the magnetic field.
- Acts to bend the particle's trajectory into a constant speed *circle*, with the magnetic field providing the necessary centripetal force.

That is:

$$F_r = qv_0B_0 = \frac{mv_0^2}{r} \tag{6.12}$$

We can, of course, solve this equation for any single unknown given the rest of the variables, but its most *common* use is to derive the so-called *cyclotron frequency* for the circulating particle:

$$\Omega_{\text{cyclotron}} = \frac{v_0}{r} = \frac{qB_0}{m} \tag{6.13}$$

Note well that this angular velocity/frequency *does not depend on the speed of the particle!* It is *fixed* by the charge of the particle, its mass, and the strength of the magnetic field *only*, which means that identical particles take the *same amount of time* to complete a circuit of their motion independent of their energy or velocity. This is the basis of the design of the *cyclotron*, one of the original particle accelerators (still) used to probe the structure of the atomic nucleus.

Example 6.2.2: The Cyclotron

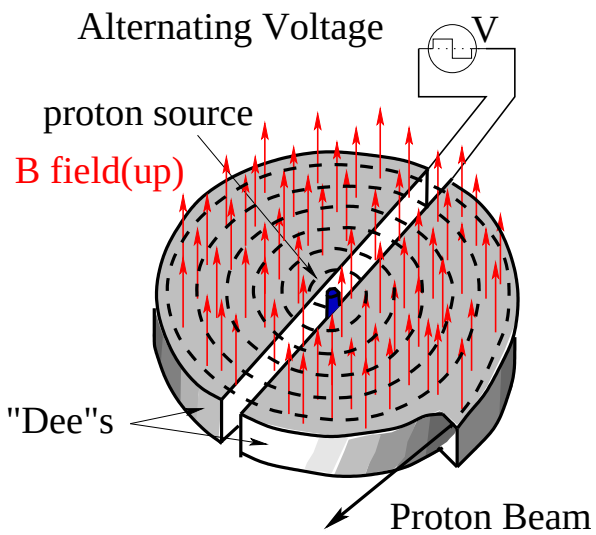


Figure 6.2: The schematic layout of a cyclotron. The electric field/potential difference between the "Dees" of the cyclotron oscillates with the same period as the period of the cyclotron frequency $\Omega_{\text{cyclotron}}$ of the particles moving in the field, so that it always pushes in the direction that speeds them up.

In figure 6.2 you can see the general design of a cyclotron. A suitable charged ion, e.g. a hydrogen nucleus (proton) is produced by e.g. an electrical arc in a source in the very center of the cyclotron with a low velocity. A powerful magnetic field bends the initial trajectory into a circular arc in the plane perpendicular to the field.

In between the upper and lower halves of the cyclotron are two copper chambers shaped like the letter "D", with a narrow slit in the plane perpendicular to the field cut along the straight segment in the middle. An alternating electric potential is applied *between* these two "Dees" that has the *same angular frequency as the cyclotron frequency of the particle being accelerated in the magnetic field in question* so that when the particle arrives at the gap between the upper and lower Dee in the figure above, it happens to point down (and hence speeds the particle up). When the particle gets to the gap between the lower and the upper Dee on the right, though, the field has *switched direction* and still *speeds the particle up* still more. Every time the particle arrives at the gap, it finds the field is there, aligned with its motion to give it yet another push.

This works because it takes *all* of the particles the same amount of time to make it around a half-circle regardless of how fast they are going. So one can have a stream of particles all falling across the gap at once at different radii from the source (with short gaps between these

”pulses” that are in phase and being accelerated together). As the particle is moving faster and faster, the radius of the circle of its motion increases until it reaches an electrostatic deflector plate at the outside edge of the magnetic field that angles it into a beam pipe where it travels through a vacuum to hit an eventual beam target.

Early cyclotrons played an important role in the development of nuclear physics, permitting the creation and discovery of the first transuranic elements past plutonium (one of which is named Lawrencium, after the inventor of the cyclotron, another of which is named Berkelium after the University where Lawrence worked).

Cyclotrons, alas, no longer work when the particles are accelerated enough to be moving at relativistic velocities. At some point the time dilation of the cyclotron period in the frame of the moving particle is enough to keep the particle from being accelerated by a Dee voltage at the cyclotron frequency that worked for a slowly moving particle. One can “fix” this problem by sweeping the frequency to match and acclerating only pulses of charge (in a *synchrotron*) but as one reaches higher and higher energies other problems emerge.

The principal limiting factor is ultimately the fact that *accelerated charges radiate*, and particles moving in circles *are accelerated all of the time* by the centripetal magnetic force. This causes a kind of “resistance” wherein the work done speeding the particle up in a cycle is balanced by radiative losses in the cycle. Only the use of very large circles can minimize the latter, which is why the extreme relativistic accelerators of modern times, such as the Large Hadron Collider (LHC) are enormous circles, the latter being *27 kilometers* in circumference.

Example 6.2.3: Cloud Chamber

In a nuclear collision, a lot of “stuff” is produced – nucleons knocked out of nuclei, electrons, positrons, gamma rays, alpha particles, and more exotic particles that help us understand the nuclear field itself. To be able to categorize and classify all of this “stuff”, it helps to be able to “see” the trajectory of a particle produced in the collision, and determine things like the ratio of its charge to its mass. A cloud chamber (and more exotic bubble chambers that work on a similar principle) is a device that makes a charged particle’s trajectory visible so that it can be photographed. It works by creating a “supersaturated” gas of e.g. alcohol, water vapor, or other substances. The charged particle in question zips through the vapor and causes it to bounce together in its wake, precipitating the vapor out as a condensation trial, much like the jet contrails one can sometimes see overhead on a clear day. In a cloud chamber the trajectories typically only last a few seconds before re-evaporating, but that is long enough to be easily seen and/or photographed for later analysis.

By putting the chamber in a *magnetic field* and right next to a nuclear target, the *positive* particles curve one way and the *negative* particles curve the other. The radius of curvature is related to the charge and mass by:

$$r = \frac{mv_0}{qB_0} \tag{6.14}$$

and can be determined directly from a photographed trajectory.

At the same time, the particle slows down because the same process that causes supersaturated gas molecules to precipitate out along its trajectory exerts a “drag” force on the particle. By looking at the rate the particle’s trajectory curvature *changes* (and various other

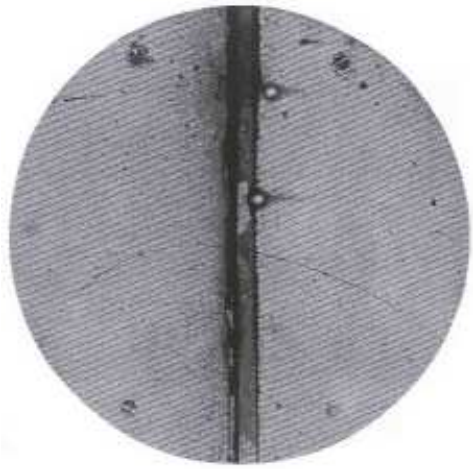


Figure 6.3: The first photograph of a positron ever taken in a cloud chamber. Note the curvature carefully. Which way is the particle travelling while slowing down? What direction does the magnetic field in the chamber point?

things), one can estimate its momentum, the charge of the particle, and its mass. Using this and many other specialized detectors, an enormous "zoo" of particles has been discovered and categorized and transformed into a quantitative model for the nuclear force that has at least some predictive power, although it is not yet a complete or perfect theory.

A simple cloud chamber is not too difficult to build – it requires a bowl, dry ice, alcohol, cotton, black paint, a light source, and a few other things, but they are all fairly readily obtainable. It is therefore a good candidate for an extra credit project, if your program has one.

Example 6.2.4: Region of Crossed Fields

Another extremely useful application of magnetic fields acting on individual charged particles is the *region of crossed fields*. A region of crossed electric and magnetic fields, when equipped with suitable collimating slits, can act as a *velocity selector*.

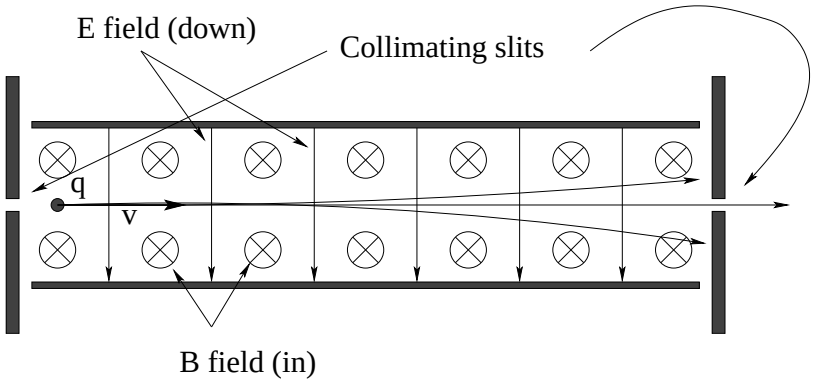


Figure 6.4: A region of crossed fields functions as a velocity selector; only particles with just the right velocity pass through undeflected.

A charged particle with charge q enters the device on the left by passing through collimating slits that ensure that its velocity is in the x-direction only. Inside the device a pair of parallel plates creates a uniform electric field $\vec{E}$ down, while a magnetic coil creates a uniform magnetic field $\vec{B}$ into the page as drawn.

From the right hand rule, the magnetic force on the charged particle is

$$F_B = qvB \tag{6.15}$$

up. The electric force, however, is

$$F_E = qE \tag{6.16}$$

down. The net force on the particle is *zero* when:

$$F_B = qvB = qE = F_E \tag{6.17}$$

or when the particle happens to have the velocity

$$v = \frac{E}{B} \tag{6.18}$$

in the x-direction. In this case the particle travels through undeflected and exits through the collimating slit on the right.

Particles that are travelling too *fast*, however, have a magnetic force that exceeds the electric force and are deflected *up*. They strike the barrier at the far end and fail to pass through the slit. Similarly particles that are travelling too slowly have an electric force that exceeds the magnetic force. They are deflected *down* and fail to make it through the second slit.

Note well that this is a *velocity* selector and passes all particles with the right velocity *regardless of their mass or their nonzero charge!* The particle can have any charge, positive or negative (except zero), or any mass – as long as it has the right *velocity* it will still make it through undeflected. This makes it very useful for preparing particle beams for certain kinds of experiments. It is also *very closely related* to the *Hall Effect* described later.

Example 6.2.5: Thomson's Apparatus for measuring e/m

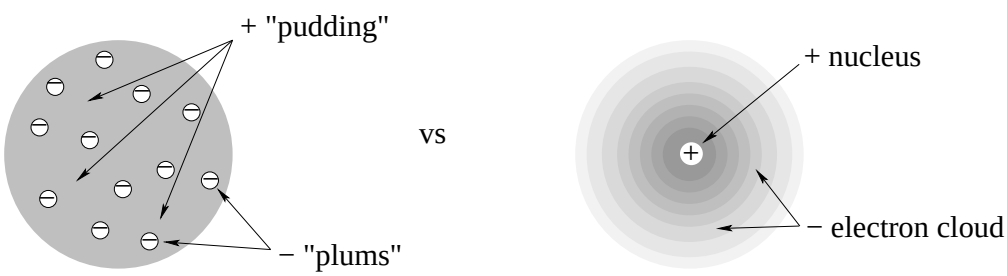


Figure 6.5: The "plum pudding model" that prevailed in 1897 on left, along with a more accurate representation of the current atomic model – a massive nucleus surrounded by a quantum "pudding" (the electron cloud).

The year is 1897. People know that matter is made up of atoms, that atoms are made up of positive and negative charge, but the human species *still does not know* if the positive

and negative charges are themselves *particles*, and if so, what the charges and masses of those particles are. There are a variety of models for atoms, most of them "static" models that have negative and positive charge glued together in some way that keeps the negative and positive charge from having to *orbit* one another the way the electrostatic force suggests that they should, as James Clerk Maxwell has shown that classical atoms made up of orbiting charged particles would radiate all of their energy away in a very, very short time and collapse. One of the favorite models is in fact called the "plum pudding model" portrayed in figure 6.5 (no kidding!) where negative charge is scattered like raisins in a gooey pudding of positive charge.

The so-called "cathode ray tube" (or Crooke's tube) has been invented for twenty or thirty years, and a mere two years earlier a gentleman named Röntgen discovered that cathode rays hitting the glass of the screen at high enough energies produce *x-rays*, capable of penetrating the human hand and forming images of the bones within (see figure 6.6 above) for which he received the *first* Nobel Prize in physics in 1901.

Figure 6.6: The first "medical x-ray" ever taken, of the bones in Anna Berthe Röntgen's hand. She was the wife of Wilhelm Röntgen, the discoverer of x-rays.

The question is: Just what *are* cathode rays? Are they particles? Do they have arbitrary amounts of charge and mass? Are they a fixed fraction of the mass of e.g. a hydrogen atom? Is the mass of a hydrogen atom split evenly between cathode (negatively charged) material and anode (positively charged) material? J. J. Thomson set out to try to answer these questions by using a specially modified Crooke's tube to deflect cathode rays *in flight* once they were produced at a heated electrical filament and accelerated by an applied potential difference so that they formed a beam.

Initially the deflection was accomplished only by the application of an electric field in between special plates built right into the tube (which was sufficient, as we shall see, to measure the ratio of e/m for cathode ray particles or *electrons* (as they turned out to be) and thereby show that they were a *tiny fraction* of the total mass of a hydrogen atom, so that nearly all of the mass was associated with the *positive* charge only. Later Thomson added a uniform magnetic field to his apparatus by means of a pair of "Helmholtz coils". As we have seen above, magnetic fields can *also* deflect moving charged particles, and indeed if a region of crossed fields is created, the $\vec{E}$ and $\vec{B}$ fields together can be used to measure the actual velocity of the

particles, which permits their kinetic energy and/or mass to be estimated and the consistency of all of the (many, not too accurate yet) measurements to be checked.

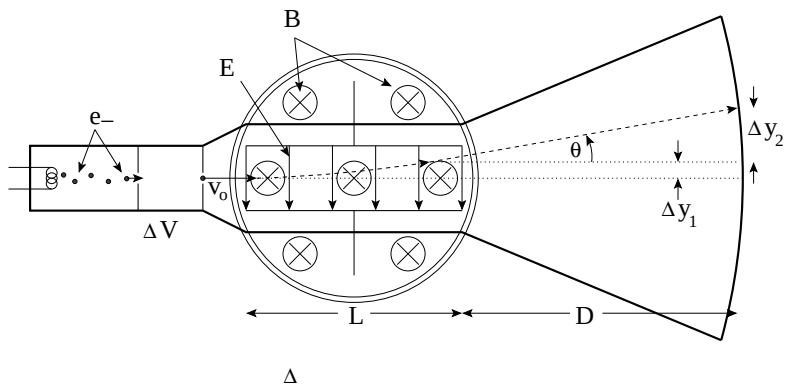


Figure 6.7: Joseph John Thomson's apparatus for measuring the ratio of the charge on the electron to its mass (improved by the addition of a magnetic velocity selector). This was a critical experiment in determining the structure of the atom, for which Thomson received the Nobel Prize (only the sixth such prize awarded in physics).

A cartoon schematic of Thomson's apparatus is shown above, although I had a very hard time finding any useful picture of how he applied the magnetic field in his second series of experiments and the magnetic part of the apparatus may be incorrect.

Let's see how Thomson used his apparatus to measure e/m for the electron. First, he cooked up some electrons using a wire heated by joule heating until electrons "boiled off". These slow electrons passed through collimating slits and fell across a potential difference maintained between the plates containing the slits to speed them up to a roughly consistent velocity. The electrons, each with (approximate) velocity v_0 then entered the region between to capacitor plates built right into the tube. The downward electric field then produced an upward, constant upward acceleration and hence deflection of the electrons (where we can completely ignore gravity in the experiment as the electrical acceleration was vastly greater) as they traversed the plate length L. On the far side they emerged from the field, travelled in a straight line for an x distance of D, and then struck the glass of the screen, where they made a glowing spot.

By measuring the *total distance* of upward deflection of the spot from the center of the screen (where they struck when the E-field was off) and the point where they struck when the E-field was turned on to some known value, Thomson could reason backwards to the ratio of e/m as follows.

First, we analyze the constant acceleration motion of the electron while it is between the plates:

$$F_x = 0 \tag{6.19}$$

$$F_y = eE = ma_y \tag{6.20}$$

from which we find (using 2D kinematics from the first semester – the problem is *identical* to

analyzing trajectories with a constant gravitational acceleration):

$$x(t) = v_0 t \tag{6.21}$$
$$v_x(t) = v_0 \tag{6.22}$$
$$y(t) = \frac{1}{2}a_y t^2 = \frac{eE}{2m}t^2 \tag{6.23}$$
$$v_y(t) = a_y t = \frac{eE}{m}t \tag{6.24}$$

We can easily find the *time* the electron is between the plates:

$$t_1 = \frac{L}{v_0} \tag{6.25}$$

from $x(t)$. Substituting this into the last two equations, we find that as it emerges from between the plates:

$$\Delta y_1 = \frac{eE}{2m}t_1^2 = \frac{eEL^2}{2mv_0^2} \tag{6.26}$$

and

$$v_y = \frac{eE}{m}t_1 = \frac{eEL}{mv_0} \tag{6.27}$$

From our knowledge of v_x and v_y when the particle emerges, we can find:

$$\tan(\theta) = \frac{v_y}{v_x} = \frac{eEL}{mv_0^2} \tag{6.28}$$

which lets us easily determine:

$$\Delta y_2 = D\tan(\theta) = \frac{eELD}{mv_0^2}. \tag{6.29}$$

Now we can relate the *measured* total y deflection to the known values of L, D, E, and our *estimated* v_0:

$$\begin{aligned} y_{\text{tot}} &= \Delta y_1 + \Delta y_2 \\ &= \frac{eEL^2}{2mv_0^2} + \frac{eELD}{mv_0^2} \\ &= \frac{eEL}{mv_0^2}(\frac{L}{2} + D) \\ &= \frac{e}{m}\frac{EL}{v_0^2}(\frac{L}{2} + D) \end{aligned} \tag{6.30}$$

Inverting this last relation we find:

$$\frac{e}{m} = \frac{y_{\text{tot}}v_0^2}{EL(\frac{L}{2} + D)} \tag{6.31}$$

We know everything on the right (where we *measure* y_{tot}), so we have measured e/m!

Of course Thomson didn't really know v_0 – he had to *estimate* it from a mix of thermodynamics and electrostatics in his first experiment. We, however, can see how the addition of a crossed magnetic field permits him to *precisely determine* v_0. With the E field turned

on, simply turn up the magnetic field B until the particle's deflection is once again *zero*. At that point, the apparatus is functioning as a velocity selector, and we know from the argument above that:

$$v_0 = \frac{E}{B} \tag{6.32}$$

this can be substituted into the expression above to obtain a much more accurate estimate for e/m, one that doesn't rely on a prior knowledge of the thermal distribution of electron energies before they are accelerated by the first potential difference:

$$\frac{e}{m} = \frac{y_{\rm tot}\left(\frac{E}{B}\right)^2}{EL(\frac{L}{2}+D)} = \frac{y_{\rm tot}E}{B^2L(\frac{L}{2}+D)} \tag{6.33}$$

This permits a measurement that is as accurate as one's knowledge of $y_{\rm tot}$ L, D, E and B; with care within a few percent even using late 19th, early 20th century apparatus. Using it Thomson was able to determine that the *relative* mass of the negative charge in a hydrogen atom compared to the mass of the positive charge was *less than 0.1%!* The electron was *extremely light* compared to the proton.

Of course, Thomson still did not know that the proton existed; the plum pudding model described the positive mass as being an "amoebic blob" that somehow bound the electron to the atom. It wasn't until Rutherford did his famous experiment a few years later that scattered alpha particles (helium nuclei) from gold foil and observed that many of the alpha particles scattered *straight back*, something that they could only do if the positive charge was tiny and extremely massive, that it became clear that the nucleus really *was* a proton, a tiny massive charge at the center of the hydrogen atom, with some 1872 times the mass of the electron.

This, in turn, spelled the death of classical physics. Plum pudding was spoiled forever. This was no great loss; it couldn't explain e.g. the spectral lines visible in light emitted by superheated hydrogen gas. However, the alternative was now a return to the classical orbital model with electrons orbiting protons the same way a planet orbits the sun, in elliptical orbits wherein the electron is constantly accelerating. Maxwell's equations had long since proven that such an atom would instantly collapse, radiating away electromagnetic energy in *all* frequencies as it did so, not in some subset of discrete frequencies. Thomson's experiment, simple as it is to us today in terms of our modern models and knowledge of electromagnetism, truly deserved the Nobel Prize because it paved the way in a critical way for the invention of quantum mechanics and our current understanding of atomic structure.

Example 6.2.6: The Mass Spectrometer

Another use of magnetism is in the construction of a *mass spectrometer*. A mass spectrometer is a device that takes some chemical "goop", perhaps a sample created in a lab, perhaps a sample obtained through some forensic process, and measures the masses of its chemically stable components.

In the schematic above, a cooker heats some unknown chemical "goop". This vaporizes it, and the vapor comes in contact with a high voltage source that ionizes it. Ions of mass m_1, m_2, m_3... and (associated) charge q_1, q_2, q_3... are then accelerated by a potential difference to an energy (respectively) of $q_1V_0, q_2V_0, q_3V_0.. >$, passed through a pair of collimating slits as

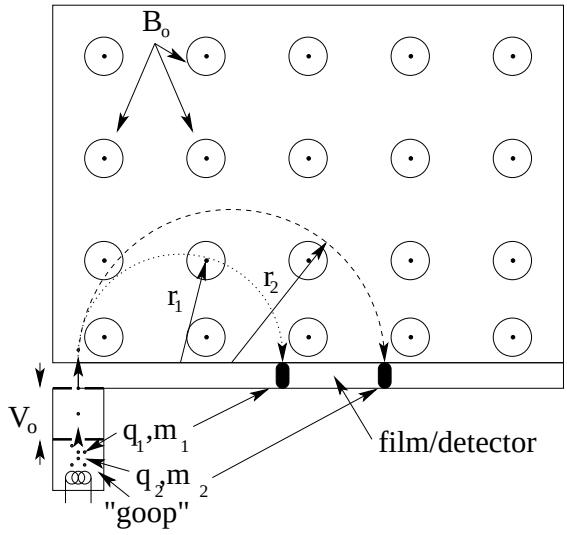


Figure 6.8: The Mass Spectrometer uses a region with a uniform magnetic field to create a *spectrum* of particles that collide with a film or other detector matrix in places that indicate the radius of the circle they are bent in by the field. This, in turn, is related to the ratio of q/m for the particle, and by assuming a charge that is a low integral multiple of e one can determine the mass.

a beam, and then piped into a region containing a uniform magnetic field B_0 (out of the page as drawn). Positive ions (for example) are then bent into circular trajectories depending on their mass, charge, and entrance energy/velocity. The ions impact on a detector of some sort – perhaps a piece of photographic film – where each particular q, m combination registers as a *distinct* signal a distance $2r$ from its entrance point (where r is the radius of curvature of the species' particular trajectory).

The molecular weight of the components of the sample is thus registered two ways. Typically a "marker" species of known weight and concentration is introduced that permits the distances from the entrance point to be calibrated and checked against a known mass, and each particular components is likely to be present in single ionized form (with charge e.g. $+e$), doubly ionized form (with charge $+2e$) etc. This appears as "similar" patterns of bands on the film or detector which permits one to tell which pattern corresponds to a particular charge, for example $+e$. From this combination it is straightforward to deduce the charge and infer the mass of the various chemical components visible in the detector fingerprint.

We can easily understand the physics behind the mass spectrometer. A charged ion of charge q and mass m produced in the goop boiler is accelerated to a kinetic energy:

$$\frac{1}{2}mv^2 = qV_0 \tag{6.34}$$

in the beam entering the magnetic field. It therefore has a velocity[90]:

$$v = \sqrt{\frac{2qV_0}{m}} \tag{6.35}$$

[90] As before in the case of the Thompson apparatus, in reality the "boiler" would produce a Maxwell-Boltzmann range of entering velocities, but we can insert a velocity selector stage to narrow the distribution to "precisely" the desired/expected v.

and experiences a centripetal magnetic force (that causes it to move in a circle of radius r) of:

$$F_r = qvB_0 = m\frac{v^2}{r} \tag{6.36}$$

so as usual:

$$\frac{v}{r} = \frac{q}{m}B_0 \tag{6.37}$$

If we solve for the radius r of its half-orbit to the film/detector, we get:

$$r = \frac{v}{B_0}\frac{m}{q} \tag{6.38}$$

Substituting for v:

$$r = \frac{\sqrt{\frac{2qV_0}{m}}}{B_0}\frac{m}{q} = \sqrt{\frac{2mV_0}{qB_0^2}} = \sqrt{\frac{m}{q}}\frac{\sqrt{2V_0}}{B_0} \tag{6.39}$$

Alternatively, since one measures r and wishes to find m (given a good guess for q):

$$m = \frac{r^2B_0^2}{2V_0}q \tag{6.40}$$

As one can see, the mass-to-charge ratio determines r, creating similar "bands" of molecular signal for different ionizations of the same collection of constituent masses. Once the charge on any given band is guessed/determined (where the lowest charge, in positive multiples of e, will have the largest radius spectral pattern for each set of m's) one can transform a knowledge of r and q directly into m.

Most of this process can be automated and computerized, and mass spectrometers based on this general principle are at this point commonplace in the laborator.

Example 6.2.7: The Hall Effect

The final object of our study of the magnetic force on single charged particles is the *Hall effect*, the tendency of a current carrying wire in a magnetic field to build up a voltage *across* the wire, or conducting strip that is based on spontaneous charge separation in the conductor to create a "region of crossed fields" where the electric field/force precisely balances the magnetic force (and simultaneously creates a potential difference).

The Hall Effect is a phenomenon that spontaneously occurs when a conductor carrying a current is placed in a magnetic field that is perpendicular to the current. The effect is easiest to observe in a ribbon shaped conductor that is relatively wide; one such is pictured above with width w (top to bottom) and cross-sectional area A.

The Hall Effect can be used to make two very important classical measurements. First, as we will easily see, we can finally determine the *sign of the charge carriers* in any given material, as positive charge carriers (the particles that are physically moving to create the current) will actually polarize the strip the *opposite way* than negative ones. Second, it enables us to directly measure n, the density of charge carriers in our basic model of conduction.

Here's how it works. The strip is placed into a magnetic field perpendicular to the strip as shown and a current is run through it. In the figure 6.9 above, we assume *positive* charge

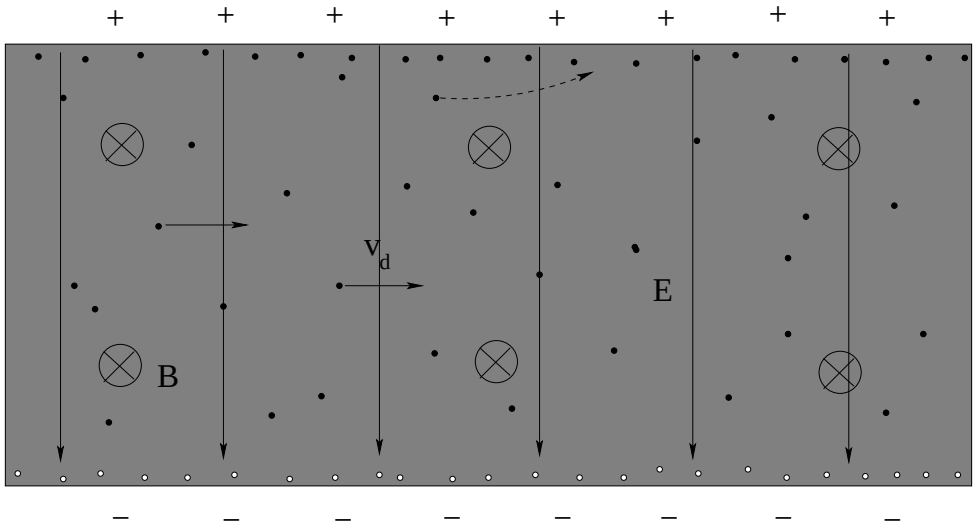


Figure 6.9: In the Hall Effect, a magnetic field causes the *mobile* charge to accumulate on the upper or lower edge of a conducting, current-carrying strip in a magnetic field. This in turn creates a potential difference across the strip that can easily be measured.

carriers as usual so that the current is in the *same* direction as the drift velocity of the carriers, from left to right.

At first these moving charges experience a magnetic force that (right hand rule!) diverts them into a curved trajectory to the *left* as indicated by the dashed arrow on one of the charges. However, charges near the top have nowhere to go and *build up* in a layer on the upper surface of the strip. This charge layer creates an electric field that begins to oppose the motion of still more charge until after a bit, the strip has equal and opposite amounts of positive (upper) and negative (lower) charge on the top and bottom edges, the latter in the form of "holes" left from which the positive charge carriers migrated.

The charges now move in a *spontaneous region of crossed fields* – the carriers in the middle move in zero net force with the electric force down equal to the magnetic force up. This, in turn, creates an electrical *potential difference* V across the strip that can be measured with a voltmeter, at the same time that the current through the strip I is measured with an ammeter.

We know that for each charge, when this situation is established:

$$qv_dB = qE \tag{6.41}$$

or

$$v_d = \frac{E}{B} \tag{6.42}$$

We also know that:

$$I = nqv_dA = nqA\frac{E}{B} \tag{6.43}$$

Finally, we known that

$$V = Ew \tag{6.44}$$

or

$$E = \frac{V}{w} \tag{6.45}$$

so that

$$I = nqA\frac{V}{Bw} \tag{6.46}$$

We can then solve for n, the desired density of charge carriers:

$$n = \frac{IBw}{qAV} \tag{6.47}$$

One can measure I and V directly. B one can compute (although the Hall effect is actually often used to *measure* B, as one can obviously turn this equation around and solve for B with a strip made from a material with *known* n). w and A can be directly measured with a ruler.

Best of all, we can finally see that the charge carriers in most metals are *electrons*, that is, they are *negative*. Suppose that the carriers in the picture above *were* electrons and negative. Then with a current travelling to the right, they would actually be moving to the left. The magnetic field would then still divert them *up*, creating a *negative* strip of charge on the upper edge of the strip and a positive one on the lower. The electric field – for the same left-to-right current – would run from the bottom to the top when the desired region of crossed fields established itself. This would make the top of the strip at a *lower* potential than the bottom, the opposite of what one gets with a positive charge carrier.

Franklin's Mistake is thus finally laid bare. Alas, the mobile charge in most conductors is made up of negatively charged electrons, the "cathode ray" particles discovered by Thomson. This is not always the case, of course. Ionic fluid solutions (like salt water) can have currents in which *both* charge carriers are present. Also, in semiconductors the carriers can easily be quantum mechanical "holes" in the electron density that have an effective positive sign.

As we can see, the magnetic force on discrete particles is a very useful thing! This by no means exhausts the utility of magnetic fields for bending streams of charged particles around to make them do our bidding.

In the last example, though, we went from a picture of single charges to one where we were working with the coarse-grained continuum limit of a charged *current* once again. Perhaps it is time to think about the magnetic force on current carrying wires!

6.3: The Magnetic Force on Continuous Currents

If we contemplate our (by now) standard model for current in a uniform wire, where the current I is given by:

$$I = nqv_dA = \int \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA \tag{6.48}$$

(where, recall, n is the density of charge carriers, q is the charge per carrier, v_d is the "drift velocity" – the average velocity of the carriers in the wire, A is the wire's cross-sectional area) then we can add up the magnetic forces on all of the charges in a short (differential) length of wire $d\ell$:

$$d\vec{\boldsymbol{F}} = nq(Ad\ell)\vec{\boldsymbol{v}}_d \times \vec{\boldsymbol{B}} \tag{6.49}$$

We now do a clever thing. We'll collect the nqv_dA *magnitudes* together and make I, and take the *direction* of $\vec{\boldsymbol{v}}_d$ and attach it to $d\ell$, making it a *vector* pointing in the direction of the current

in the wire. The result is:

$$d\vec{\boldsymbol{F}} = I(d\vec{\boldsymbol{\ell}} \times \vec{\boldsymbol{B}}) \tag{6.50}$$

for a small (differential) segment of wire carrying a current I in a magnetic field vB. Magnetic fields exert forces on current carrying wires!

To *evaluate* the total force on any given current carrying wire is not, of course, likely to be *easy* unless the wire has a very nice geometry, such as being a *straight line* in a uniform field or a *circular loop of current* in a uniform field. However, we can prove a very interesting result for *arbitrary* current loops that lets us understand how magnetic forces work on them to at least a decent approximation, especially when those loops are "small" relative to everything else that is going on. Let's procede.

Example 6.3.1: The Magnetic Force and Torque on a Rectangular Current Loop (Magnetic Dipole)

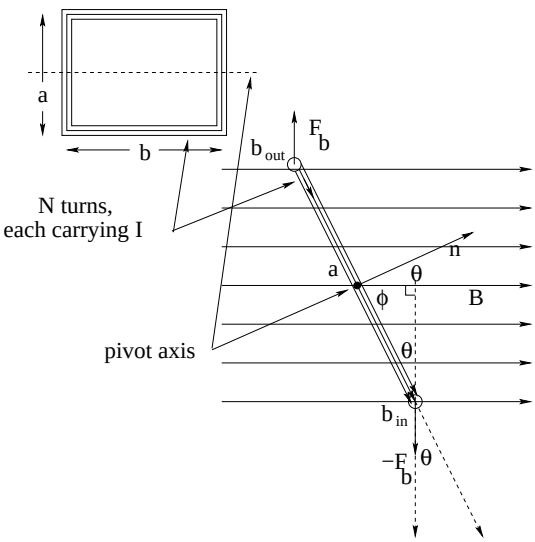


Figure 6.10: The force and torque on an $a \times b$ rectangular loop of N turns, each carrying current I, in a uniform magnetic field $\vec{\boldsymbol{B}}$ are $\vec{\boldsymbol{F}} = 0$ and $\vec{\boldsymbol{\tau}} = \vec{\boldsymbol{m}} \times \vec{\boldsymbol{B}}$ respectively.

In figure 6.10 you can see pictured a rectangular current loop with N turns, each carrying a current I. When studying electrical currents and magnetic fields, using loops with many turns is a cheap and easy way to get a larger current than one's power source can ordinarily support, as this is effectively a current of NI on each leg of the circuit. If you've ever looked inside an electrical motor, or transformer, or generator, or electronic device, you'll almost certainly see loops of reddish (epoxy or enamel insulated) copper wire wrapped into loops with many turns for just this reason.

The dimensions of this particular loop are a and b, although in the next section we'll see that these particular dimensions, and indeed the shape of the plane loop, are not terribly important. I put the loop in the "inset" to the upper left so you can visualize what it might look like lying on a table. Note that we'll imagine that the loop has an "axle" on which it can freely pivot. This too isn't strictly necessary (we can pick other piviots that will work just as well or better) but

guessing that your recollection of torque is still a bit shaky it won't hurt to draw in a simple one that is easy to understand.

In the main part of the figure I've drawn an "edge view" of the loop as it sits in a *uniform magnetic field* $\vec{\boldsymbol{B}}$ pointing to the right. The "uniform" bit is very important – we obviously would get a very different result for the force if (for example) the field on the upper b side were larger than the field on the lower b side!

Evaluating the force on each of the four sides of the rectangle is trivial. The upper and lower b sides are perpendicular to the $\vec{\boldsymbol{B}}$ field, have length b, have N turns each carrying I, and hence the magnitude of the force is:

$$F_b = NIbB \tag{6.51}$$

We can find the direction easily using the right hand rule. It is up on upper side (with current pointing out of the page) and down on the lower side.

The force on the a sides is hardly more difficult. Let's consider the one closest to us in the figure, with the current slanting down and to the right. The directed current makes an angle of ϕ with the magnetic field, so the force on it is:

$$F_a = |NI\vec{\boldsymbol{a}} \times \vec{\boldsymbol{B}}| = NIaB\sin(\phi) \tag{6.52}$$

with a direction (right hand rule again) of out of the page. The hidden a side on the other side (where the current slants up and to the left) has the same magnitude force and the opposite direction.

The sum of these forces is this clearly

$$\vec{\boldsymbol{F}}_{\rm tot} = (F_b - F_b)\hat{\boldsymbol{y}} + (F_a - F_a)\hat{\boldsymbol{z}} = 0 \tag{6.53}$$

where I've fairly arbitrarily popped a coordinate system onto the picture with x to the right, y up, and z out of the page.

Does this ($\vec{\boldsymbol{F}} = 0$) mean that nothing interesting happens to the loop in the field? Not at all! The two F_a forces are indeed uninteresting, as they act along the same line (along the axle, in fact) and exert neither force nor torgue on the system. The two F_b forces, however, do *not* act along the same line. They exert a *torque* on the loop!

How large a torque? Recalling that $\vec{\boldsymbol{\tau}} = \vec{\boldsymbol{r}} \times \vec{\boldsymbol{F}}$ where $\vec{\boldsymbol{r}}$ is a vector from the pivot to the force, the torque from the upper b side using the pivot shown (so that $r = a/2$) is:

$$\tau_b = \frac{a}{2}F_b\sin(\theta) \tag{6.54}$$

and points *in* to the page. The torque from the lower b side is identical in magnitude and *has the same direction* (into the page). The total torque thus has magnitude:

$$\tau = aF_b\sin(\theta) = NI(ab)B\sin(\theta) \tag{6.55}$$

into the page.

Now take a moment to look carefully at the geometry of this figure. The angle θ we used is the one between the direction of $a/2$ in each case and F_b. I've drawn this angle in for the

lower side to make it easy to see, but it is the same for the upper side too. If you follow θ from the angle in between to the angle in the right triangle with the dashed side, use $\phi = \pi/2 - \theta$, you can see that the angle between the *right handed normal to the plane loop* $\hat{\boldsymbol{n}}$ drawn and the magnetic field will *always* be the very θ that we want. The right handed normal is the unit vector perpendicular to the plane of the loop that points in the direction your right hand thumb points when your fingers curl around the loop in the direction of the current.

This (and the highly suggestive form of τ) suggests that we define the *magnetic dipole moment* of this loop to be:

$$\vec{\boldsymbol{m}} = NI(ab)\hat{\boldsymbol{n}} \tag{6.56}$$

in which case the torque takes the familiar form:

$$\vec{\boldsymbol{\tau}} = \vec{\boldsymbol{m}} \times \vec{\boldsymbol{B}} \tag{6.57}$$

which looks *just like* the torque on an *electric* dipole, $\vec{\boldsymbol{\tau}} = \vec{\boldsymbol{p}} \times \vec{\boldsymbol{E}}$! In fact, since the force on an electric dipole also vanished in a uniform field, we can *instantly adopt* (reasoning by algebraic analogy or formally rederiving it all as we prefer) all of the results in table 3 below. But first, let's generalize our expression for the magnetic dipole moment a bit and consider a more general plane current loop instead of just a rectangle.

Example 6.3.2: The Magnetic Moment of an *Arbitrary* Plane Current Loop

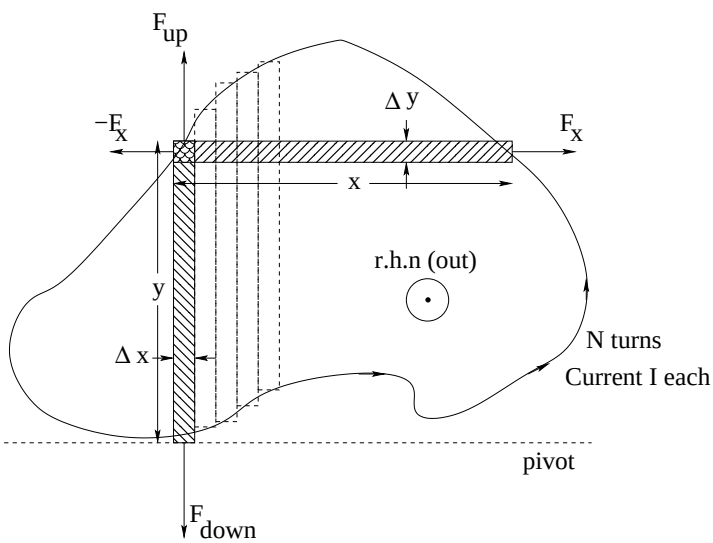


Figure 6.11: Arbitrary plane loop of current can be broken into small pieces that are aligned with or perpendicular to torque axis.

In figure 6.11 we see a golf-putting-green shaped loop of current carrying wires in a plane. As before, there are N turns carrying a current I, and I've drawn an arbitrary rotation axis/pivot that is perpendicular to the $\vec{\boldsymbol{B}}$ field that the loop will be in and located at the end of the (each) loop rectangle for convenience.

As you can see, one can take the curve and break it up into perpendicular segments that approximate the curve arbitrarily closely as the Δx and Δy segments are made smaller and smaller. If one considers just *one* such opposing pair of segments each (the shaded/textured

areas in the figure), the forces F_x between the Δy parts of the curve are equal and opposite and along a common line parallel to the axis of torque. They contribute no force and no torque in a uniform field so we don't even bother to sum over them, we just ignore them.

The forces between the Δx parts of the curve (the direction that would have been into or out of the page in the rectangular figure above) are *also* equal and opposite, but they are typically offset so that they do not act along a common line but rather one with a perpendicular displacement of $y\sin(\theta)$, where θ is the angle between the $\vec{\boldsymbol{B}}$ field and a right handed normal to the figure. y (for this small segment of current) thus acts like the a coordinate in the rectangular figure above, Δx acts like a very short piece of the b segment. This pair of forces *does* contribute a net torque (magnitude) for just this little strip of the total wire of:

$$\Delta\tau = y\Delta x NIB\sin(\theta) \tag{6.58}$$

Summing over all of the strips of width Δx, the total torque on this plane loop is thus:

$$\tau = NI\lim_{\Delta x\to 0}\left(\sum y(x)\Delta x\right)B\sin(\theta) = NI\left(\int y(x)dx\right)B\sin(\theta) = NIAB\sin(\theta) \tag{6.59}$$

or (including the vector direction from the right-hand-rule applied both to the torque and the right handed normal to the loop):

$$\vec{\boldsymbol{\tau}} = \vec{\boldsymbol{m}}\times\vec{\boldsymbol{B}} \tag{6.60}$$

with

$$\vec{\boldsymbol{m}} = NIA\hat{\boldsymbol{n}} \tag{6.61}$$

We see that our rule for the rectangular loop above is thus *general* and applies to any plane loop of current, no matter what the shape.

6.4: Potential Energy of a Magnetic Dipole

As before with electric dipoles, we must do ***work*** rotating a magnetic moment from one angle to another in a magnetic field, working against the torque. The work *we* do to rotate the dipole equals the potential energy stored in the system (the magnetic dipole and field combined). We can compute this potential energy by following the derivation we used for electric dipoles, using as before a zero of the potential energy when the dipole is at right-angles to the magnetic field. That is (given $\tau = -mB\sin(\theta)$, with sign opposite to the sign of θ):

$$\begin{aligned} U &= -\int \tau\, d\theta \\ &= -\int_{\pi/2}^{\theta}(-mB\sin(\theta))\ d\theta \\ &= -mB\cos(\theta) \end{aligned}$$

or

$$U = -\vec{\boldsymbol{m}}\cdot\vec{\boldsymbol{B}} \tag{6.62}$$

Note that as before, $U(\theta)$ is minimum (negative) when the magnetic dipole is aligned with the field, maximum (positive) when antialigned.

From this, we can also find the force acting on a magnetic dipole in a *non*-uniform magnetic field:

$$F_x = -\frac{dU}{dx} \tag{6.63}$$

(with similar expressions for the other force components, where this derivative should really be a *partial derivative* for those of you who have taken multivariate calculus).

We can now construct a table of the analogies between electric and magnetic dipole moments and their associated fields, forces, and torgues. It is quite strong:

Quantity	Electric Dipole	Magnetic Dipole
Dipole Moment	$\vec{\boldsymbol{p}} = q\vec{\boldsymbol{\ell}}$	$\vec{\boldsymbol{m}} = NIA\hat{\boldsymbol{n}}$
Force in Uniform Field	$\vec{\boldsymbol{F}} = 0$	$\vec{\boldsymbol{F}} = 0$
Torque in Uniform Field	$\vec{\boldsymbol{\tau}} = \vec{\boldsymbol{p}} \times \vec{\boldsymbol{E}}$	$\vec{\boldsymbol{\tau}} = \vec{\boldsymbol{m}} \times \vec{\boldsymbol{B}}$
Potential Energy	$U = -\vec{\boldsymbol{p}} \cdot \vec{\boldsymbol{E}}$	$U = -\vec{\boldsymbol{m}} \cdot \vec{\boldsymbol{B}}$
Force in Non-Uniform Field	$\vec{\boldsymbol{F}} = -\vec{\boldsymbol{\nabla}} U$	$\vec{\boldsymbol{F}} = -\vec{\boldsymbol{\nabla}} U$

Table 3: Similarity of results for the electric and magnetic dipoles in (or later, as the source of) their respective fields.

6.4.1: Advanced: Comment on Magnetic Fields and Work/Potential Energy

As noted above, magnetic fields do no work on classical charged particles, no matter how many of them there are or how they are moving. This is a direct consequence of:

$$\boxed{P = \frac{dW}{dt} = \vec{\boldsymbol{F}}_m \cdot \vec{\boldsymbol{v}} = q(\vec{\boldsymbol{v}} \times \vec{\boldsymbol{B}}) \cdot \vec{\boldsymbol{v}} = 0} \tag{6.64}$$

as an *identity*, plus the superposition principle.

However, we just derived an expression for the potential energy of a rotating charge distribution or collection of charges moving in a loop to form a (coarse-grained) current loop in a magnetic field, using our usual definition of potential energy being the negative work done by the magnetic force/torque ***and got a nonzero answer!*** What gives?

There are two distinct cases to consider. In the first case, the magnetic moment in question is created by *moving physical charge around in a loop* while the charge itself is constrained to *stay* in the loop of motion by internal electro*static* forces. This is the case for our classical currents in conducting wires, but it is equally the case for static electric charge constrained to stay on the rim of an insulating ring by electrostatic forces while the ring is rotated – our very next example.

In this case, all magnetic force is generated by the action of the magnetic field on the microscopic charged particles that *transfer* that force to the rigid object via electric forces that hold the charge in place, or in the case of a ring of electric current, inside the conductor. As we'll show in a specific example in a couple more chapters, the actual work done on the rotating ring of matter or rotating ring of current turns out to be done by a mix of internal electric forces plus (in the case of the ring of current) the electric forces inside the battery that maintains the current against the resistance of the wire.

The second case has no proper classical analog, but we'll build some semi-classical models that come close. It turns out that charged elementary particles – quarks, charged leptons such as electrons and μ-mesons, and the $W_\pm$ heavy vector boson in the standard model – have intrinsic *spin* angular momentum and an associated *non-classical* magnetic moment. The magnetic moment of an electron, for example, cannot be due to physical charge rotating in a physical open loop of any sort held together by some combination of forces because in quantum mechanics any such loop must have *integer* spin rather than the *half-integer* spin observed in the case of the electron. Also, the electron is a "true point particle" as far as our ability to resolve such things goes (and indeed *must* be pointlike according to a very simple argument where elementary particles are elementary precisely because they are *not* bound states of other 'more elementary' particles with some internal interaction).

In this case the classical argument we give later fails, but ***it is still the case empirically that*** $U = -m \cdot \vec{B}$ ***holds!*** Indeed, electron spin resonance and nuclear spin resonance are (these days) commonplace examples of magnetic energy and magnetic forces in action – the latter in industrial production as the basis for *magnetic resonance imaging* described below. Furthermore, as noted, electron beams passed through inhomogeneous magnetic fields *split* into two beams[91] according to the spin of the electrons even in the complete absence of electric fields! Even electrically neutral *neutron* beams split in the Stern-Gerlach experiment! It is very difficult to ascribe the work done splitting a beam of uncharged particles passing through a space with no electrical field present at all to anything but the magnetic field, especially when the results are accurately *predicted* by the magnetic interaction between that field and their intrinsic magnetic moments!

Finally, all of the current theories and models for paramagnetism and ferromagnetism involve unbalanced intrinsic spin of (usually) electrons resulting in a given kind of atom or molecule having an intrinsic magnetic moment that again does not arise from the electrostatically constrained physical motion of charge so that the theorem above, based on magnetic force acting on a charged particles in the *absence* of any intrinsic magnetic moment attached to those particles, does not hold. Hence every time you let a magnet "snap" onto the side of your refrigerator, magnetic forces are doing work.

Fortunately, we don't really have to choose to use $U = -\vec{m} \cdot \vec{B}$ only when $\vec{m}$ is an intrinsic moment. It turns out that – electrostatic or not – the *net* work *gained or lost by the system of particles* that make up a rotating ring of charge or current carrying loop is precisely described by this formula even when the work is *really* due to e.g. internal electrostatic forces binding the material all together plus external energy sources such as batteries or generators. Indeed, you'd do just fine in the course if no one ever told you that (classical) "magnetic forces do no work" (on classical charged particles constrained to move in non-free trajectories by the *combination* of magnetic forces and internal non-magnetic forces) and just accepted $U = -\vec{m} \cdot \vec{B}$ as being "true".

From this point on, then, that's just what we'll do!

[91] Wikipedia: http://www.wikipedia.org/wiki/Stern-Gerlach experiment.

6.5: The Magnetic Moments of Rotating Charged Objects

Not all current carrying wires or current densities will have magnetic dipole moments that are easy to compute. In fact, *most* current densities will have moments that are *too difficult* to compute with anything less than a computer! Imagine a spool of wire tangled up like fishing line with a current running through it – this is only one of the infinity of arbitrary shapes to consider, most of which cannot even be expressed as a simple function of three dimensional coordinates! Still, our plane figure result above appears to be *very useful* because when we as humans design a magnetic apparatus (say, a motor) we can certainly *choose* to wrap our coils in a plane (at least approximately). Also, we can see how to at least *formulate* the problem for arbitrary currents. We are therefore done (for this level of instruction) with current loops per se until your next course in electrodynamics (if there is one) where you will learn how to compute moments from integrals over current densities expressed nastily in multivariate calculus.

However, there is one more generic distribution of moving charge that has an easily computable magnetic moment that we very much need to consider before quitting. A surprisingly common occurrence in physics is to have a "particle" (that is microscopically more or less a ball with a mass and a charge) that is *rotating* about some axis. A proton, for example, can be modelled as a ball of some radius $r_p \approx 10^{-15}$ meters, containing a mass m_p and a charge e. The proton also has a *spin* – an intrinsic angular momentum – of $L_z = \hbar/2$ where $\hbar$ is Planck's constant over 4π (a number that need not concern us in this course – it is very small in macroscopic terms but is *large* as far as the proton's physical dimensions are concerned).

We can then build a *classical model* for a proton, where we imagine the proton to be a uniform ball of charge with radius R, with total (uniformly distributed) charge $Q = e$, with (uniformly distributed) mass M, spinning about some axis through its center at an angular velocity $\vec{\boldsymbol{\Omega}}$ so that:

$$\vec{\boldsymbol{L}} = I\vec{\boldsymbol{\Omega}} = \left(\frac{2}{5}MR^5\right)\vec{\boldsymbol{\Omega}} \tag{6.65}$$

Hopefully it is clear that the proton will *also* have a *magnetic moment* parallel to $\vec{\boldsymbol{L}}$. What is *not* so obvious is that this magnetic moment will be directly proportional to the angular momentum in a way that is independent of the shape of the proton (or even that it is a proton), so that

$$\vec{\boldsymbol{m}} = \frac{Q}{2M}\vec{\boldsymbol{L}} = \mu\vec{\boldsymbol{L}} \tag{6.66}$$

for *any* symmetric spinning particle with identically distributed charge and mass, where I have defined the ratio:

$$\mu = \frac{Q}{2M} \tag{6.67}$$

as the classical equivalent of the "Bohr magneton" in the quantum physics of the electron[92].

Let us understand this, starting with a simpler example than a ball.

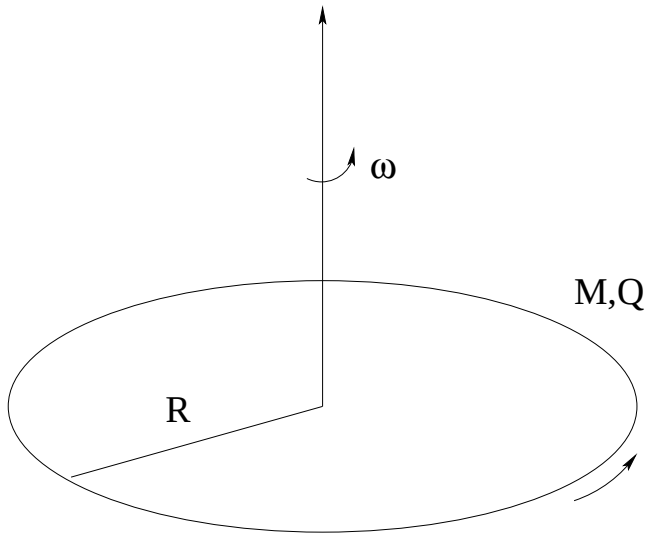


Figure 6.12: A rotating ring of charge with mass M, radius R, and charge Q has a magnetic moment of $\vec{\boldsymbol{m}} = Q/2M(MR^2\vec{\boldsymbol{\Omega}}) = \mu\vec{\boldsymbol{L}}$.

Example 6.5.1: The Magnetic Moment of a Rotating Ring of Mass and Charge

Suppose we have a ring of charge Q, mass M, and radius R spinning at angular speed Ω about its axis of symmetry as drawn in figure 6.12

The "current" in such a ring can easily be evaluated. The total charge in the ring goes around exactly one time in one period of its revolution. Thus:

$$I = \frac{Q}{T} = \frac{Q\Omega}{2\pi} \tag{6.68}$$

The magnetic moment of the ring in the (right handed) z-direction is thus just:

$$m_z = IA = \frac{Q\Omega}{2\pi}\pi R^2 = \frac{Q\Omega}{2}R^2 \tag{6.69}$$

If we multiply the expression on the right by $\frac{M}{M}$ (one!) and rearrange the terms, we get:

$$m_z = \frac{Q}{2M}\left(MR^2\Omega\right) = \mu L_z \tag{6.70}$$

using $L_z = MR^2\Omega$ for a ring of mass M rotating symmetrically about the z-axis.

That was almost too easy!

Example 6.5.2: Magnetic Moment of a Rotating Charged Massive Disk

Next consider a rotating disk of total charge Q, total mass M, radius R. The charge of a differential ring of charge of radius r and thickness dr is just

$$dq = \sigma dA = \sigma(2\pi r dr) \tag{6.71}$$

[92]The Bohr magneton of the electron is $\mu_B = \frac{e\hbar}{2m_e}$, which we recognize as our μ, but with the units of $\hbar$ appended to make the remaining parameter a dimensionless quantum number.

where $\sigma = Q/\pi R^2$ is the surface charge density of the uniformly distributed charge. The current in just this differentially thick ring is:

$$dI = dq\frac{\Omega}{2\pi} \tag{6.72}$$

just as it was for the ring example above. The area inside the differential ring is $A = \pi r^2$, so its differential magnetic moment is:

$$dm_z = dIA = 2\pi r dr\frac{Q\pi r^2}{\pi R^2}\frac{\Omega}{2\pi} = \frac{Q}{R^2}\Omega r^3 dr. \tag{6.73}$$

Integrating (to cover the disk) from 0 to R we find:

$$m_z = \int_0^R \frac{Q}{R^2}\Omega r^3 dr = \frac{Q}{4R^2}R^4\Omega = \frac{Q}{4}R^2\Omega \tag{6.74}$$

Once again we multiply by $\frac{M}{M} = 1$, do some rearrangement, and *presto, change-o:*

$$\boxed{m_z = \frac{Q}{4M}MR^2\Omega = \frac{Q}{2M}\left(\frac{1}{2}MR^2\Omega\right) = \mu L_z} \tag{6.75}$$

Part of your homework for this week will be to re-prove these two cases and several others to show that:

$$\vec{\boldsymbol{m}} = \frac{Q}{2M}\vec{\boldsymbol{L}} \tag{6.76}$$

is quite general, *subject to the condition that Q and M are proportionally distributed,* so that:

$$\frac{\rho_Q}{\rho_M} = \frac{Q}{M} \tag{6.77}$$

at all points inside the object, and (if you are do the advanced problems) that the object is rotating around a principle axis (an axis with enough symmetry that $\vec{\boldsymbol{\Omega}}||\vec{\boldsymbol{L}}$. This is simple enough if you require that both the mass and the charge have densities that are *identical functions of coordinates* and write dm_z correctly in terms of those densities.

So fine, fine, fine. Given this result we can now see that our classical model proton *should* have a magnetic moment that is related to its angular momentum by the simple relation:

$$\vec{\boldsymbol{m}}_p = \mu_p\vec{\boldsymbol{L}}_p \tag{6.78}$$

where $mu_p = \frac{e}{2m_p}$, the magnetic moment of a classical electron should be the same with $\mu_e = \frac{e}{2m_e}$ and so on, and it isn't *that* difficult to directly integrate over a solid sphere of charge/mass as we did in these examples to prove it! Indeed, this result works adequately in the case of *quantum* magnetic moments of elementary particles as well, as long as we remember to use the *intrinsic spin* of the particles in question.

Why do we care? It is because we can *use* this result in a clever way by taking advantage of the *motion* that results when we place a proton in a strong magnetic field. The motion, as we shall see, is a *precession* of the magnetic moment of the proton in a cone *around* the applied magnetic field that has a precession frequency $\Omega_p = \mu B$ independent of the relative angle between the angular momentum or spin of the proton and the magnetic field.

While precessing in this way, we can easily trick the magnetic dipole moments of the charged protons to *absorb or emit* electromagnetic radiation of the same angular frequency as Ω_p. By detecting the signal produced by the protons in various clever ways (beyond the scope of this course to detail, but *within your capabilities of understanding* if you master the next section) we can measure the *density* of bare protons in almost any substance and create a *three dimensional map* of that density at a remarkably fine resolution.

Protons, of course, are the nuclei of *hydrogen atoms* and water is dihydrogen oxide, with two protons just waiting to be mapped. And what are we? Well, mostly water! The precession of magnetic moments of protons around strong applied fields is the basis of *magnetic resonance imaging* (MRI), one of the most important technologies in use in hospitals around the world today. With MRI one can safely map out soft tissue densities of the human body in a lovely complement to x-rays (that map out dense tissues but that go right through soft tissue without much differentiation). My wife is a physician, and she orders MRIs on patients on at least a weekly basis, if not a daily one.

Spin resonance is also a very important experimental probe for physicists, as this trick works for more than "just protons". Whether you are a potential physics major or engineering student or a premedical student, you really *must* master the next section, then, as it is actually directly important to your future planned career. To encourage this mastery, I typically tell my students that a problem on magnetic resonance and precession *will* be on at least one quiz, hour exam, or on the final. This is usually enough incentive to motivate them to take the time to plow through the complexities of torque as the time rate of change of the *vector* angular momentum.

I present this result two distinct ways below – the first suitable for any student, the second perhaps better for students that have mastered the concept of the cross product in cartesian coordinates. I strongly suggest that *all* students at least *try* to master both, but at the very least get to where you fully understand the first one.

6.6: The Precession of Magnetic Moments: Magnetic Resonance

In figure 6.13 above, you can see a cartoon classical proton in a *strong* external magnetic field $\vec{B}_0$. The proton (we imagine) is spinning like a little planet – very little indeed given that its radius is order of 10^{-15} meters – and hence has an *angular momentum* $\vec{L}$ pointing in the direction up and to the right along its axis of rotation. Because its charge is *positive*, it has a magnetic moment that is parallel to its angular momentum, and in the previous section we argued strongly (leaving actual proof to the student) that as long as its charge and its mass are identically distributed, its magnetic moment can generally enough be written:

$$\vec{m} = \frac{e}{2m_p}\vec{L} = \mu_p\vec{L} \tag{6.79}$$

where $e = 1.6 \times 10^{-19}$ Coulombs is its charge and $m_p = 1.67 \times 10^{-27}$ kilograms is its mass (in SI units).

As we have derived above, the magnetic field exerts a *torque* $\vec{\tau}$ on the magnetic dipole of the proton. This torque is out of the page at the particular instant drawn above and is

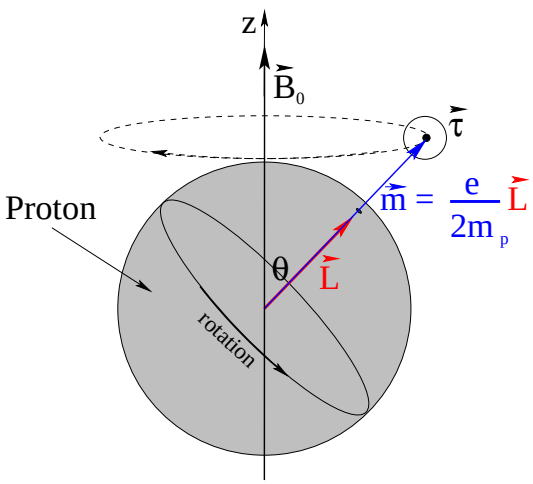


Figure 6.13: A *classical model* for a rotating proton with a (blue) magnetic moment $\vec{m} = \mu_p \vec{L}$ aligned with (red) its angular momentum along its rotation axis. This proton *precesses* around an applied magnetic field $\vec{B}_0$ with a precession frequency $\Omega_p = \mu_p B_0$ independent of the particular angle θ between $\vec{m}$ and $\vec{B}_0$. Note that *classically*, we expect $\mu_p = \frac{e}{2m_p}$ from the previous section.

quantitatively given by:

$$\vec{\tau} = \vec{m} \times \vec{B}_0 \tag{6.80}$$

or (using the fundamental definition of the torque as the time rate of change of the angular momentum):

$$\frac{d\vec{L}}{dt} = \mu_p(\vec{L} \times \vec{B}_0) \tag{6.81}$$

This is a very important *first order, linear, homogeneous, ordinary differential equation*. As we (hopefully) learned in the previous mechanics course where we studied precessing bicycle wheels and gyroscopes, the solution to this equation leads to the *precession* of the angular momentum around $\vec{B}_0$ in the direction indicated, at an angular velocity Ω_p. Ideally, you remember how to solve it (two or three ways!), but we will nevertheless review the more intuitive/graphical solution below. Consider figure 6.14.

The magnitude of the torque is given by:

$$\tau = \mu_p L B_0 \sin\theta = \frac{d\vec{L}}{dt} \approx \frac{\Delta\vec{L}}{\Delta t} \tag{6.82}$$

in finite sized steps (to make it easier to visualize). We break $\vec{L}$ up into two instantaneous components, L_z parallel to $\vec{B}_0$ (and hence the z axis as drawn) and $L_\perp$ perpendicular to it and hence lying in the x-y plane as drawn in figure 6.14. It should be obvious that:

$$L_z = L\cos\theta \qquad\qquad L_\perp = L\sin\theta \tag{6.83}$$

and that the direction of the torque is tangent to the circle with radius $L_\perp$ shown in both figures 6.14 and 6.15. Recalling our discussion of cross products in mechanics, we can easily see that the torque only comes from, and only changes, the $L_\perp$ component.

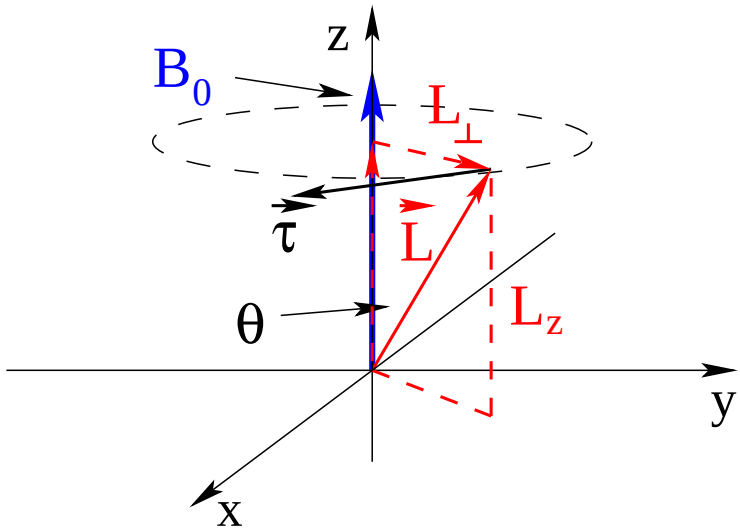


Figure 6.14: The torque causes (red) $L_\perp$ to precess around the (blue) $\vec{B}_0$ field in the direction shown, perpendicular to $L_\perp$, causing $L_\perp$ to swing around in a circle. (Red) L_z is constant as it is parallel to $\vec{B}_0$! θ is the angle between $\vec{L}$ and $\vec{B}$.

Since the torque is always *perpendicular* to $\vec{L}_\perp$ it changes its *direction* but not its *magnitude*. This is a familiar situation in physics – obviously $\vec{L}_\perp$ turns in a circle of radius $L_\perp$ where $\vec{\tau}$ is always perpendicular to it. This situation is pictured in an "overhead view" in figure 6.15 at an instant when $\vec{L}_\perp$ has both x and y components.

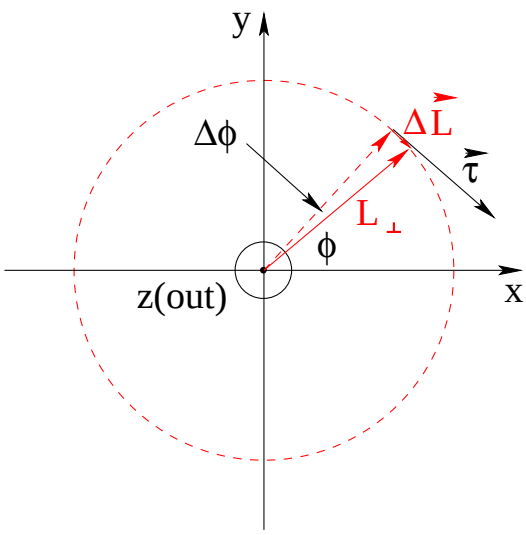


Figure 6.15: The torque causes (red) $L_\perp$ to precess around the $\vec{B}$-field parallel to the z-axis (out of the page). $L_\perp$ swings in a (red dashed) circle, and in a short time Δt it moves through an angle $\Delta\phi$ and hence changes the angular momentum by (red) $\Delta\vec{L}$ as shown.

In a short time Δt, the angular momentum changes a small amount with magnitude $\Delta L = \Delta L_\perp$ that is the length of the arc on the $L_\perp$ circle subtended by the angle $\Delta\phi$ through which it turns in that time, as shown in figure 6.15. In the figure it is obvious that:

$$\Delta L = L_\perp \Delta\phi = L \sin\theta \, \Delta\phi \tag{6.84}$$

or (dividing both sides by Δt) the magnitude of the rate of change must be:

$$\frac{\Delta L}{\Delta t} = L \sin\theta \frac{\Delta\phi}{\Delta t} \tag{6.85}$$

This must *also* equal the magnitude of the torque in terms of the field, in the limit that we let $\Delta t \to dt$, so (combining the two pieces):

$$\tau = \frac{dL}{dt} = \mu_p L \sin\theta B_0 = L \sin\theta \frac{d\phi}{dt} = L \sin\theta \Omega_p \tag{6.86}$$

where we define:

$$\Omega_p = \frac{d\phi}{dt} \tag{6.87}$$

to be the ***angular speed of precession***[93], commonly referred to as the ***Larmor frequency***[94] in the context of spin resonance.

Solving for the angular (Larmor) speed of precession (cancelling $L \sin\theta$), we find that:

$$\Omega_p = \mu_p B_0 \tag{6.88}$$

independent of the angle θ between $\vec{m}$ and $\vec{B}_0$! This latter fact, that the frequency of precession is independent of the angle, is a key aspect of magnetic resonance as it allows us to match an external rotating magnetic field to this frequency to make some magic happen – the "resonance" part.

Note well that this derivation, while correct enough for the moment, doesn't directly result in equations of motion for the individual components of the angular momentum – it is at least somewhat heuristic and relies on the pictures and visualization as much as the algebra. It is easy enough, however, to write down the *three coupled equations of motion for* L_x, L_y, L_z using the *cartesian* form for the cross product. One of these is trivial as there is no torque in the direction of $\vec{B} = B_0\hat{z}$. The other two first order coupled differential equations become ***second order equations*** for the ***oscillatory*** motion of L_x and L_y separately. The solutions, however, are not independent, as the *phase* of one is determined by the phase of the other. Indeed, the solution describes $\vec{L}_\perp$ tracing out an explicit circle at the precession frequency.

Instead of covering this solution in the text, this is left as an "advanced" (optional) homework problem for the interested student or a required problem for physics/engineering/math majors in the homework for this chapter. The math for this, note well, is very similar to the math used to derive the wave equation for the electric and magnetic field components from Maxwell's equations in a few chapters, so it isn't completely crazy to give this a try now even if you don't "have" to to make it easier on yourself then!

[93]This is a case where one can equally well call it an angular velocity, as the angular momentum sweeps out a real, physical angle per unit time, or an angular frequency since one *measures* the angular frequency of the *radio waves* emitted by the precessing angular momentum in e.g. nuclear magnetic resonance experiments or MRI. Most of the literature will call this ω_p or $\omega_{\rm larmor}$, in other words, but we are sticking with calling angular speeds involving actual angles Ω and true angular frequencies involving no actual angles ω, to avoid confusing the student in problems where both occur.

[94]Wikipedia: http://www.wikipedia.org/wiki/Larmor_precession. This article introduces and defines the gyromagnetic ratio and g-factor discussed in the next section, but is otherwise a bit quantum and complex for this course.

6.6.1: Advanced: Spin Echoes and Magnetic Resonance Imaging

One of the primary reasons for many students to take a course in electricity and magnetism is to learn enough about how magnetic fields and moments work that they can understand ***Magnetic Resonance Imaging*** (MRI). MRI is one of the most important non-invasive diagnostic tools available to physicians practicing modern medicine. It is also not terribly easy to understand even for physics majors because to completely understand it one has to understand a *lot* about both quantum mechanics and spin relaxation to do a completely proper job of it. This is especially true given that there are multiple somewhat distinct methods (that provide some degree of choice in contrast and resolution) that all come under the general heading of MRI and are all options on the hardware that can accomplish different purposes.

However, at this point you *should* know *enough* to understand a sort of a "toy model" of just how at least one or two of the MRI methods work, including the one that is arguably the most important (conceptually) to understand. This section is devoted to presenting just such a toy model. It deliberately omits most of the discussion of the quantum mechanics involved, while necessarily introducing certain very general terms and describing in a qualitative manner the key processes related to those terms. This section should very definitely be viewed as "optional" for most students but may serve as an introductory reference for students who are interested or who are confused by other descriptions.

Note well that this presentation is *my own* conception of the process, and while I do have some research experience with the related quantum theory of photon resonance and photon echos, I am far from being an expert on MRI and nuclear spin echos in particular in the context of MRI or otherwise. Those who are more expert than I who read this and find errors are encouraged to contact me to correct them, as long as the correction preserves the general semi-classical, functional presentation I am attempting that is as appropriate for introductory physics non-major students interested in the life sciences or medicine as it might be for future physics majors.

Although it can detect and map a number of nuclei, MRI is used medically to map primarily the density of *hydrogen nuclei* – protons – in the human body. This is because hydrogen is *by far* the most abundant element in addition to being one of the most responsive (in terms of having a large γ, defined below). Many of these protons are bound up in *water molecules* and are screened to some extent from electromagnetic fields and radiation by the surrounding molecular electron clouds. One can then treat them as "isolated" protons and add a phenomenological correction that describes the effect of small variations in their local fields magnetic fields.

From our discussion above, we *classically* expect the magnetic dipole moment of an isolated proton to be given by an expression such as:

$$\vec{\boldsymbol{m}}_p = \frac{e}{2m_p}\vec{\boldsymbol{S}} \tag{6.89}$$

where $\vec{\boldsymbol{S}}$ is the ***spin*** (intrinsic, quantum, mechanical) angular momentum of the proton. There are several problems with this equation. The proton is not, in fact, a homogenous ball of spinning mass and charge. It is a composite particle made up of three quarks bound together with the strong nuclear force, and *they*, not the proton per se, have spin angular momentum (a purely quantum mechanical kind of angular momentum), not the classical "orbital" angular momentum described by $\vec{\boldsymbol{L}}$. Classically, the angular momentum $\vec{\boldsymbol{L}}$ we have studied can have

any value, but in quantum theory the spin angular momentum $\vec{S}$ is *quantized* so that it can only take on certain values, generally integer or half-integer values of $\hbar = 1.05 \times 10^{-34}$ joule-seconds. We will model such a spin as a *classical* fixed angular momentum that can point in any direction but not change its magnitude.

It is well beyond the scope of this course to go into any more detail about the quantum theory of angular momentum used to effectively set that magnitude, but at the same time we *do* want to connect the spin angular momentum of a proton, whatever it might be, to its magnetic dipole moment. This is accomplished by using a *semi-empirical parameter* γ such that:

$$\vec{m}_p = \gamma \vec{S} \tag{6.90}$$

In this equation, γ is the so-called ***gyromagnetic ratio*** for the proton relating its spin $\vec{S}$ (whatever value it might have) to its magnetic moment. It is commonly written as:

$$\gamma = \frac{ge}{2m} \tag{6.91}$$

where the dimensionless g is called (creatively enough) the "g-factor" and simply adjusts the classically expected gyromagnetic ratio (where $g = 1$) to the correct tabulated (measured and/or calculated) g-factor for the particle with spin under consideration. A concise discussion of this and accepted values for g/γ are available on Wikipedia[95] [96] .

We'll begin by borrowing a key *concept* from statistical mechanics without introducing the Boltzmann constant or defining more precisely what we mean by high and low temperatures. Qualitatively, then, at "high" temperatures, *all* of the possible energy states available to a collection of many particles with spin in an external magnetic field are equally populated. As one lowers the temperature of the system, it becomes more and more likely to find spins in energy states with lower energies instead of higher energies, out of the finite range of energies the particles can have. At "low" temperatures, most of the spins will therefore be in those states with the lowest available energies.

Suppose, then, we have a large number of protons in a coarse-grained chunk of matter that we will call a "spin packet", so named because all of the spins in that chunk experience "the same" external magnetic field from some external arrangement of coils carrying currents, to some resolution. In the *absence* of any strong external magnetic field (when those coils are turned off, that is) and neglecting any internal spin-spin interactions, all of the spins have *no* magnetic potential energy – so all temperatures become "high" temperatures relative to the range of available energies – and the spin angular momentum (and magnetic moment) of any given proton in the chunk is ***equally likely to point in all possible directions***. The *total* spin angular momentum and magnetic moment of the spin packet should therefore be very nearly zero, independent of temperature.

[95] Wikipedia: http://www.wikipedia.org/wiki/G-factor_(physics). This is provided primarily for physics majors, who will one day be expected to understand most of the omitted details in this discussion. Note well that this article uses $\pm\hbar/2$ as its spin angular momentum, which is technically the magnitude of S_z, not its total angular momentum which would usually be given as $\sqrt{s(s+1)\hbar^2} = \sqrt{3}/2\hbar$ for a spin-$\frac{1}{2}$ particle. This is irrelevant to my purpose here, though, which is to give you a decent *conceptual* idea of what *happens* to generate a spin echo, given the appropriate resonance frequency.

[96] Wikipedia: http://www.wikipedia.org/wiki/Proton_magnetic_moment. This is specific to the NMR/MRI/spin echo theory for a model proton that we are discussing.

Following from this, if we put our spin packet of protons into a *strong external magnetic field* in (say) the z-direction:

$$\vec{\boldsymbol{B}} = B_0 \hat{\boldsymbol{z}}, \tag{6.92}$$

(where B_0 is the magnitude of the field within that packet, but might be slightly different for neighboring spin packets) the potential energy of each proton suddenly *does* depend on its direction relative to the field:

$$U = -\vec{\boldsymbol{m}}_p \cdot \vec{\boldsymbol{B}} = -m_z B_0 \tag{6.93}$$

where m_z is the z-component of its magnetic moment, $m_z = \gamma S_z$. The magnetic potential energy of any given proton is minimized when its spin is *aligned* with the external field, and is maximized when it is *antialigned* with the external field. In general this makes it *more probable* that spins will be found at least partially aligned with the field than either antialigned or randomly aligned where the lower the temperature of the spin packet (or the stronger the external field!), the greater the expected degree of alignment in thermal equilibrium.

If one starts from no field and then "suddenly" turns on a field, the spin packet is initially *not* in thermal equlibrium. Nor will the external magnetic field itself *put* it into thermal equilibrium – the conservative magnetic interaction itself will just cause individual spins to precess *around* the magnetic field in the z-direction with a constant potential energy and won't alter the angle between the spin and the external field at all!

There are two ways the spin can alter its orientation *relative* to this simple precession around the strong external field. The first is that each spin interacts with the bulk material it is a part of, referred to as the "lattice" of atoms and molecules, and can absorb energy from or lose energy to this lattice. This is called the "spin-lattice" interaction and leads to spin-lattice relaxation towards thermal equilibrium with that lattice. In time, this interaction causes the spins in the packet to align more with the external field than against it, developing a macroscopic collective magnetic moment as it does so in the direction of the field. The relaxation process is best described by our "saturating exponential" solution like that describing the charging of a capacitor in an RC circuit, with a time constant called T_1, the spin-lattice decay time.

Typical values for T_1 for protons in living tissue at fields of 1.5 tesla and normal body temperatures are observed to fall between 1 and 2 seconds and get longer at higher field strengths. One achieves 99% of the possible peak magnetization of a sample in $4.6 \times T_1$ seconds after turning on the field with the spins initially in an random, unmagnetized state (a useful number to remember for exponential saturation processes).

The second way a spin can alter its orientation is via local spin-spin interactions – one spin reacting to the magnetic field produced by a second, nearby spin as *both* interact systematically with the strong external field and more or less randomly with the lattice. These interactions are conservative and can only exchange energy, but they do cause the components of the spins *perpendicular* to the direction of the strong field to spread out and experience a transverse exponential "spin-spin relaxation" with a characteristic exponential decay time referred to as T_2.

A third way spins in a *bulk sample* can experience a kind of "reversible" transverse relaxation is because the strong external field experienced by all of the spin packets that make up the sample is not exactly the same or equal to the "average" magnetic field in the sample!

Some spin packets will experience slightly stronger fields than the average and will systematically pull ahead of the precession of a spin packet in the average field. Others will be in slightly weaker fields than average and will systematically get left behind. This, too leads to an energy-conserving exponential decay of the average component of the magnetization (if any) perpendicular to the external field called *inhomogeneous relaxation* $T_{2,\mathrm{inhomogeneous}}$.

Even this isn't everything. Spins can interact with the more or less "fixed" static magnetic field produced by their local *chemical* environment – the orbiting electrons, for example to produce an effect similar to and mixed in with spin-spin T_2. The T_2 decay times can depend on the field strength itself, and hence on T_1. When we (eventually) hit a system with a pulse of radiation at some fixed frequency, that frequency is "broadened" by purely Fourier effects, twisting the simple rotations described below sideways. We will lump all of the transverse decay times into one "average" time and call it T_2^*, where a significant fraction of T_2^* will be due to $T_{2,\mathrm{inh}}$, and leave it at that.

Now let's trace out of the mechanics of spin echoes and magnetic resonance by considering a chunk of protons (made up of many spin packets with slightly different resonant frequencies) that have been in a strong $\vec{B}$-field for many times T_1 and hence have more or less completely relaxed to the low energy state of maximum equilibrium alignment with the strong external $\vec{B}$ field.

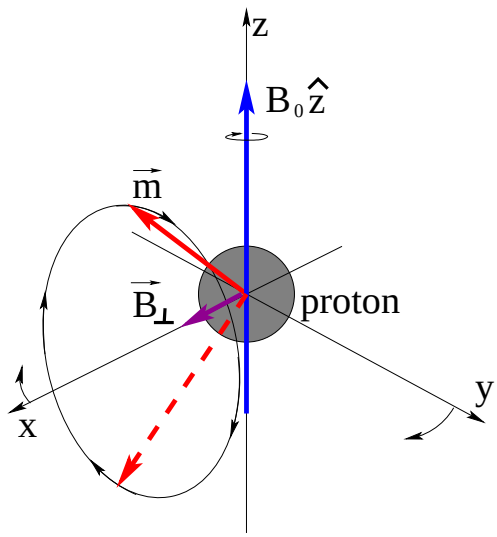


Figure 6.16: A single proton precessing around a strong $\vec{B}$-field aligned with the z-axis will *also* precess around a *weak* $\vec{B}$-field rotating around the z-axis at the resonant frequency in the "rotating frame".

Suppose we look at a single proton in this very strong field oriented in the z-direction, where the field is represented as the thick blue arrow in figure 6.16. Its spin/magnetic moment is classically portrayed as a solid red arrow initially "mostly" parallel to the $\vec{B}$-field but with a small $\hat{x}$ component. In the absence of any other fields:

$$\vec{\tau}_{\mathrm{tot}} = \frac{d\vec{S}}{dt} = \vec{m} \times \vec{B} = \gamma \vec{S} \times (B_0 \hat{z}) \tag{6.94}$$

We solved this above with a few minor tweaks – $\vec{m}$ precesses rapidly around the $\vec{B}$-field aligned with the z-axis with angular Larmor frequency $\Omega_p = \gamma B_0$, as we derived above for

the *classical* angular momenta of spinning charged balls. Fortunately, this classical behavior is almost perfect preserved in the case of *quantum* spins, so our classical picture is good enough for us to conceptually understand what is going on!

Note that so far, we are assuming that this proton is in a locally uniform field that is *precisely* equal to the average field for the entire sample, and are momentarily neglecting both spin-lattice relaxation (T_1) and the spin-spin interaction between neighboring protons and other T_2 stuff because they are *slow* relative to the period of precession of the spin in an e.g. 1.5 T field, $T_p = \dfrac{2\pi}{\Omega_p}$. We'll add all of this stuff back later.

Now, at $t = 0$ we mentally turn on a *much weaker* magnetic field:

$$\vec{\boldsymbol{B}}_\perp(t) = B_\perp \left(\cos(\Omega_p t)\hat{\boldsymbol{x}} - \sin(\Omega_p t)\hat{\boldsymbol{y}}\right) \tag{6.95}$$

while leaving the strong field in the z direction on. Note that this field is perpendicular to $B_0\hat{\boldsymbol{z}}$, initially points (at $t = 0$) in the $\hat{\boldsymbol{x}}$ direction, and ***rotates around the z-axis at the same (average) Larmor frequency Ω_p and in the same direction that the magnetic dipole moment of our typical proton precesses around the strong field!***

Technically, then, we should try to solve the differential equation involving *all three field components, two of them time dependent*:

$$\vec{\boldsymbol{\tau}}_{\rm tot} = \frac{d\vec{\boldsymbol{S}}}{dt} = \vec{\boldsymbol{m}} \times \vec{\boldsymbol{B}} = \gamma\vec{\boldsymbol{S}} \times \left(B_\perp \cos(\gamma B_0 t)\hat{\boldsymbol{x}} - B_\perp \sin(\gamma B_0 t)\hat{\boldsymbol{y}} + B_0\hat{\boldsymbol{z}}\right) \tag{6.96}$$

to obtain $\vec{\boldsymbol{S}}(t)$ and hence $\vec{\boldsymbol{m}}(t)$, but *in general* this highly nonlinear differential equation is a *mess* and trying to solve it analytically will make you ***very sad!***

Fortunately, however, a mathematical "miracle" happens in the special case when $B_\perp \ll B_0$. When this is true, it turns out we can describe the motion of the magnetic moment *fairly* accurately by transforming our entire description to the so-called *rotating frame*[97] – a frame that is rotating around the z-axis at the Larmor frequency $\Omega = \gamma B_0$ and in the same direction that both $\vec{\boldsymbol{B}}_\perp$ and $\vec{\boldsymbol{m}}$ are precessing. In this (primed) frame:

$$\vec{\boldsymbol{B}}'_\perp(t) = B_\perp \hat{\boldsymbol{x}}' \tag{6.97}$$

is *stationary*. If $B_\perp = 0$, the magnetic dipole moment of our ideal proton $\vec{\boldsymbol{m}}'$ (primed as it is expressed in the rotating frame) is *also* stationary in this frame.

In this special case, we can more or less ignore B_0 altogether in the rotating frame and solve for the precession of $\vec{\boldsymbol{m}}'$ *around* $\vec{\boldsymbol{B}}_\perp$ as if it were the only field present! This precession will be "slow" compared to the rate the moment is whipping around the z axis, which is why the motion can be separated in this way. In the end, we can find the solution in the rotating frame and then rotate that solution back into the lab frame via Ω_p to get a very respectable description of the state in the lab frame at any particular time!

[97] The easiest way to think of the rotating frame is to imagine a flat merry-go-round with an x'-y' axis painted onto its floor and the center pole marked as the $z = z'$ axis. When this frame rotates around the z/z' axis *at* the Larmor frequency in the right direction, a spin precessing around a strong $\vec{\boldsymbol{B}}$-field lined up with z' appears – to riders on the merry-go-round – *stationary*, with *constant* $S_{x'}, S_{y'}, S_{z'}$ components! The trick, then, is to arrange it so that $\vec{\boldsymbol{B}}_\perp$ is *also* stationary in the rotating frame and weak enough that the spins that were stationary in the frame now slowly precess around $\vec{\boldsymbol{B}}_\perp$ – *in that frame* while the entire frame rotates.

Makes you kind of dizzy, doesn't it...

Understanding (and quantitatively describing) the motion is now easy! In the rotating frame $\vec{m}'$ (comparatively) *slowly* sweeps out a cone around the x'-axis at frequency $\gamma B_\perp$ while at the same time, in the lab frame *both* $\vec{m}$ *and* $\vec{B}_\perp$ *continue whirling around the* z*-axis at the resonant frequency* $\Omega_p = \gamma B_0$*!*

By varying the time you leave the perpendicular field $\vec{B}_\perp$ *turned on*, you can cause this ideal proton's magnetic moment $\vec{m}'$ to sweep out a cone through the *controlled* azimuthal angle $\pi/2$ (a so-called "$\pi/2$ pulse") or π (a "π pulse") around the rotating x' axis! We can even *compute* just how long the rotating field should be turned on to accomplish either one because we know that one period of *slow* precession in the rotating frame should be:

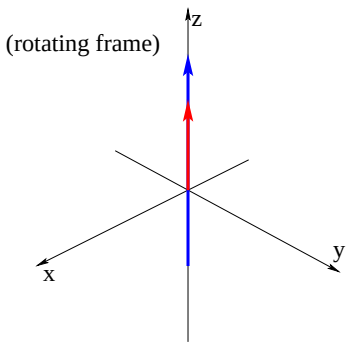
$$T_{\rm rot} = \frac{2\pi}{\gamma B_\perp} \tag{6.98}$$

so that a rotating field turned on for a quarter of a period, $T_{\rm rot}/4$ would be a $\pi/2$ pulse while a half a period $T_{\rm rot}/2$ would be a π pulse!

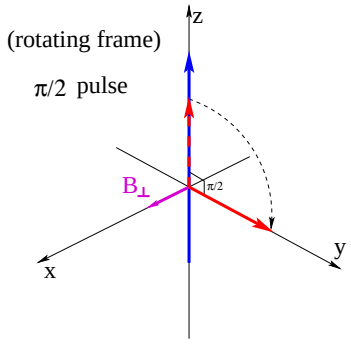
Figure 6.16 illustrates the effect of a π-pulse, taking the (initial) solid red $\vec{m}'$ vector from mostly aligned with z to the (final) dashed red $\vec{m}'$ vector that is mostly *anti*-aligned with z. Note well that this corresponds to the proton *absorbing energy* from the *electromagnetic radiation* associated with the rotating $\vec{B}_\perp$-field! For times that are *short* relative to the spin-lattice relaxation time T_1, the proton will *remain* in this new orientation, still precessing around the z-axis at the Larmor frequency Ω_p but *in a higher energy state than it began in at time* $t = 0$*!*

Now consider a *collection* of protons in a spin packet in thermal equilibrium with the strong field. By regulating the amount of time you leave on the rotating transverse field, $\vec{B}_\perp$, one can take the entire spin packet from the initial low energy state where its spins are mostly parallel to $B_0\hat{z}$ into almost any desired state of spin polarization in the rotating frame, where they will remain when the transverse field is turned off until spin-spin and spin-lattice relaxation first destroys the transverse component of the collective magnetization and as it does so, gradually rebuilds the collective moment in the z direction!

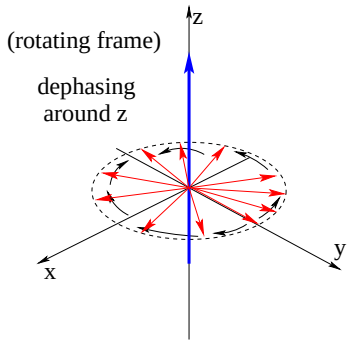
We are now finally ready to understand *spin echoes* and how they enable **Magnetic Resonance Imaging** (MRI). Here is the sequence of events.



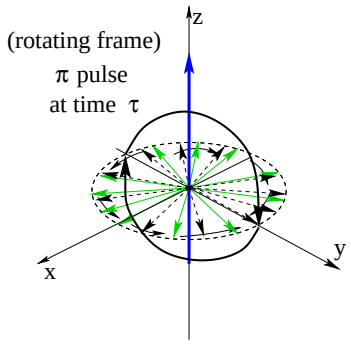
First, a collection of protons is placed in a *very strong* B*-field* oriented in (say) the z-direction. The protons quickly (in, say, more than $4.6 \times T_1$ to achieve over 99% magnetization) "relax" so that they are predominantly in the *lowest potential energy state* with their spins mostly aligned with this field consistent with thermal equilibrium at the temperature of the sample.



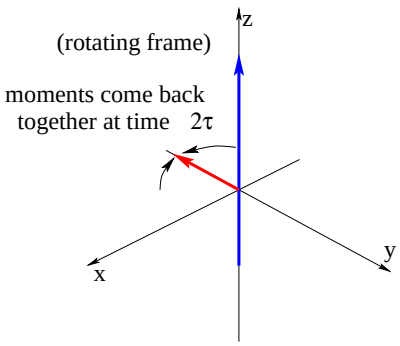
Next, $\vec{\boldsymbol{B}}_\perp(t)$ (in the x'-direction in the rotating frame) is turned on in a $\pi/2$ pulse, rotating the collective $\sum_i \vec{\boldsymbol{m}}_i$ of many protons all at once so they end up pointing in the y'-direction in the rotating frame. For a moment electromagnetic energy is *strongly* radiated out of the sample by the *large* collective rotating magnetic dipole, but that energy is mixed in with the $\pi/2$ pulse and difficult to detect.



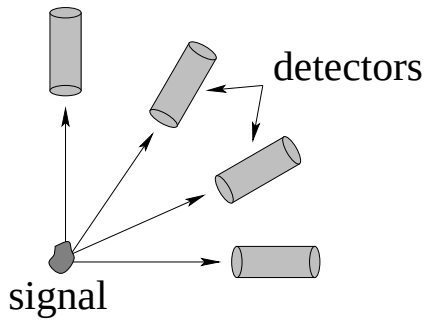
Third, *individual* spin packets with collective moment $\vec{\boldsymbol{m}}_i$ have very slightly different precession frequencies and systematically "dephase" due to $T_{2,\mathrm{inh}}$ and get ahead of or behind the "ideal spin" that is precessing at precisely the Larmor frequency Ω_p (and hence is stationary in the rotating frame). The total moment $\sum_i \vec{\boldsymbol{m}}_i$ summed over the spin packets in a given coarse grained chunk of the material being scanned rapidly averages to zero.



At a time T_R (where $T_2 < T_R < T_1$) after the $\pi/2$ pulse, the protons are hit with a π pulse of $\vec{\boldsymbol{B}}_\perp(t)$ that rotates them around the x'-axis so that their y' component (only) inverts in the rotating frame. The spin packets remain in the same slightly inhomogeneous field, however, so their collective moments continue to get ahead or behind in the same direction in the rotating frame, *unwinding* the total angle they accumulated since the $\pi/2$ pulse!



At a time $2T_R$ after the original $\pi/2$ pulse, the magnetic moments of all of the spin packets in the coarse grained chunk *all come back together* into a *single collective magnetic moment*, in a quiet environment with *no competing radiation!* This large collective rotating dipole (somewhat reduced by the spins that relaxed back into the mostly aligned state during $2T_R$) ***strongly radiates electromagnetic energy***, in a so-called ***spin echo*** pulse.



The protons in each coarse-grained chunk of an extended sample being scanned thus give off an easily detectable spin-echo *pulse* of electromagnetic radiation with a width roughly equal to $2T_2$ centered at time $2T_R$. The pulses from all of the chunks that "cover" the sample are picked up by a surrounding *array of detectors* at times that differ according to the time of flight from any given radiating chunk of protons given the distance of the chunk from each detector and the speed of light. By working backwards, the array signal is ultimately transformed into a 3D spatial map of ***proton density in the sample!***

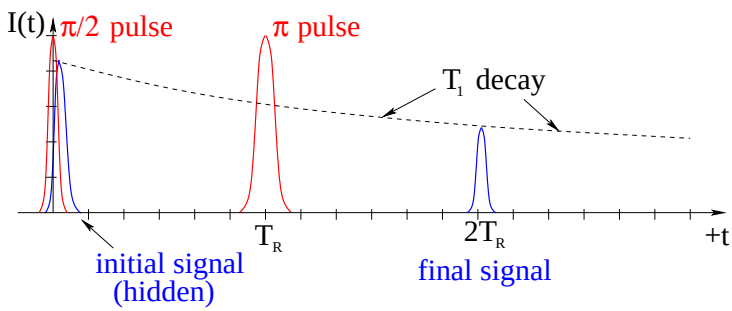


Figure 6.17: The key steps in nuclear spin echoes. The two red pulses show the intensity of an *external* rotating magnetic field applying $\pi/2$ and π pulses. The two blue pulses show the intensity of the radiated signal from the rotating spins. The black dashed curve shows the gradual loss of intensity in the second "echo" pulse due to T_1 (and the irreversible part of T_2) decay back into a state aligned with the external field during the entire process.

In figure 6.17, an approximate graph of the electromagnetic intensity as a function of time, $I(t)$, in the vicinity of protons shows the key features of the process. At $t = 0$, a (red) $\pi/2$ pulse creates a large y'-polarized rotating moment in the sample, that radiates an initial signal (in blue) that is quickly cut off at T_2 inhomogeneous dephasing spreads out the spins in the x'-y' plane. A time T_R later, a (red) π pulse effectively reflects each spin across the x'-z' plane, but *preserves* the direction and rate of the spin's slow dephasing precession in the x'-y' plane. This "unwinds" the dephasing so that the spins come together again at time $2T_R$, producing a final signal in the absence of the rotating field. This pulse is slightly attenuated by T_1 exponential decay of some of the spins back into the low-energy state mostly aligned with the strong external field, as well as spin-spin T_2 relaxation that mixes up the spins in way not reversible by anything as simple as a π pulse.

This concludes our discussion of magnetic spin resonance, MRI, spin echoes, and the like. Students interested in the subject are strongly encouraged to visit Wikipedia: http://www.wikipedia.org/wiki/Relaxa especially to see some excellent dynamical graphics that can help you visualize the process steps illustrated statically above, as well as provide you with more detailed descriptions of T_1, T_2, free induction decay, and more than I am able to include in what was supposed to be a *simple* walkthrough of Wikipedia: http://www.wikipedia.org/wiki/Magnetic Resonance Imaging.

As you will see if you peruse this more general article (or any of the other resources available on the Internet) medical MRIs exploit a vast range of parameter space associated with T_1,

T_2, $T_{2,\mathrm{inh}}$, and T_R that let them home in on specific diagnostic techniques and scans for different organs and disorders. The general *idea* of these scans is the generally the same, however, and a student who understands the general manipulations of collective spin-packet magnetization via the strong field and the rotating weak transverse field described above should be able, with some work, to understand them all.

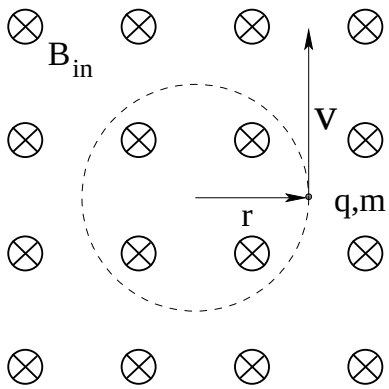
Homework for Week 6

Problem 1.

Physics Concepts

Make this week's physics concepts summary as you work all of the problems in this week's assignment. Be sure to cross-reference each concept in the summary to the problem(s) they were key to. Do the work carefully enough that you can (after it has been handed in and graded) punch it and add it to a three ring binder for review and study come finals!

Problem 2.



A particle with mass m and charge q a has a velocity $\vec{v}$ perpendicular to a uniform magnetic field $\vec{B}$ (with magnitude $B = |\vec{B}|$). Find:

a) the radius r of its orbit;

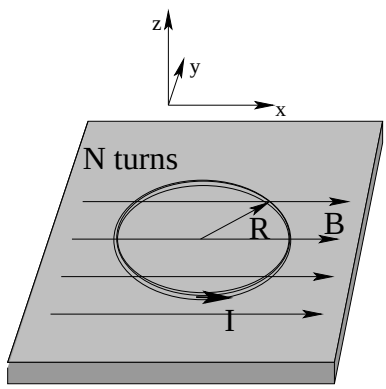
b) the period of the orbit;

c) the momentum of the particle;

d) the kinetic energy of the particle.

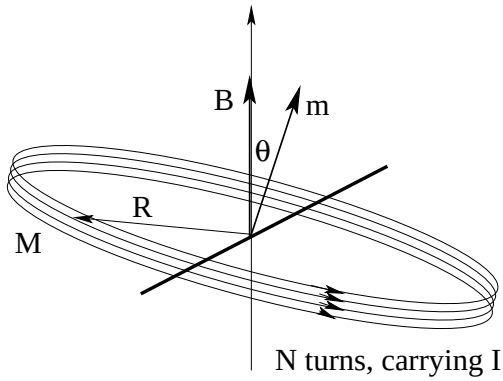
All answers but the first should be in terms of q, m, B and r – no v should appear in b-d

Problem 3.



A rigid circular loop of wire with mass m, N turns and radius R carries a current I in each turn and is sitting on a rough table. There is a horizontal magnetic field B that is parallel to the surface of the table in some direction (call it x). What is the minimum value of B sufficient to lift an edge of the loop off of the table? On your figure, clearly indicate which edge lifts relative to the directions you select for I and B.

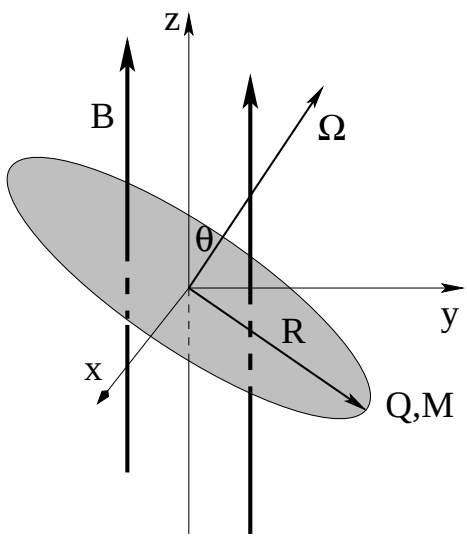
Problem 4.



A circular loop of wire with radius R, N turns, and total mass M carries a current I. It is pivoted about a line that passes through the loop as shown, then placed in a uniform magnetic field $\vec{\boldsymbol{B}} = B_0\hat{\boldsymbol{z}}$ so that its magnetic moment makes an initial angle of $\theta_0 \ll \pi$ with the z-axis at time $t = 0$, and is then released.

Describe its small-angle motion quantitatively. Note well that this arrangement has *no* angular momentum to speak of in the $\vec{\boldsymbol{m}}$ direction and will *not* precess!

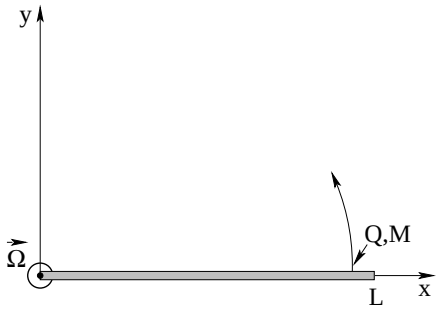
Problem 5.



A disk of uniformly distributed mass M, charge Q, and radius R is spinning at angular frequency Ω about its axis. Its axis, in turn, makes an angle θ with a powerful uniform magnetic field $\vec{\boldsymbol{B}} = B_0\hat{\boldsymbol{z}}$.

Find (in terms of these givens) the frequency Ω_p with which the magnetic moment *precesses* around the magnetic field, and *unambiguously indicate the direction of precession* at the instant portrayed to the left, as in $\vec{\boldsymbol{\Omega}}$ precesses "into the page" or "out of the page" or "into the $+\hat{\boldsymbol{z}}$ direction" or "$-\hat{\boldsymbol{y}}$ direction" etc.

Problem 6.



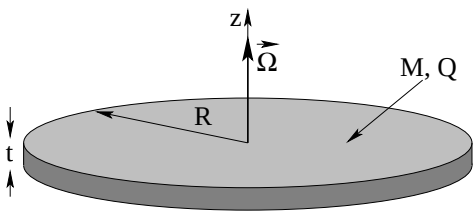
A nonconducting rod of total mass M and length L has a charge Q uniformly distributed along it. It is pivoted around one end and is rotating in the $x-y$ plane around the z-axis at angular frequency ω.

a) Consider a small bit of charge dq a distance r from the pivot and compute its average magnetic moment in the z-direction, dm_z.

b) Integrate this result and find the total magnetic (dipole) moment of the rotating rod m_z

c) Show that the result can be expressed as $m_z = \frac{Q}{2M} L_z$ where L_z is the angular momentum of the rod about the pivot (that is to say, in the z-direction).

Problem 7.



A disk of radius R and thickness t, with uniform charge density ρ_q and uniform mass density ρ_m is rotating at angular velocity $\vec{\omega} = \omega\hat{z}$.
Consider a tiny differential chunk of the disk's volume $dV = dA\ t$ located at r, θ in cylindrical polar coordinates. Note that this chunk is orbiting the z-axis at angular frequency ω in a circular path.

a) Find the magnetic moment dm_z of this chunk in terms of ρ_q, ω, dV and its coordinates.

b) Find the angular momentum dL_z of this chunk in terms of ρ_m, ω, dV and its coordinates.

c) Doing the two (simple) integrals, express them in terms of the total charge and total mass of the disk, respectively, and show that the magnetic moment of the disk is given by $\vec{m} = \mu_B \vec{L}$, where $\mu_B = \frac{Q}{2M}$.

d) What do you expect the magnetic *field* of this disk to look like on the z axis for $z \gg R$? (Answer in terms of $\vec{m}$ is fine.)

Advanced Problem 8.

Consider an "arbitrary" object with angular velocity $\vec{\boldsymbol{\Omega}} = \Omega\hat{\boldsymbol{z}}$ that has ***identical charge and mass distributions***, where these distributions have ***sufficient symmetry relative to the z-axis*** that $\vec{\boldsymbol{L}} = L_z\hat{\boldsymbol{z}} = I_z\Omega\hat{\boldsymbol{z}}$ (review the section on (un)balanced rotation in the Intro Physics 1 companion textbook if the meaning of this is not clear – the distribution(s) must either have reflection symmetry across the x-y plane or be reflection symmetric across the z axis itself – or both, like a sphere).

Find a relationship between dL_z (the z-component of the angular momentum of small chunk of mass $d\Updownarrow$ at a cylindrical coordinate radius r from the axis of rotation) and dm_z (the magnetic moment of the *same* small chunk of charge dq at the *same* cylindrical radius r) and integrate over the object on both sides to show that for all sufficiently symmetric (relative to the axis of rotation z) distributions with identical charge and mass distributions, $m_z = \frac{Q}{2M}L_z$. This result therefore holds for spheres, cylinders, disks, rods (in a plane), spherical or cylindrical shells, etc, ***as long as they rotate around a (principle) axis of symmetry!***

Advanced Problem 9.

For a presumed e.g. proton in a magnetic field, the equation of motion for the torque/angular momentum is:

$$\vec{\boldsymbol{\tau}} = \frac{d\vec{\boldsymbol{L}}}{dt} = \mu_p\vec{\boldsymbol{L}} \times \vec{\boldsymbol{B}} \tag{6.99}$$

where $\mu_p = e/2m_p$ and $\vec{\boldsymbol{B}} = B_0\hat{\boldsymbol{z}}$.

Solve these equations of motion for $\vec{\boldsymbol{L}}(t)$ ***in Cartesian coordinates*** when $\vec{\boldsymbol{L}}$ at $t = 0$ is arbitrary, that is:

$$\vec{\boldsymbol{L}}(0) = L_{x0}\hat{\boldsymbol{x}} + L_{y0}\hat{\boldsymbol{y}} + L_{z0}\hat{\boldsymbol{z}}$$

Your answer should prove that the angular momentum does indeed precess around the applied magnetic field with the constant angular velocity $\Omega_p = \mu_p B_0$ independent of $\vec{\boldsymbol{L}}$'s magnitude or direction. **Hint:** Don't forget the general solutions to the SHOE from Intro Physics 1!

Week 7: Sources of the Magnetic Field

- Magnetic and electric fields are clearly connected in many ways (some of them still to be learned). A perfectly reasonable question is: “Are magnetic fields created by a magnetic charge the same way electric fields are created by electric charge?” We will refer to such an isolated ”north” or ”south” pole as a *magnetic monopole*.

  It is empirically the case that *no isolated magnetic charges have been experimentally observed*, in spite of an electromagnetic theory that “begs” for them, a quantum theory that can explain charge quantization if a *single* magnetic monopole exists in the Universe, and in spite of an intense experimental search for them. It is probably safe to say that magnetic monopoles are at the very least *rare* in all the places we’ve looked for them!

- We express this (lack of discovderable monopoles, so far) by means of *Gauss’s Law for Magnetism*:

$$\oint_S \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA = 4\pi k_m Q_{m,\text{in S}} = \mu_0 \int_{V/S} \rho_m dV = 0 \tag{7.1}$$

  where the magnetic field constant $k_m = 10^{-7}$ tesla-meter/ampere *exactly*. Note that this constant is exact because SI units define the coulomb of charge as an ampere-second, not the other way around – Ampere “got there first”.

- The magnetic field constant can be written as:

$$k_m = \frac{\mu_0}{4\pi} \tag{7.2}$$

  where

$$\mu_0 = 4\pi \times 10^{-7} \text{ tesla-meter/ampere} \tag{7.3}$$

  is the ***magnetic permeability of free space*** and is the magnetic constant analoguous to ϵ_0, the dielectric permittivity of free space, just as k_m is the magnetic equivalent of k_e.

-

$$d\vec{\boldsymbol{B}} = k_m \frac{I d\vec{\boldsymbol{\ell}} \times \hat{\boldsymbol{r}}}{r^2} = \frac{\mu_0}{4\pi} \frac{I d\vec{\boldsymbol{\ell}} \times \hat{\boldsymbol{r}}}{r^2} \tag{7.4}$$

  is known as the *Biot-Savart Law* (Bee-oh Sa-var Law) for the magnetostatic field produced by stationary currents in a wire. In this expression, $d\vec{\boldsymbol{\ell}}$ is a differential length of the wire with a direction pointing *in the direction of the current* I in the wire. It must be integrated over the wire(s), presumed to carry a *constant* (or very slowly varying) current I.

- You should be able to use the Biot-Savart law to find the field of a straight wire segment, a current carrying loop on its axis of symmetry, and on the axis of symmetry of a rotating

ring or disk of charge. From either of the latter two (far from the disk or ring) you should be able to guess the *general* magnetic field of a magnetic dipole in terms of its dipole moment in analogy with the field of an electric dipole. Note that the *results* of doing this are not included in this summary because you aren't to memorize them, you are to learn how to find them!

- The field of a long straight wire carrying a current I is:

$$\vec{\boldsymbol{B}} = \frac{2k_m I}{r}\hat{\boldsymbol{\phi}} \tag{7.5}$$

where $\hat{\boldsymbol{\phi}}$ curls around the wire in the direction given by the *right hand rule*. In words, if you let your right-handed thumb point in the direction of the current (as you "grasp the wire" with your right hand) your *fingers* will curl around the wire in the direction of the magnetic field lines. This can be derived *either* from Ampere's Law (below) *or* the Biot-Savart Law, with the former being by far the easiest way but the latter is useful as well as a means to check that both lead to the same result and to be able to find the field of *short* straight wire segments carrying a current. We will use this particular result so often that it is worth remembering (and hence is in the summary) even though you should be able to easily find it if requested.

- The actual source for magnetic fields (in the absence of monopoles) is *moving charge*, either in the form of a more or less continuous current or from moving discrete point charges. The correct description of the field produced by a moving point charge requires the theory of relativity and hence is beyond the scope this course, but one can obtain an *approximate* result for the field produced by a ***slowly moving point charge*** (relative to the speed of light) from the Biot-Savart Law by treating the current in a differential chunk of the wire as a "moving point charge", that is $I\,d\vec{\boldsymbol{\ell}} \Leftrightarrow q\vec{\boldsymbol{v}}$ (a coarse-grained equivalence we have used before). The result is[98]:

$$\vec{\boldsymbol{B}} = k_m \frac{q\vec{\boldsymbol{v}} \times \hat{\boldsymbol{r}}}{r^2} = \frac{\mu_0}{4\pi}\frac{q\vec{\boldsymbol{v}} \times \hat{\boldsymbol{r}}}{r^2} \tag{7.7}$$

- Ampere's Law – for magnetostatic/steady-state currents – is

$$\oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} = \mu_0 I_{\text{thru C}} = \mu_0 \int_{S/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA \tag{7.8}$$

This is our *third* Maxwell equation, but...

[98]**Note Well!** This result is *not* general and will *not* work for charges moving at any appreciable fraction of the speed of light or at points in space that are "distant" from the source charge! The correct magnetic field produced by a point charge at the origin moving in the z direction at speed v at a point $P = (r, \theta, \phi)$ in spherical polar coordinates is:

$$\vec{\boldsymbol{B}} = k_m \left(\frac{1 - \frac{v^2}{c^2}}{1 - \frac{v^2}{c^2}\sin^2\theta} \right) \frac{q\vec{\boldsymbol{v}} \times \hat{\boldsymbol{r}}}{r^2} \tag{7.6}$$

which obviously reduces to the form above when $v \ll c$. You are not responsible for knowing this form (or the related form for the electric field of a moving charge, when the finite speed of propagation of both fields are taken into account) but far too many textbooks give the non-relativistic Biot-Savart result for the "field of a moving point charge" without the $v \ll c$ qualification.

- ...there is a *conceptual error* in this "broken" version Ampere's Law. The current I through an open surface S bounded by a closed curve C is *not invariant* as evaluate the current through *different* surfaces bounded by the *same* closed curve C!

  As a challenge for physics or math majors (or others who just like a challenge!): From this one observation, plus your knowledge that *charge is conserved* (so that the net flow of charge out of any closed volume must equal the rate at which the charge inside that volume decreases in time:

  $$\frac{dQ}{dt} = -\oint_S \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA \tag{7.9}$$

  you should be able to *deduce* the necessity for an additional term known as *Maxwell's Displacement Current* which makes the total current *invariant* as we select surfaces bounded by the closed curve C to compute the current through.

  If you can do this on your own without looking ahead in the textbook, well, you're just a bit late for a Nobel prize, but this is the general idea for how you will eventually go about winning one: find an inconsistency in a physical theory and resolve it. Unify two fields! Explain something mysterious in a way that agrees with observation! You too can have your name on something one day...

- On a more mundane level, all students taking this course should learn to use Ampere's Law to find the magnetic field of any steady state: ***cylindrically symmetric current distribution, a (long) solenoid, a toroidal solenoid, and a plane sheet of current*** (and trivial integrable or summable variations of these four special geometries). These are all taught and reinforced in lecture, in the examples below, and in the homework (and in-class problems if your particular class is using active Team Based Learning).

- Useful **True Fact**: We do not usually deduce a *scalar* magnetic potential analogous to the electric potential, because magnetic fields do no work on isolated point charges, so our entire method for deriving a potential fails! In a future, more advanced, electrodynamics class physics majors will learn about a *vector* potential that leads to the magnetic field by virtue of a different form of multivariate *vector* differentiation (taking the "curl" of the vector potential) rather than taking the *gradient* of the potential as is the case for finding the electric field from the electric potential. This is why we will stop with "direct" evaluation of the magnetic field via the Biot-Savart law or via Ampere's Law for the cases where we can exploit symmetry to make the evaluation easy (enough) to teach the laws without undue mathematical hardship and omit any further discussion of magnetic potentials in this textbook.

7.1: Gauss's Law for Magnetism

At this point we know a rather lot about the magnetic field. We know that *moving charges* experience a magnetic force when they move through a magnetic field, and we further know that that force is "odd" compared at least to the Coulomb electrostatic force which (like gravity) acted on the "right line" connecting two charges. As you can see (looking over the chapter summary above), there are *other* surprises associated with the magnetic field waiting to be

learned, some of which will eventually lead us toward the theory of special relativity and more! For now, though, it is time to search for the *sources* of the magnetic field.

It is perfectly reasonable to begin our search by saying to ourselves: "Gee, we just spent all of this time learning about electrostatic fields coupled to monopolar electrical charges that behave like $k_e q_e/r^2$. I know about the gravitational field too, which behaves like Gm/r^2. Is it just barely possible that there is a quantity that behaves like a gravitational mass or an electrical monopolar charge that is similarly a source of the magnetic field so that:

$$\vec{\boldsymbol{B}}(\vec{\boldsymbol{r}}) = k_m \frac{q_m}{r^2}\hat{\boldsymbol{r}} \qquad (???) \tag{7.10}$$

for a magnetic point charge located at the origin?"

If there is, then we would expect the field of a collection of "magnetic monopolar" charges q_m[99] to be given by:

$$\vec{\boldsymbol{B}}(\vec{\boldsymbol{r}}) = \sum_i \frac{k_m q_{mi}(\vec{\boldsymbol{r}} - \vec{\boldsymbol{r}}_i)}{|\vec{\boldsymbol{r}} - \vec{\boldsymbol{r}}_i|^3} \tag{7.11}$$

and the magnetic force between a pair of monopoles to be given by:

$$\vec{\boldsymbol{F}}_{12} = \frac{k_m q_{m1} q_{m2}(\vec{\boldsymbol{r}}_1 - \vec{\boldsymbol{r}}_2)}{|\vec{\boldsymbol{r}}_1 - \vec{\boldsymbol{r}}_2|^3} \tag{7.12}$$

and so on and so forth (magnetic potential energy, magnetic potential, etc), where I've introduced a magnetic force constant k_m equivalent to k_e to set an SI scale for the units of magnetic charge and force.

If monopoles such as these existed, clearly I could derive a *Gauss's Law for Magnetism*:

$$\oint_S \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA = 4\pi k_m Q_{m,\text{in S}} = \mu_0 \int_{V/S} \rho_m dV \tag{7.13}$$

proceding *exactly* as I did before for an isolated electrical charge! This would make the static electrical and magnetic fields, at least, identical to one another, and even would suggest that there would be a force on a magnetic charge moving in an *electrical* field that has that pesky velocity dependent cross product in it, to maintain the symmetry even further.

As we'll see later, we even know what the magnetic field constant k_m would have to be. It is:

$$k_m = 1.00000... \times 10^{-7} \text{tesla} - \text{meter/ampere} \tag{7.14}$$

exactly (exactly because it defines the coulomb, not the other way around) as was determined and defined by Ampere in his experiments on magnetism!

At the time of Maxwell and for twenty or so years afterwards, the lack of any evidence for monopoles was simply an accepted fact. Pierre Curie, on the other hand, pointed out in the 1880's that Maxwell's equations could be made *consistent* with magnetic monopoles, so

[99] I meditated for quite a time what symbol to use for magnetic charge in this book. There are no particularly good choices. The one I initially leaned towards is g, which is sort of like a q but backwards, but this conflicts with the gravitational field. I finally went with q_m, even though this will require me to sometimes refer to *electrical* charge as q_e when I'm discussing the two kinds of charge together. This is tedious, however, in the long run, so be warned: q by itself will generally refer to electrical charge; I will always add the subscript m when discussing magnetic monopoles.

they *might* exist, but again lacking evidence that they did, nobody took the possibility seriously. Finally, in 1931 Paul Dirac published a theoretical demonstration that *if* at least *one* magnetic monopole existed in the accessible Universe surrounding a point electric charge, it would explain charge quantization, something that at the time was a complete mystery.

Well, dangle bait like that in front of a bunch of physicists and they'll be haring off to the laboratory to search for magnetic monopoles, visions of Nobel Prizes and trips to Stockholm to meet the king dancing through their minds. For at least 90 years at this point, intense effort has been expended searching experimentally for magnetic monopoles using a variety of ingeneous methods. This search *aggressively continues* to this day since magnetic monopoles turn out to be more or less *required* for most of the current "grand unified theories" (GUTs) or "theories of everything" (TOEs) to be consistent. In particular, they would have needed to be present in the early Universe during the time electromagnetic charges "condensed" out of the unified-field "soup" that is hypothesized to have existed at the earliest moments of the Big Bang. It is *still* a more or less guaranteed Nobel Prize to the researcher who first finds them in an experimentally reproducible way[100].

Alas, no *isolated magnetic monopoles* have been experimentally observed, in spite of an electromagnetic theory (and GUTs, and TOEs) that "beg" for them. Physicists would *love* for at least one magnetic monopole to exist in the Universe because if it did, quantum theory could explain charge quantization and much more! However, given the lack of concrete evidence for their existence at this point, it is probably safe to say that magnetic monopoles are at the very least *rare*. We express this lack of discoverable monopoles (so far) in Gauss's Law for Magnetostatics (GLM) as:

$$\boxed{\oint_S \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA = 4\pi k_m Q_{m,\text{in S}} = \mu_0 \int_{V/S} \rho_m dV = 0} \tag{7.15}$$

and this is just the way that you should learn it for this course.

Believe it or not, this is yet another one of *Maxwell's equations*, and we need to learn this equation just as well as we learn its electrostatic equivalent, Gauss's Law for Electrostatics (GLE). It actually tells us some *very useful things* about the magnetostatic field. In vector differential form (something you will learn later, if you continue on in physics) it is a key differential equation that you will need to be able to solve field problems. In *this* class, its implications can be summarized as:

- Magnetic field lines *cannot* begin or end at a point (recall that they could only begin or end at a point for electric field lines if the point contained an *electric charge*). Nor can they cross. This leaves only one alternative:
- Magnetic field lines must form *closed loops*.
- As we'll shortly see, those closed loops must be caused by something. That something is *moving charge passing through the loops*, at least for the next two chapters.

[100]Personally, I think the only places they are likely to find them are at the bottom of gravitational wells – the centers of asteroids or Lagrange points, given that the centers of planets are inaccessible. Isolated magnetic monopoles should not bind to ordinary charged matter although they would strongly interact with it, so they should diffuse "down" if ever brought into equilibrium with ordinary matter.

To repeat: Gauss's Law with no monopoles is an *empirical* rule, and lack of evidence isn't positive evidence of lack! We don't know if there are, or are not, magnetic monopoles somewhere in the Universe; we only know that we haven't seen any *so far* when we've looked for them quite hard. At any moment, though, a reproducible experiment that observed them would require us to modify GLM (as well as Faraday's Law from the next chapter) to include monopolar terms and we'd all have to work a bit harder to learn electrodynamics. This would be well worth the price, however, as it would enable us to understand why charge is quantized, it would make various TOEs more consistent, and besides, they are so *rare* that they would not practically complicate problem solving in *introductory* physics classes, however much physics *graduate* students might be tortured...

7.1.1: The Units of Magnetic Flux

GLM above contains an integral that amounts to the flux of the magnetic field through a surface. We will encounter this kind of flux again in Faraday's Law covered in the next chapter, so we will at this point explicitly define the magnetic flux through a surface S and the SI units associated it.

The definition of magnetic flux through a surface S is (as should already be clear):

$$\phi_m = \int_S \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA \tag{7.16}$$

(where in context we might well omit the m subscript). Its SI units are called *Webers*, where 1 Weber is one Volt-Second or one Joule/Ampere or 1 Tesla-Meter^2, as you prefer.

There. That's done. From a practical point of view, we will almost never use Webers as units per se, as we will work directly with the equations in which magnetic flux occurs to get quantities that are of a lot more directly useful to us, for the most part. You should still know what they are.

7.2: The Biot-Savart Law

In the previous section we tried to generalize Gauss's Law for Electrostatics into a Gauss's Law for Magnetostatics, where static magnetic fields could be created by magnetic "charges" (magnetic monopoles) much the same way that static electric fields are created by electric charges. Using our imagination, we readily succeeded, but alas when we went out into the world to search for magnetic charges we didn't find any. Yet magnetic fields exist in abundance; otherwise how could pictures and newspaper articles ever be stuck to our refrigerator doors?

When we go to search for sources of these magnetic fields, we find that they all have something in common: The static magnetic fields we can readily generate and observe in the lab are created by *moving electrical charges*, usually in the form of a static (or slowly varying) *electrical current in a wire*[101].

[101] Reality is slightly more complicated than that, however. For one thing, as we discussed in the last chapter, point-like electrons (and several other elementary particles) and many nuclei have a magnetic dipole moment due to their quantum "spin" even though this spin can *not* be correctly modeled as a moving/rotating electrical charge

As it turns out, the magnetic field produced by a *single* charged particle has proven to be very interesting – *too* interesting, in fact, for us to give a really complete treatment of it in an introductory course. We will therefore defer a full non-relativistic discussion of the *origins* of the magnetic field of a moving point charge at least until we have completed our discussion of the Maxwell Displacement Current in a few weeks.

Doing the field of currents per se first actually recapitulates the historical order of things. In 1819 Øersted discovered that the currents in a wire connecting two poles of a battery (invented by Alessandro Volta in 1800 – hence the SI unit *volt*) would cause the deflection of a magnetized compass needle. One year later André-Marie Ampere discovered that two current carrying wires exerted a force on one another and published his findings, within *a week* of when Jean-Baptiste Biot and Flix Savart studied the damped oscillations of a compass needle deflected by a wire to conclude what amounts to the same result for the causal "magnetic field" produced by a wire. At this time the nature of the charged "fluid" flowing through matter was almost entirely unknown!

This situation lasted almost seventy years! The magnetic fields produced by individual charged particles were beyond the experimental reach of eighteenth and nineteenth century physicists until the very end of the nineteenth century. Between 1838 and 1874, various models were proposed for atoms, some of which did involve pointlike charged particles and "cathode rays" produced by hot negatively charged wires were hypothesized to be one of the particles, but it wasn't until 1887 that the charge-to-mass ratio of the electron was first measured by J. J. Thomson and the electron as a pointlike charged particle was officially "discovered". The value of the e (and hence m_e) was not measured until 1909 by Robert A. Millikan, although George Johnstone Stoney deduced the value of e (correctly) from experiments in *chemistry* without assigning it to any particular particle other than atoms themselves way back in 1874. Finally, it wasn't until 1908 when the atomic nucleus was discovered by Ernest Rutherford!

In other words, it wasn't until the *early 20th century* that a reasonably accurate model for the atom as a building block for molecules and ordinary matter was developed!

Yet most of the electromagnetism we learn in this book was discovered and more or less "finished" as far as its fundamental equations are concerned (through Poynting's Theorem, the complete "work-energy-momentum" equation of the electromagnetic field) by 1884, so it is clear that we don't *really* need to understand the magnetic field of a point particle to understand how to make a magnetic field at least in the non-relativistic limit of more or less continuous currents in e.g. wires. Relativity theory itself was only developed in the late nineteenth/early twentieth century, largely to help make electromagnetism *consistent*!

As the short history lesson above suggests, the original magnetic fields studied were generated one of two ways:

- They were "natural" fields generated by magnetic *objects*: Bar magnets, compass needles, magnetite mineral chunks, the Earth itself. These natural magnets and the forces they generate have been known almost from prehistoric times – it is probable that primitive compasses were used in China and in India as long ago as 1000 BCE, and a Greek

with finite extent. For another, as we shall see in the next chapter, *changing electric fields make magnetic fields* even in the *absence* of moving electrical charges. Still, we will be able to understand *both* phenomena in terms of electrical currents at least at first.

legend dates it as much as 1000 years before that. The ancient Olmec culture in the Americas likely had discovered magnetism by 1000 BCE as well, as some of the artifacts from their culture are strongly magnetized. These fields were not really *studied*, but they were used to *engineer* compasses and curios.

- Beyond this, the original *controlled, artificial* magnetic fields that *were* systematically studied after the Enlightenment (that is, with the scientific method and producing published results) were all generated by *current carrying wires* in the laboratory, under human-controlled conditions.

The results of the early experiments by Biot and Savart (who used their apparatus to literally map out the field strength and direction near the wire) can be summarized as follows:

- The magnetic field produced by a small (differentially short) segment of wire carrying an electrical current is proportional to the current (and by implication, proportional to the length of the segment).
- The magnetic field of the segment diminishes like the distance from the charge to the point of observation squared, just as was the case of the electrostatic field.
- The *direction* of the magnetic field produced by that same current-carrying segment is that it forms *circular loops* around the direction of the current, in the direction the fingers of your right hand curl around your thumb if you line your thumb up along with the direction of the current in the segment.
- For any given value of the current, the field strength produced by a differential chunk of the wire is proportional to $\sin\theta$ where θ is the angle between the direction of $\vec{v}$ and the direction of I in the chunk.

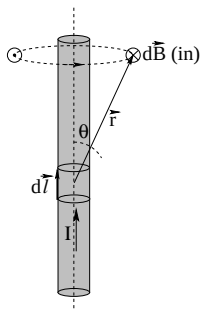


Figure 7.1: The geometry represented by the observations of Biot and Savart.

The geometry of $d\vec{\boldsymbol{\ell}}$ and the resulting magnetic field is illustrated in figure 7.1, where for simplicity we center the coordinate frame on $d\vec{\boldsymbol{\ell}}$ in such a way that the z-axis of a spherical polar coordinate frame is aligned with the direction of the current. We can transform the list of observations of Biot and Savart into the following formula. If we let $d\ell$ be the length of the segment and give it the direction of the current I in the wire the segment is a part of, then:

$$\boxed{d\vec{\boldsymbol{B}} = k_m \frac{I(d\vec{\boldsymbol{\ell}} \times \hat{\boldsymbol{r}})}{r^2}} \tag{7.17}$$

should be the magnetic field of *just that small segment!*

This result is not particularly general. First, the wire in question is straight; real wires can easily be bent into curves. We can solve this by using the superposition principle, and this is the main reason we insisted on finding the field of a *differentially short* segment of wire, so we can integrate over an arbitrary wire and thereby add up the field (in more general coordinates). Also note that this expression *makes no sense in isolation* – because of *charge conservation*, we can *never* observe the magnetic field of an *actual* microscopic segment in isolation because charge cannot be created at one end and destroyed at the other – the current has to get to the ends of the segment *through more wire!*

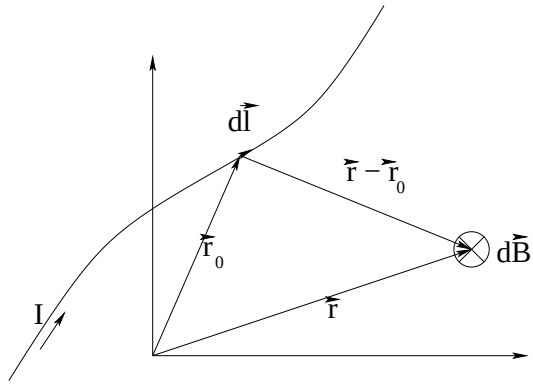


Figure 7.2: The Biot-Savart Law in general coordinates.

We will usually write this more properly in more general coordinates instead of coordinates that centered (as these ones did) on the chunk $d\vec{\ell}$, and for a possibly curved conductor instead of just a straight one. The result (and its associated figure 7.2 becomes:

$$\boxed{d\vec{\boldsymbol{B}} = k_m \frac{I(d\vec{\boldsymbol{\ell}} \times (\vec{\boldsymbol{r}} - \vec{\boldsymbol{r}}_0))}{|r - r_0|^3} = \frac{\mu_0}{4\pi} \frac{I(d\vec{\boldsymbol{\ell}} \times (\vec{\boldsymbol{r}} - \vec{\boldsymbol{r}}_0))}{|r - r_0|^3}} \tag{7.18}$$

which is known as the *Biot-Savart Law* (Bee-oh Sa-var Law).

The clever and observant student will have noticed that k_m appearing in this expression looks *exactly* like the k_m I introduced to scale the magnetic field to bare magnetic monopolar charge (if it were to exist). In fact, the two constants are the same! We are still using

$$k_m = \frac{\mu_0}{4\pi} = 10^{-7} \text{ tesla-meter/ampere } \textit{exactly}$$

although the value of this constant was set not by Biot and Savart, but by Ampere. I have also taken the liberty of establishing early on the relationship between this "magnetic constant" (equivalent to the electric constant k_e) and

$$\mu_0 = 4\pi k_m = 4\pi \times 10^{-7} \text{ tesla-meter/ampere},$$

a new quantity known as the ***permeability of free space***, which will be the magnetic equivalent of ϵ_0, the *permittivity* of free space. As before, the electric and magnetic constants are easy to remember, while the permittivity and permeability are easy to remember numbers multiplied or divided by 4π which are *not* particularly easy to remember.

Wow! That looks a lot more complicated than our integral expressions for the electrostatic field! And so it is... we will have to work much harder to evaluate the magnetic field directly

from a current (distribution), and the need to do twisty integrals of directed differential vectors as they are worked along a curve *and* formed into a cross product with a relative vector to the point of observation (in some system of coordinates) will severely limit the problems we can solve analytically by just doing sufficiently straightfoward integrals.

This is good news and bad news. From the student's point of view, it means that things at the intro level are relatively easy. If you learn to do all of the examples presented in the textbook, the homework, and any in-class examples or problems, well, that is *close* to all of the examples anyone (including, very likely, your instructor) can do without being an integration *god*[102]. On the other hand, it presages bad things for the more advanced student, where sooner or later some of the more difficult problems must be faced (at which point you will need to have made some progress on the road to calculus-deity or have learned to integrate complicated functions numerically on a computer).

For this course, we will cheerfully take the easy road and work through a nice set of relatively simple examples that make up, as promised, most of or close to *all* of what one can do, period, before things get very complicated indeed.

7.3: Examples of Using the Biot-Savart Law to Find the Magnetic Field

Example 7.3.1: Magnetic Field of a Straight Wire Segment

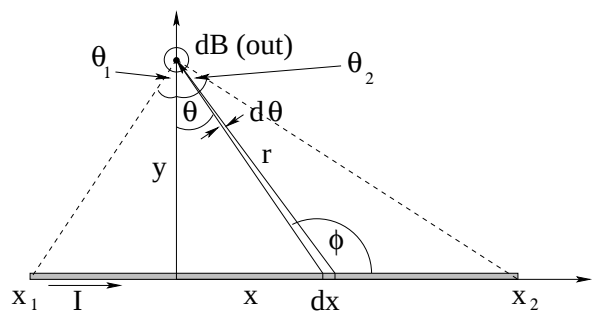


Figure 7.3: The geometry and coordinates that make it simplest to evaluate the magnetic field of a straight segment of wire carrying a current I.

In figure 7.3 we see the geometry of a single, straight, segment of wire relative to an arbitrary point in space. We wish to use the Biot-Savart Law to find the field of this wire at the point on the y-axis indicated. Note well that this is a *general point* as the y-axis itself is located at a general point on the x-axis, and no matter where we locate it we can do an integral between x_1 and x_2.

We therefore begin by considering the geometry of the cross product. If we let the fingers of our right hand line up with dx in the direction of I and rotate through the small angle to line up with the vector $\vec{r}$ between dx and the point of observation, our thumb picks the perpendicular direction *out* of the page as indicated. We can thus easily write the magnitude of the magnitude

[102]These do exist, by the way, some of them wandering in the halls of physics departments. Be warned.

for the field produced by this small differential chunk of the current as:

$$dB = \frac{k_m I \sin(\phi)\ dx}{r^2} \tag{7.19}$$

and hence (formally integrating both sides) we get:

$$B = \int_{x_1}^{x_2} \frac{k_m I \sin(\phi)\ dx}{r^2} \tag{7.20}$$

Alas, we have an embarrassment of variables in this. As we vary x in the integral, both r and ϕ vary! To do the integral we have to use *exactly* the same methodology we used to evaluate the *electric* field of a straight line of *charge*. We change variables, in other words, so that they are consistently all (r, θ). That is:

$$\begin{aligned} x &= y \tan(\theta) \\ dx &= \frac{y\ d\theta}{\cos^2(\theta)} \end{aligned} \tag{7.21}$$

and

$$\sin(\phi) = \cos(\theta) \tag{7.22}$$

(think about it for a minute, it will make sense). The integral becomes:

$$B = \int_{\theta_1}^{\theta_2} \frac{k_m I y \cos(\theta)\ d\theta}{\cos^2(\theta) r^2} = \int_{\theta_1}^{\theta_2} \frac{k_m I y \cos(\theta)\ d\theta}{y^2} \tag{7.23}$$

(using $y = r\cos(\theta)$) and we finally obtain:

$$\begin{aligned} B &= \frac{k_m I}{y} \int_{\theta_1}^{\theta_2} \cos(\theta)\ d\theta \\ &= \frac{k_m I}{y} \left(\sin(\theta_2) - \sin(\theta_1)\right) \end{aligned} \tag{7.24}$$

(out of the page, as noted). This is almost exactly identical to our expression for the electric field of a long straight line of charge as evaluated in week 2, so it should be easy enough to remember or rederive.

As was the case then as well, we can find the magnetic field produced by an *infinite* straight line of current by taking the limits $\theta_1 \to -\pi/2$ and $\theta_2 \to \pi/2$, where the sines become -1 and 1 respectively. Note that this result will actually be relevant and useful any time we seek the field *close enough* to a wire carrying current that the angles to the end points approach $\pi/2$. The field of such an "infinite" wire is just:

$$B_\infty = \frac{2k_m I}{y} \tag{7.25}$$

Again note the analogy with electric field, with $k_e \to k_m$ and $\lambda \to I$, but note well, the *geometry* of the field is entirely different! The magnetic field, for a finite or infinite wire carrying current, flows in *circular loops around the wire!* In our picture, the field goes out of the page above the wire, into the page below the wire, and in general **if we let the thumb of our right hand line up with the direction of the current in the wire, the field circulates around the wire in the same sense as our fingers**.

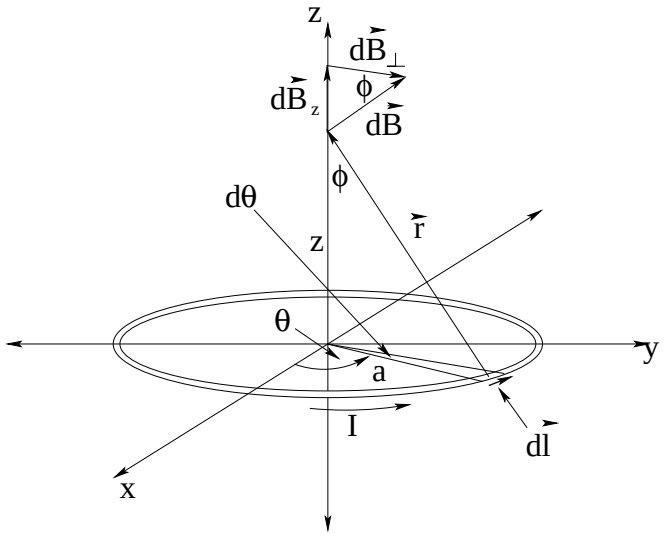


Figure 7.4: The geometry and coordinates used to compute the magnetic field of a circular loop wire carrying a current I on its axis of symmetry.

Example 7.3.2: Field of a Circular Loop on its Axis

In the figure above, we have the geometry of a circular current loop in the xy-plane. Finding the magnetic field of this loop at an arbitrary point in (say) spherical polar coordinates is not impossible, but neither is it easy – it is a chore best left for your next course (if any) in electromagnetism. In *this* course, however, we can easily find the field at an arbitrary point on the z-axis, because there we can use the *cylindrical symmetry* of the arrangement to our advantage.

We begin by writing the Biot-Savart Law for the small chunk of current in the segment of the wire labelled $d\vec{\boldsymbol{l}}$:

$$d\vec{\boldsymbol{B}} = k_m \frac{I d\vec{\boldsymbol{l}} \times \hat{\boldsymbol{r}}}{r^2} \tag{7.26}$$

As you can see, the direction of this infinitesimal field element is in the plane formed by $\vec{\boldsymbol{r}}$ and the z-axis, perpendicular to $\vec{\boldsymbol{r}}$ in the right-handed direction. The magnitude of this field element is:

$$dB = k_m \frac{I dl}{r^2} \tag{7.27}$$

We can easily find the components of $d\vec{\boldsymbol{B}}$ parallel and perpendicular to z using the angle ϕ:

$$dB_z = k_m \frac{I dl}{r^2} \sin(\phi) \tag{7.28}$$

$$dB_\perp = k_m \frac{I dl}{r^2} \cos(\phi) \tag{7.29}$$

We can evaluate the two trig functions using the right triangle with sides of a, z, and r (which has the same angle ϕ in its apex) – $\sin(\phi) = a/r$ and $\cos(\phi) = z/r$:

$$dB_z = k_m \frac{I dl\ a}{r^3} \tag{7.30}$$

$$dB_\perp = k_m \frac{I dl\ z}{r^3} \tag{7.31}$$

At this point we could be lazy and invoke symmetry. The problem has *azimuthal symmetry* – if we walk around the ring and look at it from arbitrary angles, the problem does not change with our perspective, so we know that the *total magnetic field cannot have a component that changes as we walk*, that is, one in the x or y direction. The field can point in the z direction only on the z-axis. This allows us to evaluate B_z only to get the total field.

However, one day you might need to *show* that the $\perp$ field vanishes the *hard way* by actually integrating it. Fortunately, this really isn't that hard. If you look carefully at the picture, you can see that:

$$dB_z = k_m \frac{Idl\ a}{r^3} \tag{7.32}$$

$$dB_x = k_m \frac{Idl\ z}{r^3} \cos(\theta) \tag{7.33}$$

$$dB_y = k_m \frac{Idl\ z}{r^3} \sin(\theta) \tag{7.34}$$

The only thing remaining is a variable we can integrate over. Hopefully it is obvious that integrating over x and y is a really bad choice, while integrating over θ is a good one. We note that $dl = ad\theta$, substitute this in, and we are ready to go:

$$\begin{aligned} B_z &= k_m \int_0^{2\pi} \frac{Ia^2 d\theta}{r^3} \\ &= k_m \frac{I2\pi a^2}{r^3} \end{aligned} \tag{7.35}$$

$$B_x = k_m \int_0^{2\pi} \frac{I\ az}{r^3} \cos(\theta)\ d\theta = 0 \tag{7.36}$$

$$B_y = k_m \int_0^{2\pi} \frac{I\ az}{r^3} \sin(\theta)\ d\theta = 0 \tag{7.37}$$

We conclude that:

$$\vec{\boldsymbol{B}} = k_m \frac{I2\pi a^2}{(a^2+z^2)^{3/2}} \hat{\boldsymbol{z}} \tag{7.38}$$

It is instructive to write this in terms of the *magnetic moment* of the loop, $\vec{\boldsymbol{m}} = I\pi a^2 \hat{\boldsymbol{z}}$:

$$\vec{\boldsymbol{B}} = \frac{2k_m \vec{\boldsymbol{m}}}{(a^2+z^2)^{3/2}} \tag{7.39}$$

which is *exactly the same form* as that of the electric field on the axis of an electric dipole, $\vec{\boldsymbol{E}} = 2k_e \vec{\boldsymbol{p}}/(a^2+z^2)^{3/2}$, that we derived several weeks ago, with the substitution of k_m for k_e and $\vec{\boldsymbol{p}}$ for $\vec{\boldsymbol{m}}$. This (hopefully) continues to motivate the idea that electric and magnetic fields have certain *characteristic shapes* – those of monopoles, dipoles, quadrupoles, and so on – and that if we ever learn to evaluate their *multipolar moments* for arbitrary charge-current distributions, we will be able to easily reconstruct at least a good approximation to the total electromagnetic field of those distributions.

In that spirit, we can easily find the form of the field when $z \gg a$:

$$\vec{\boldsymbol{B}} = \frac{2k_m \vec{\boldsymbol{m}}}{z^3} \tag{7.40}$$

where we used the binomial expansion, sort of – we only had to keep the leading term after factoring out the z so it was pretty easy.

Evaluating the magnetic field using the Biot-Savart Law becomes increasingly difficult from here on. At the very least, it becomes an exercise in increasingly difficult *calculus*, even though the physical *concept* is the same and you can always write down an integral that – if you could do it – would lead you to the answer. There is one more worth at least laying out to help get you set up for your homework.

Example 7.3.3: Field of a Revolving Ring of Charge on its Axis

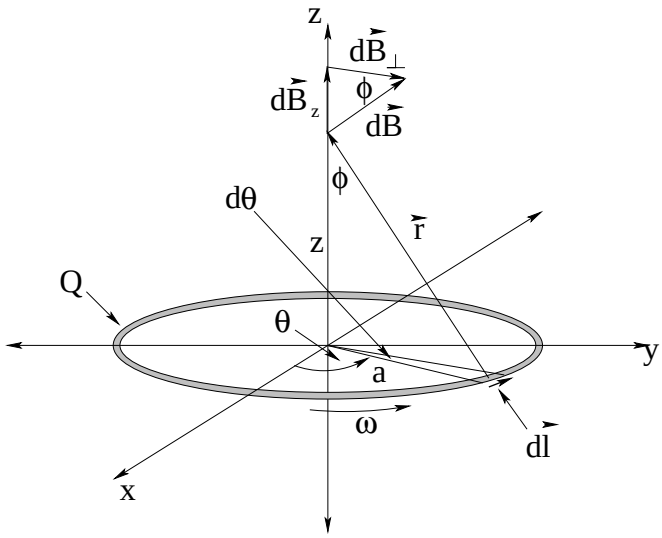


Figure 7.5: The geometry and coordinates used to compute the magnetic field on the axis of symmetry of a circular ring of charge Q revolving at angular velocity Ω.

In the figure above, a circular ring of charge with charge Q, radius a, and angular velocity Ω (which really points in the vector z-direction, recall – I'm just indicating the direction of rotation in the figure above) is in the xy-plane concentric with the z-axis. Our job is to find the field on the z-axis once again.

If this figure reminds you of the one in the last section, it should – they are the *same*. In fact, the solution is going to be the same, except that we have to figure out the *current* in the case where we have a revolving ring of charge instead of an actual current in a wire. To do so, we note that all of the charge in the ring moves past an arbitrary point on the circle of its motion – say, where it crosses the x-axis – in one period of its revolution. The total charge per unit time passing that point is thus:

$$I = \frac{Q}{T} = \frac{Q\Omega}{2\pi} \tag{7.41}$$

The field is thus obtained by doing the exact same integrals as before:

$$\vec{\boldsymbol{B}} = k_m \frac{I 2\pi a^2}{(a^2+z^2)^{3/2}}\hat{\boldsymbol{z}} = k_m \frac{Q\Omega a^2}{(a^2+z^2)^{3/2}}\hat{\boldsymbol{z}} \tag{7.42}$$

The main reason to do this here (since it no doubt seems trivial) is that it is a key step along the way to finding the magnetic field of a rotating *disk* of charge Q and radius R on your homework. On your homework problem you will want to draw the disk of charge and select out

a thin ring of that charge of radius r and thickness dr. The field of this rotating ring will depend on r (which is the same as a in this example, the radius of the ring, not the distance from the ring to the point z). With a bit of care, you can integrate B_z over r to find the total field on the z axis. Then you can investigate the $z \gg R$ limit and (with luck and the use of expressions you derived for m of a rotating disk in the last chapter) show that it is still $k_m 2\vec{\boldsymbol{m}}/z^3$ in this much more complicated case.

We aren't quite "done" with Biot-Savart. There are a few problems I could reasonably give you and expect you to be able to at least formulate them as integrals and – with a bit of skill – integrate to find the total field. Some of them are on your homework, but you can imagine others – the field on the axis of a rotating rod of charge. The field on the axis of a solenoid. The field of a rotating sphere, or spherical shell of charge. Some of these may seem daunting, but in all cases with a bit of work you could get them.

However, learning to do these increasingly difficult *integrals* won't teach you the *physics* any better. To get a better grip on the physics, we have to leave the Biot-Savart Law behind, or better yet, convert it into a more general equation the same way that we converted the field of a point charge into Gauss's Law for Electrostatics. In the next section we will do just that – we will turn the Biot-Savart Law into our next Maxwell equation, *Ampere's Law*.

But first, let's have a discussion on the magnetic field produced by "slowly" moving point charges. As promised, this section will be very short, because we will come back to it later in the textbook armed with Ampere's Law, but there are some *really interesting aspects* of magnetism that one first confronts when thinking about the magnetic field of a moving point charge, and if nothing else, they'll give you something to think about in the weeks ahead.

7.4: The Magnetic Field of a Point Charge

The Biot-Savart Law looks as though it is *begging* to be turned into a very simple rule for finding the magnetic field produced by a point charge. When we consider the geometry of a short segment of wire with length $d\ell$ and cross-sectional area A, we learned that:

$$I = nqv_d A \quad \Rightarrow \quad Id\ell = nqv_d(Ad\ell) = nq(Ad\ell)v_d$$

where $Ad\ell$ is the *volume* of the small chunk of wire, so:

$$\Delta Q = nq(Ad\ell) \quad \Rightarrow \quad Id\ell = \Delta Q v_d$$

In words, the current times the small length $d\ell$ equals the free conduction charge in that small chunk times its drift velocity.

It seems pretty *reasonable* to think that the field of the point-like segment, then, equals the field produced by the point-like charge ΔQ *of* the segment times the actual (classical) speed the charge is moving with down the wire in the direction of the current! In that case we would expect:

$$d\vec{\boldsymbol{B}} = k_m \frac{I(d\vec{\boldsymbol{\ell}} \times \hat{\boldsymbol{r}})}{r^2} \quad \Leftrightarrow \quad \vec{\boldsymbol{B}} = k_m \frac{q\,(\vec{\boldsymbol{v}} \times \hat{\boldsymbol{r}})}{r^2} \tag{7.43}$$

should be the "Biot-Savart Law" magnetic field of a point charge q moving at a velocity $\vec{\boldsymbol{v}}$! In fact, this expression is approximately correct, but *only if the charge q is moving at a speed*

$v \ll c$*, slowly relative to the speed of light*, and *only if* r *is "small enough"* that we can treat the emission of the field and its observation to be effectively "instantaneous".

Let's try to understand this, as it is one of the things that strongly motivates the later development of *the theory of special relativity*. Let's start with the geometry of the field lines we

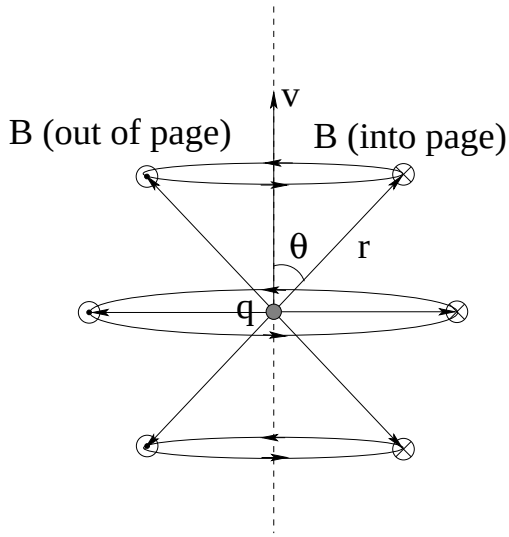


Figure 7.6: The geometry of the magnetic field lines going in circles *around* the (dotted) line of motion of the charge in the right-handed sense. Note that the direction of $\vec{v} \times \vec{r}$ is into the paper on the right, out of the paper on the left, for $\vec{v}$ and any of the $\vec{r}$ vectors shown.

expect for the point charge, as shown in figure 7.6. In spherical polar coordinates centered on the charge with the z-axis aligned with its direction of motion, the cross product simplifies to give us a field magnitude

$$B = k_m \frac{qv}{r^2} \sin\theta \qquad (v \ll c) \text{ only.} \tag{7.44}$$

The direction of the field is obtained as usual from the RHR – if you let your right-handed thumb point in the direction of motion of the charge, your fingers will curl around your thumb in the same sense that the field lines are directed around the velocity vector of the charge in an axially symmetric way.

7.4.1: Finite Field Propagation Speed for E and B

This rule is simple enough, and is *almost* the rule you will learn in much more advanced courses in electrodynamics than this one. The only thing we are leaving out (that is, of course, very important) is that neither the electric nor the magnetic field appears instantaneously in all space. When one of their sources is "turned on" by a suitable rearrangement or motion of charges, the fields propagate outward from the charge at the speed of light, establishing their value at a point of observation point at a lagged time *after* the charge or current appears at any given point of field emission.

This is true even for the apparently "static" electric field we worked on in the first chapter! For this field evaluated "near" a more or less stationary (slowly moving) electric charge this didn't matter much, but we still called the field the electro*static* field to emphasize the requirement. For the magnetic field, things are still more surprising! Let's consider a very simple

example. Suppose you have a point charge at rest at the origin. As we can see from the expression above, it should have no magnetic field at all, and should produce an electrostatic field at the point $\vec{r}$ (as usual) of:

$$\vec{E}_{\text{rest}} = \frac{k_e q}{r^2}$$

Now imagine that we *change reference frames to a new inertial reference frame* moving a velocity $\vec{v} = -v_0\hat{z}$ relative to the first frame. In this frame, q is moving up with speed $v_0\hat{z}$ at $t = 0$, the time it is located at the common origin of both frames! At that instant and in this moving frame there are *both* electric *and* magnetic fields at the point $\vec{r} = \vec{r}\,'$!

This results in a complete mess! For example – if we have a *neutral* current-carrying wire with a segment at $\vec{r}$, it will experience no electric force or magnetic force in the first frame. In the second, it will *still* experience no electric force, but (depending on the current direction relative to $\vec{v}$) it may well experience a magnetic force in the second frame! This isn't the appearance or disappearance of a *pseudo*force in an *accelerating* frame, this is a serious blow to either Newton's Laws, or to Galilean relativity, as the *actual* force in the two frames changes so that there are *different accelerations* depending on which frame you are in! This makes no sense, of course. The *trajectories* observed must be the same in both frames, within the *kinematics* of the transformation between frames!

It will take a bit of work, in some future E&M course, to see that the fundamental problem is with the galilean frame transformation itself and that the theory that consistently permits it all to work out is the *theory of special relativity*. In fact, *both* $\vec{E}'$ *and* $\vec{B}'$ in the moving frame differ from what they are in the original stationary charge-centered frame – and so are space and time – but in just the right way that ultimately, observers in the two frames don't see fundamentally different realities, but just two different ways of observing just one reality!

7.4.2: Violation of Newton's Third Law

Another problem that we can understand right *now* follows from combining this empirical law for building the field with the Lorentz force law for the magnetic force on a moving charge. Put them together and we find that for two charges q_1 and q_2 travelling at velocities $\vec{v}_1$ and $\vec{v}_2$ and with $\vec{r}_{12}$ the vector from q_1 to q_2, the force on q_1 due to q_2 is:

$$\vec{F}_{12} = -q_1\vec{v}_1 \times k_m q_2 \left(\frac{\vec{v}_2 \times \hat{r}_{12}}{r_{12}^2}\right) = -\frac{k_m q_1 q_2}{r_{12}^2}\left(\vec{v}_1 \times (\vec{v}_2 \times \hat{r}_{12})\right) \tag{7.45}$$

Similarly, the force on q_2 due to q_1 is:

$$\vec{F}_{21} = q_2\vec{v}_2 \times k_m q_1 \left(\frac{\vec{v}_1 \times \hat{r}_{12}}{r_{12}^2}\right) = \frac{k_m q_1 q_2}{r_{12}^2}\left(\vec{v}_2 \times (\vec{v}_1 \times \hat{r}_{12})\right) \tag{7.46}$$

There is just one wee problem with this result. $\vec{F}_{12} \neq \vec{F}_{21}$! Newton's Third Law has just bitten the dust, never to return! It is *not correct* the way it is usually taught, and its failure has profound implications! In case the inequality of these two terms (except for a minus sign) isn't obvious, consider the geometry in figure 7.7: As you can easily see by inspection, the magnetic field at the position of q_2 in this figure is zero, because $\vec{v}_1$ is *parallel* to $\vec{r}_{12}$ (so the cross product is zero). However, $\vec{v}_2$ is perpendicular to $\vec{r}_{12}$ and the cross-product is not zero;

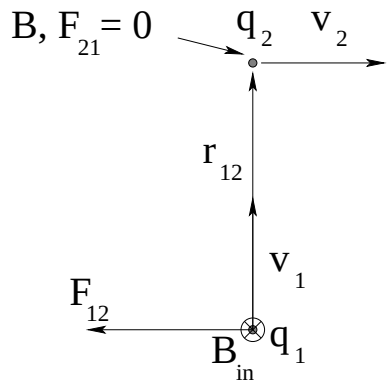


Figure 7.7: Newton's Third Law fails for this arrangement because the field (and hence force) at q_2 due to q_1 is *zero* while the field (and hence force) at q_1 due to q_2 is *not zero!*

the magnetic field at q_1 is nonzero and in to the page, perpendicular to $\vec{v}_1$. Consequently the force on q_1 is nonzero and points to the left while the force on q_2 is zero!

If you think for a moment, you will recall that we *used* Newton's Third Law to derive a very important physical principle: *The Law of Conservation of Momentum!* In the picture above, the total momentum of the interacting pair of particles is *changing in time*. If we worked harder and computed the way the total *energy* of the particle pair is changing as a function of time we would find that it is changing too. The pair of particles is literally "lifting itself up by its own bootstraps"!

This is, of course, offensive to all right minded individuals, who quite correctly view the failure of energy or momentum conservation to be the failure of all of physics, a global inconsistency that would cause us to observe all sorts of "magic" that we do not, in fact, observe in the world.

Historically, whenever momentum or energy have *appeared* to be lost in a collision, we have (when we looked carefully) *found* them again, often in an unexpected form. This case is no exception; in fact it was probably the original case of the rule. Physics is quite safe, because momentum *is* conserved, even though the momentum of a collection of massive particles interacting only electrically and magnetically is *not*! Where do you think the missing momentum and energy might be found?

7.5: Ampere's Law

It is difficult to know the best way to show you the path from the Biot-Savart Law to ***Ampere's Law*** or vice-versa. In a sense, this is conceptually one of the most difficult things to see, because the usual connection is established using multivariate calculus that is beyond the scope of this course, and it is *Ampere's Law* that is viewed as being the more fundamental!

The Biot-Savart Law *can* can be used to derive a limited "magnetostatic" version of Ampere's Law using relatively simple arguments that we will give below, but the only method I know of for deriving the Biot-Savart Law from Ampere's Law using more or less *elementary* calculus at the level of this course requires that we *start* from Ampere's law corrected by the "Maxwell Displacement Current", something we will not cover for two more chapters.

We will therefore defer until then a very interesting demonstration (due to Robert Buschauer) for obtaining the Biot-Savart Law from Ampere's law using only the displacement current of a point charge moving at speed $v \ll c$ along the z-axis b – a demonstration that also reinforces the idea that the magnetic field produced by a point charges in a frame where that charge is moving must be an *electric* field only in a frame where it is not – something that forces us to conclude that electromagnetism with the electric and magnetic fields represented by simple vectors is ***not invariant*** under the Galilean transformation between these two reference reference frames!

The point of this is that the Biot-Savart Law is itself not strictly correct outside of the context of magnetostatics where Ampere's Law – with Maxwell's eventual addition – is one of the two equations that lead us to *electrodynamics* – the fully unified dynamic electromagnetic field. As we've seen, it doesn't seem to work correctly for "isolated" current-carrying wire segments or their equivalent moving point charges.

In fact, my favorite way of presenting this whole chapter is as a sort of detective story, where I lay down hints along the way and give you a chance to win a prize[103]. The goal is to see the *flaw* in Ampere's Law as we soon write it, to see how it *must* fail to be mathematically consistent for certain geometries of currents, and – naturally – to correctly derive the *fix* for it: the Maxwell Displacement Current that (with Faraday's Law) unified Electricity and Magnetism.

Of course, any student who wishes can skip ahead a few chapters and "cheat", but that would be no more satisfying than reading the last chapter of a mystery novel first.

So come on. You're pretty smart. You're taking a no-kidding physics course. Think you can slam-dunk like James Clerk Maxwell? Then *bring it*. Figure out why Ampere's law isn't consistent and make it right, without peeking! If you can do that, you can do anything, and the knowledge that you can do anything is more valuable than you can imagine when later you hit some really difficult problems in life if not physics.

Thus we will start our discussion by thinking once again about a single, infinitely long, straight line of current I. We recall from our Biot-Savart Law based example problem done in the text above that:

$$B = \frac{2k_m I}{r} = \frac{\mu_0 I}{2\pi r} \tag{7.47}$$

where the direction of the magnetic field is around the current I in the right-handed sense. The geometry of this is drawn in figure 7.8.

Note well that the field *drops off like* $1/r$. The circumference of the circle just happens to *increase* like r. In week 3 we saw that if we multipled a field that went down like $1/r^2$ by an area that went up like r^2, we got a quantity that only depended on the charge (and Gauss's Law for Electrostatics). Here we can clearly do the same thing and write:

$$B \times 2\pi r = \frac{\mu_0 I}{2\pi r} \times 2\pi r = \mu_0 I \tag{7.48}$$

So far, this is only suggestive. However, consider the geometry of figure 7.9. where the

[103] A prize of no value whatsoever and of the greatest value you can imagine. In my own class, I up the ante of the "no value whatsoever" part and give any student that manages it a piece of candy or a small prize picked out of a treasure chest of cheap prizes. This, of course, makes the total value of the prize *greater* than the greatest value you can imagine, if you can imagine that. Of course – ontologically – you can't, hmmm...

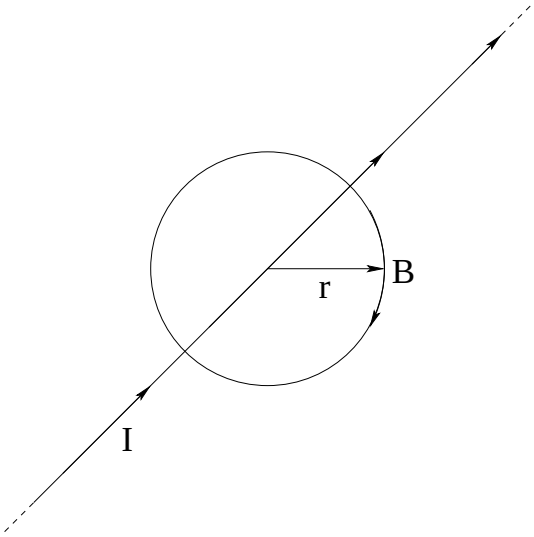


Figure 7.8: An infinitely long straight wire carries current I and has a magnetic field that goes *around* the wire in circular loops of constant magnitude in the right-handed sense.

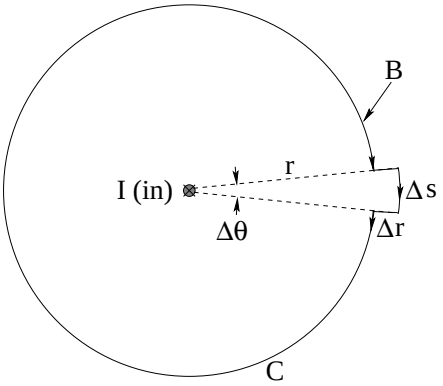


Figure 7.9: A circular path of radius r around a long straight wire with a "notch" of angular width $\Delta\theta$ and radius $r + \Delta r$.

current I is drawn directly into the page so we can concentrate on the plane in which the magnetic field lies.

The B-field is of course constant in magnitude on the circle of radius r as before, and so we can multiply it by the length of the circular arc at the same radius right up to the notch. However, our path now steps *out* by Δr along the radius (and perpendicular to the field). Along the curved path Δs, the field is somewhat weaker, but the path itself is somewhat longer.

In fact, we can cleverly add up the following:

$$\begin{aligned} B(2\pi - \Delta\theta)r + B\Delta s &= B(2\pi - \Delta\theta)r + B(r + \Delta r)\Delta\theta \\ &= \frac{\mu_0 I}{2\pi r}(2\pi - \Delta\theta)r + \frac{\mu_0 I}{2\pi(r + \Delta r)}(r + \Delta r)\Delta\theta \\ &= \frac{\mu_0 I}{2\pi r}2\pi r - \frac{\mu_0 I}{2\pi}\Delta\theta + \frac{\mu_0 I}{2\pi}\Delta\theta \\ &= \mu_0 I \end{aligned} \tag{7.49}$$

and we see that deforming the circle with the notch *did not alter the value of the sum* we got

from multiplying the field times the length of the curved path C (circle with notch) *along* the magnetic field, while ignoring the part of C perpendicular to the magnetic field. Again, this should be reminding you of what we did for Gauss's Law only it is a bit simpler.

Well, we can add more notches; in fact, we can deform the curve C in, we can deform it out, we can deform it so that it goes along the wire and no longer lies in a plane, and as long as we break up C into teensy segments that either lie perpendicular to the field or follow a curved arc tangent to the field, we only get a contribution from the piece of length ds tangent to the field, and that contribution is always of the form:

$$Bds = \frac{\mu_0 I}{2\pi r} r d\theta = \frac{\mu_0 I}{2\pi} d\theta \tag{7.50}$$

If we clean up the geometry of this, picking a path element along C with a *vector* length $d\vec{\boldsymbol{\ell}}$ and selecting only the component parallel to $\vec{\boldsymbol{B}}$ at that point with the dot product, we get:

$$\oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} = \oint_C \frac{\mu_0 I}{2\pi r} \hat{\theta} \cdot r d\theta \hat{\theta} = \frac{\mu_0 I}{2\pi} \int_0^{2\pi} d\theta = \mu_0 I \tag{7.51}$$

which is true for *any* curved path C that goes around the infinitely long straight wire precisely one time, so that the integral of $d\theta$ is eventually 2π (with the r in the length element along $\vec{\boldsymbol{B}}$ always cancelling the $1/r$ dependence of $\vec{\boldsymbol{B}}$).

Note two things. One is that current carrying wires that do *not* pass through the closed loop C do not contribute to the loop integral – the integral of $d\theta$ around such a loop always adds up to zero because it doesn't go, and stay, around. If there were many wires and not just one, we could use superposition and show that this equation would still be true as long as we only add up the total current I that *passes through* C on the right hand side.

The other is something that I can't precisely show, but which you can *kind* of see is true. It turns out that this equation works *even if the wire(s) aren't infinitely long or straight!* In fact it works for *any* steady-state (static) current passing through the closed loop C. We can imagine trying to prove this by (for example) leaving C as a circle and starting to deform the path followed by the current I and noting how the result depends on the angle subtended by each point on the circle, but in the end visualizing it would be too difficult because of the cross product. For that reason, this is one of the few times in this book where I'll ask you to just trust me because I can't *quite* show you how the result doesn't change as we, for example, bend the infinitely long straight wire around into an arbitrary loop itself, while maintaining the fact that it passes "through"[104].

This leads us to write the previous equation in the following carefully selected form:

$$\oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} = \mu_0 I_{\text{through C}} = \mu_0 \int_{S/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA \tag{7.52}$$

which we will call (the incorrect form of) *Ampere's Law*. Ampere's Law is our third Maxwell equation, and is the equation that Maxwell in fact "fixed" to get his name on the entire set. Some fix!

[104]This is your first hint. Exactly what does it mean for a current to pass "through" an *arbitrary* closed loop? It's easy to answer this when the line is straight and the loop is a nice plane circle, but Ampere's Law holds for curves C that are topologically equivalent to a kilometer of fishing line with the ends tied together (to form a closed curve) and the balled up onto the biggest, worst fishing tangle you ever saw! What does it mean for current to go "through" that? And yet it can, and if you stuff the entire snarl into a pipe carrying water, you can completely imagine that some of the water does, in fact, go *through the loop* as it flows along.

Note that I wrote the current "through C" in a mathematically correct way as the *flux of the current density through an open surface S bounded by the closed curve C!*[105]. This is the way I'd like you to practice writing Ampere's Law, although in *application* below finding the field of certain highly symmetric constant (or slowly varying) currents we'll often (but *not always*) be able to just add up the total current through any given loop "by inspection".

Another Maxwell Equation! We're now up to three, with one to go. Gosh, seems as though it should be good for *something*, doesn't it?

Indeed it is. Even in its slightly broken form above, we can use it to *find the magnetic field* in problems with just enough of the right kind of symmetry. There are only a handful of problems that fit the bill, but they are all useful and important. In the end, though, the purpose of Ampere's Law (fixed) is that it is a law of nature (where the Biot-Savart Law per se is not). In fact, we can derive the non-relativistic Biot-Savart Law *from Ampere's Law, using the Maxwell Displacement Current* as I'll show in a chapter or two once we know what the latter is. Just bear in mind that this introductory treatment (but not Maxwell's Equations per se) is always going to be incomplete without special relativity and differential vector calculus covered in a more advanced course on electrodynamics.

For now, though, and in *this* course, we'll content ourselves with *the* handful of problems where Ampere's Law can be used to evalute the magnetic field. We're basically going to do all of them. On your homework, I'll ask you to do them again (without looking, before you are done) and will throw you a few simple enough variants of the problems.

7.6: Applications of Ampere's Law

There are basically four problem geometries where Ampere's Law can be used to find the field. As was the case with Gauss's Law for Electricity, each of them has an associated *symmetry* that permits the path integral on the left to be evaluated "once and for all", so that solving the problem amounts ot finding the total current through the Amperian Loop (topical equivalent of the Gaussian Surface) in question.

Those categories are:

a) Infinitely long straight wire, or cylinder, or cylindrical shell, or anything else where the current has cylindrical symmetry. These examples will be like the argument we used to justify Ampere's Law above, only backwards.

b) Infinitely long solenoid.

c) Toroidal solenoid (which also has cylindrical symmetry, but in a different way).

d) Infinite plane sheet of current (which may or may not be a "thick" sheet).

[105] Hint: There are *many* – in fact, an infinite number of – surfaces S that are bounded by *any particular* closed curve C. Is the value of this integral independent of which surface you choose? If it is, is that a problem?

Example 7.6.1: Cylindrical Current Density – Infinitely Long Thin Wire

The simplest example of this is the infinitely long straight *thin* wire, so we'll do that just as a warm-up.

Take an infinitely long, straight wire carrying a current I. We know *from symmetry* (not just because we used the fact to sort-of-derive Ampere's Law in the first place) that the magnetic field is constant in magnitude on and tangent to a circle of radius r because the problem doesn't change as we walk around the wire. We therefore choose the Amperian Path C to be a *circle of radius* r and do so once and for all for this kind (symmetry) of problem. Then:

$$\begin{aligned}
\oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} &= \mu_0 I_{\text{thru C}} \\
B_t \oint_C d\ell &= \mu_0 I \\
B_t 2\pi r &= \mu_0 I \\
B_t &= \frac{\mu_0 I}{2\pi r}
\end{aligned} \tag{7.53}$$

Big surprise. We find that the field tangent to the circle at all points B_t is exactly what we know it to be as we more or less invert our "derivation" of Ampere's Law. The one *important* lesson to take from this is that the left hand side and concluding algebra for this little mini-derivation will never change! For *every* cylindrical problem we will *always* use a circular Amperian Path and the left hand integral will (because B is constant on and tangent to the circle) *always* evaluate to $B_t 2\pi r$.

The right hand side, on the other hand, we may have to work for. Specifically, we will often have to work to find the actual current through the *particular* C we have drawn.

Still, this seems a hell of a lot easier than setting up and evaluating the Biot-Savart integral we did earlier this week. Maybe there *is* something useful in here, after all!

This is more apparent in the next example.

Example 7.6.2: Cylindrical Current Density – Field of an Infinitely Long Thick Wire

Suppose we have an infinitely long straight wire that has some finite radius R and is carrying a current I that is uniformly distributed across the wire cross-section as shown in figure 7.10. We would like to compute the magnetic field everywhere in space, both inside and outside of the wire.

Our first step is to transform I into a current density $\vec{\boldsymbol{J}}$ into the page:

$$\vec{\boldsymbol{J}} = \frac{I}{\pi R^2}\hat{\boldsymbol{z}} \tag{7.54}$$

where the z-axis is into the page.

Next, we have to think just a bit about the field we expect to get. This step is *essential* – most people who get this problem wrong get it wrong because they omit it, they haven't thought about the problem enough. The current has cylindrical symmetry, so the field will too.

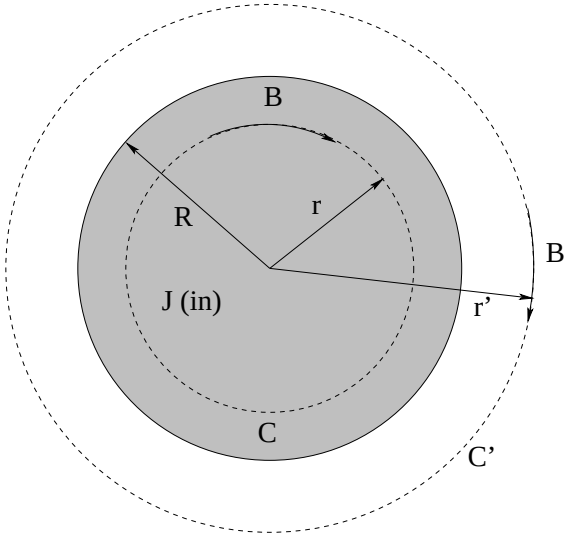


Figure 7.10: An infinitely long thick wire of circular radius R carries a current I into the page as drawn. We would like to find the magnetic field in all space.

We expect the field lines to run in circles of constant magnitude around the center of symmetry in the middle of the wire, in the clockwise direction as drawn from the right hand rule. But we do *not* expect them to have the same form inside the wire and outside of the wire. We therefore have to draw *two* Amperian Paths, one (C) of radius r in region I $r < R$, the other (C') of radius r' in region II $r' > R$. We have to apply Ampere's Law *twice*, once in each region.

Let's do region I ($r < R$):

$$\begin{aligned}
\oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} &= \mu_0 \int_{S/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA \\
B_t \oint_C d\ell &= \mu_0 J \int_{S/C} dA \\
B_t 2\pi r &= \mu_0 J \pi r^2 \\
B_t &= \frac{\mu_0 J \pi r^2}{2\pi r} = \frac{\mu_0 I r}{2\pi R^2}
\end{aligned} \tag{7.55}$$

where we have selected a right-handed normal $\hat{\boldsymbol{n}}$ into the page so that the dot product of $\vec{\boldsymbol{J}}$ and $\hat{\boldsymbol{n}}$ is just the magnitude J. The right hand side, as you can see, computes the total current that flows *through* the curve C (inside the radius r)! The left hand side is identical to what it was for the thin wire (and what it will be for all other cylindrical problems).

Then, region II ($r' > R$):

$$\begin{aligned}
\oint_C' \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} &= \mu_0 \int_{S/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA \\
B_t 2\pi r' &= \mu_0 I \\
B_t &= \frac{\mu_0 I}{2\pi r'}
\end{aligned} \tag{7.56}$$

which is the same as for a long straight thin wire. The field *outside* of any cylindrical current will be the same as the field of a current of the same strength all concentrated in a thin wire at

the origin. This should all be very reminiscent of Gauss's Law and fields outside of cylinders or spheres.

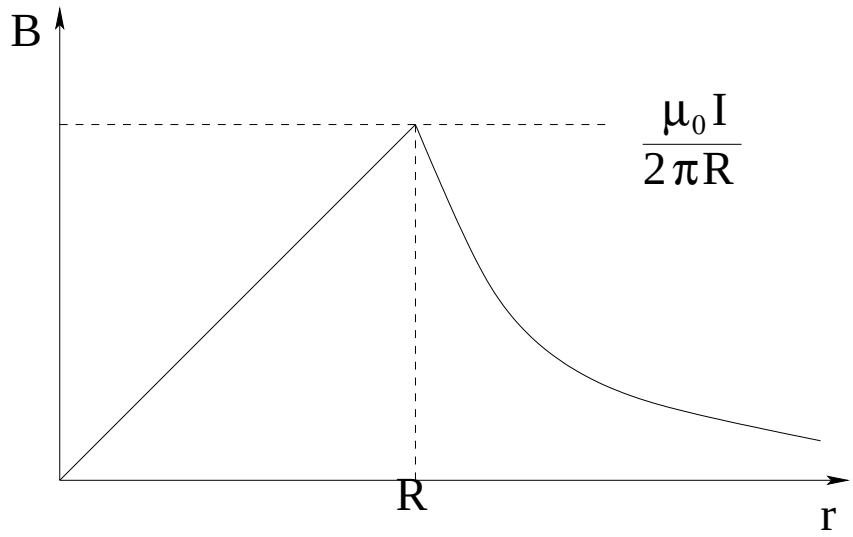


Figure 7.11: $B(r)$ for a long thick wire of radius R carrying a current I. Note that the field increases linearly inside of the wire and reaches a maximum value on the surface of the wire. Outside it drops off like $1/r$. Although the field is continuous, its derivative (slope) is not; it jumps at $r = R$.

We crudely plot the field as a function of r in figure 7.11. Remember, the field circulates around the current (density) in a clockwise direction as determined by the right hand rule.

We could, of course, do more complicated problems now that have this symmetry as long as we can figure out how to do the integrals (or otherwise figure out the amount of current that passes through C) on the *right* hand side of Ampere's Law. The left hand side is always the same. Variations include: Finding the field in a thick cylindrical shell carrying a current I; a coaxial cable; a thick wire with a cylindrical hole, a thick wire with a current density that is *not* uniform. The latter is particularly relevant for alternating currents – when an alternating current is sent through a thick wire the current is *not* uniformly distributed, it tends to concentrate near the surface and die off in the middle. This has implications for computing the resistance and actually affects the design of high voltage power transmission lines and wave guides.

Example 7.6.3: The Solenoid

The solenoid pictured above in figure 7.12 is a classic problem in magnetism – it is (as we will see) the moral equivalent of a capacitor for the storing of *magnetic* energy. A solenoid is also our ideal model for "permanent magnets" as well as electromagnets of all flavors.

In order to apply Ampere's Law to a solenoid – which is basically a cylindrical coil of wire with many (N) turns and cross-sectional area A carrying a current I – we need the solenoid to have enough *symmetry* that we can figure out a suitable Amperian Path. To accomplish this, we will assume that the solenoid is *tightly wrapped* – so much so that the coils form a more or less *continuous* current around the interior volume – and that it is *infinitely long*. Both are idealizations, but both of these assumptions are *good* idealizations – they will work well

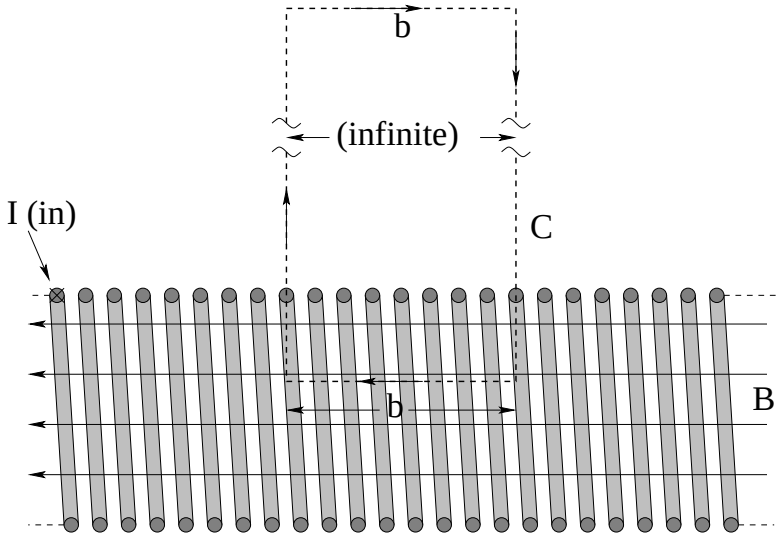


Figure 7.12: A cross-sectional view of an infinitely long solenoid with n turns per unit length, cross-sectional area A, carrying current I in each turn. The field both inside and outside of the solenoid is parallel to the axis of the solenoid (from symmetry), leading to the Amperian Path shown.

enough for any snugly wrapped coil that is (much) longer than its diameter.

If you examine figure 7.12, you can see from symmetry that the magnetic field inside must travel parallel to the axis of the solenoid from right to left. The general right to left direction follows from the right hand rule given the current into the page on the tops of all of the wires and out at the bottom. The fact that it must be *paralle* follows from the fact that every point is in the middle of an infinite line, so there can be no up or down or in or out compontent because it wouldn't be symmetric with respect to either inversion or translation down the solenoid to another "central" point. Furthermore, the field strength must be *constant* along any straight line parallel to the axis for the same reason – it cannot vary from its value in "the middle", whereever you choose to put that middle.

Outside the same is true but opposite. The field (if any) must flow from left to right and be parallel to the axis of the solenoid. This determines a good Amperian Path C. We select a rectangle of side b (inside the solenoid) with *infinitely long sides!* The field is everywhere perpendicular to the sides so we get *no contribution* to the path integral of the field from them. By making the sides infinite, we can also make the field zero on the upper horizontal chunk. We only get a contribution from the side of length b inside the solenoid. That is:

$$\begin{aligned}
\oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} &= B_z b + 0(\text{left}) + 0(\text{top}) + 0(\text{right}) = \mu_0 I_{\text{thru C}} \\
B_z b &= \mu_0 n b I \\
B_z &= \mu_0 n I = \frac{\mu_0 N I}{L}
\end{aligned} \tag{7.57}$$

where we computed the total current *through* C by multiplying the number of turns per unit length by the length of C through which the turns passed times their current.

Note well that this tells us that the field is *zero* outside of an ideal solenoid – all magnetic field lines are confined to live inside the solenoid tube and none can escape to the outside.

It also tells us that the field inside is uniform – there is no dependence of the answer on any spatial coordinates, so it doesn't vary with coordinates beyond being non-zero on the inside and zero on the outside.

The final form is given as you might use it for a solenoid with a finite number of turns N and of finite length L, where (recall) L needs to be much larger than the radius or diameter of the solenoid and where we are finding the field not too near the ends. Usually we will idealize even finite size solenoids as having the field of an infinite solenoid inside, and will neglect end effects. That is, we will assume that the field is uniform but drops to zero "instantly" at the solenoid ends. Of course this isn't physical, but the field *does* drop off very rapidly at the ends, so it is a good approximation once again, as was neglecting fringe fields for capacitors (the moral equivalent in the electrostatic case).

That was certainly *very* easy compared to any sort of Biot-Savart Law integration. The latter *can* be done with some work, but it isn't easy and requires more calculus than you are likely to have so far; maybe some day in a future class you'll do it.

Simple, easy or not, the solenoid is an enormously useful and important example, so be sure you learn it completely.

Example 7.6.4: Toroidal Solenoid

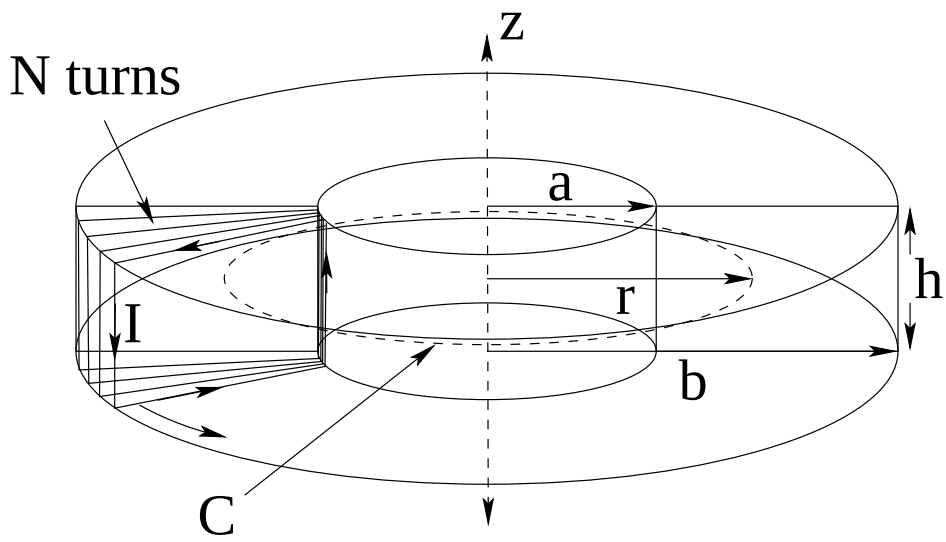


Figure 7.13: A cross-sectional view of a toroidal solenoid with N turns, and a rectangular cross-sectional with inner radius a, outer radius b, and height h, carrying current I in each turn. The field both inside the solenoid is concentric to the vertical axis of the torus (from symmetry and the right hand rule), leading to the Amperian Path shown.

In figure 7.13 above a toroidal[106] solenoid is drawn. The particular one we will look at has a rectangular cross-section although (as we will see) this doesn't really matter as far as finding the field in all of space is concerned – any uniform cross-sectional shape (such as a circle or ellipse or outline of Homer Simpson) would do. We choose a rectangle with nice coordinates

[106]Wikipedia: http://www.wikipedia.org/wiki/Torus. A torus is a "doughnut shape", usually with a circular cross section.

mostly to make it easy to compute the self-inductance of this solenoid *next* week, not because it matters *this* week and this way we can just reuse the figure as well as the Ampere's Law result.

The wires in the figure (drawn on the left) have to be visualized wrapping the whole torus (fairly tightly). If one lays one's right hand thumb mentally along the direction of the current in each leg of a loop around the torus, you can easily convince yourself that *each* wire produces a field nearby that is generally cylindrically "around" the torus in the direction given by laying your thumb in the direction of the *inside* wires, the ones closes to the z-axis of symmetry. In this case the B-field is counterclockwise, then, viewed from our perspective above, and our Amperian Path (along which the field should be constant in magnitude and tangent to the path or anything you like and perpendicular) is a circle of radius r.

We locate the circle *inside* the solenoid at first. Ampere's Law then gives:

$$\begin{aligned} \oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} &= \mu_0 I_{\text{thru C}} = \mu_0 \int_{S/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA \\ B 2\pi r &= \mu_0 N I \\ B_t &= \frac{\mu_0 N I}{2\pi r} \end{aligned} \tag{7.58}$$

where we discover that the current "through C" is just the current in a single wire times the number of wires but *only when the curve C lies inside the torus!* For circles C outside of the torus the current through the any surface bounded by C is zero, as every wire goes (at best) into the surface one time and right back out out one time.

Our conclusion is that the toroidal solenoid *confines* the magnetic field to live inside the torus, and the geometry of the field causes it to drop off like $1/r$! How useful! How interesting! Solenoids in general seem to like to *trap* magnetic field lines and keep them from escaping. If we bend them around in curves, they keep the field inside (and cause it to vary by getting weaker on the outside edges of the curves). If we wrap them back into themselves (making a torus or a topological knot of some sort then the magnetic field cannot get out into the room and remains confined to the inside of the coil.

This property will turn out to be very useful next week when we consider making inductors out of solenoids, as a toroidal solenoid will have the helpful property of having *very little* mutual inductance with nearby current loops, where finite length regular solenoids produce a pesky "fringe field" at their ends that can induce unwanted voltages in conductors or loops close to those ends.

If you look inside a computer or other electronic device, you will usually see a few toroidal inductors soldered into the motherboard, and that is exactly *why* they are shaped the way they are shaped – it is very "bad" for computer motherboards to pick up inductive signals from processes that have nothing to do with their function, especially if the voltages involved approach the threshold that can trigger flips and flops in its enormously complex bit processing structure.

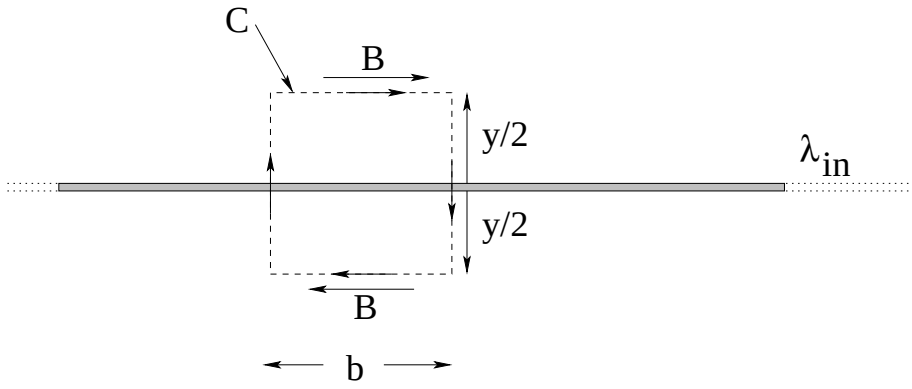


Figure 7.14: A side view of an infinite sheet of conductor carrying a current (per unit length) λ into the page. The field due to the sheet is symmetric up and below the sheet as drawn, and must point parallel to the sheet because every point is in the middle of the infinite plane (as usual). Any up-down asymmetry would violate mirror symmetry about that "middle" because the problem would not change but the solution would. This leads us to the Amperian Path shown, which should remind you of that of the infinte solenoid, with sides perpendicular to the field.

Example 7.6.5: Infinite Sheet of Current

In figure 7.14 we see our final example, an infinite conducting sheet of negligible thickness (exaggerated in the picture) carrying a uniform *current per unit transverse length* into the paper. We then follow a familiar ritual. Every point is in the middle of an infinite sheet, so our picture is located in the middle. If we flip the picture over (maintaining the direction of the current into the paper) the field has to be the same, so we know that the field has to have the same *magnitude* equal distances above and below the plane. We know that the picture has *mirror* symmetry around any vertical line. We know that there is much current to the right of that line (which produces a field with an upward directed component above the sheet) as there is to the left of the line (which produces a field with a symmetric downward directed component), so our right hand tells us that the only possible direction for the field is to the right *parallel* to the sheet above it, and to the left parallel to the sheet below it. A sensible Amperian Path is then a rectangle symmetric about the sheet with sides perpendicular to the field and ends parallel to it, traversed in the right handed direction as shown.

It is now simple to apply Ampere's Law, as we get no contribution from the sides of C and equal positive contributions from the upper and lower legs of C:

$$\begin{aligned} \oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} &= \mu_0 I_{\text{thru C}} \\ 2B_{||} b &= \mu_0 \lambda b \\ B_{||} &= \frac{\mu_0 \lambda}{2} \end{aligned} \tag{7.59}$$

where $B_{||}$ is the magnitude of the component of $\vec{\boldsymbol{B}}$ parallel to the sheet a distance $y/2$ above or below it. Of course we note that this field *doesn't depend on* y so the field above and below the sheet is *uniform* to the right and left respectively.

There is a bit of insight to be gained from thinking about *two* sheets, one carrying current

in, one carrying current out, separated by a distance d. In this case the superposition principle suggests that the field above the two sheets and below the two sheets will be zero, as the contributions from the two sheets cancel. In between, though, they *add* to a total magnitude of:

$$B_{||} = \mu_0 \lambda \tag{7.60}$$

If we imagine that λ is made up of the field in a lot of very closely spaced single wires each carrying some current I, then you can see that:

$$\lambda = nI \tag{7.61}$$

or, the number of *wires* per unit length times the current per wire equals the amount of *current* per unit length. The field in between is thus:

$$B_{||} = \mu_0 \lambda = \mu_0 nI \tag{7.62}$$

which looks *just like the field of a solenoid!*

Recall that our computation of the field inside an infinitely long solenoid didn't depend on the cross-sectional shape of the solenoid. In fact, it could have been rectangular! If we imagine that the top and bottom sides of the rectangle get longer and longer, eventually we can imagine that they become *infinitely* long and close only *at* $\pm\infty$ so that the current that goes in at the top returns on the bottom (say). In this way we can see that our result for the pair of infinite sheets *makes sense* and is completely consistent. We could have guessed this result by mentally deforming a solenoid until it looked in our minds like two infinite sheets in close to where we were actually measuring the field.

It also tells us that even though we have been quite careful to make the sheets we have been considering be planar, all we really need is for them to be *straight* in the left-right direction, and continue on to infinity (and "close") there in the direction in and out of the page. Two e.g. hyperbolic sheets of current that stretch to infinity would have exactly the same field in between them as we obtained in this example. This sort of conceptual understanding can be very useful later on, as can the ability to think in terms of *topological deformations* of the sort we have just considered, so don't be surprised if a quiz question probes whether or not you "get it" well enough to answer simple conceptual questions.

7.7: Concluding Discussion

Yes, this week is long enough, and has enough content, that it is worth a bit of a wrap-up at the end. We have covered one and a half Maxwell equations, after all!

At this point you should be aware that unless and until somebody positively discovers magnetic monopoles in an experimentally reproducible setting so that everybody agrees that they are real (and ideally, learns enough about them to incorporate them into our general picture of physics) Gauss's Law for magnetism will tell us that magnetic field lines produce no net flux through a closed surface S and consequently must form closed loops in space.

The Biot-Savart Law for currents tells us how to compute the magnetostatic field produced by a steady-state current distribution, if we can manage the complexity of dealing with vectors, cross-products, and multivariate integral calculus simultaneously.

The "Heaviside" form for the magnetic field of a point charged particle q travelling at some velocity $\vec{v} \ll c$, consistent with the Biot-Savart Law, led us to some serious puzzles, enough to make us doubt the consistency of classical physics itself. For one thing, we were able to show that the interaction forces between two charged particles interacting with this field violated Newton's Third Law and hence (apparently) the Law of Conservation of Momentum for the pair! For another, Biot and Savart only obtained their experimental law by studying steady state currents, and a charged particle exists only at a single point in space and isn't smeared out into a "continuous" current; we effectively assumed that the magnetic field propagates *instantaneously* from the moving charge in the form we wrote down, and as it will turn out, this is incorrect.

This was apparent when we thought about the appearance and disappearance of magnetic fields (and hence magnetic *forces* when we do nothing but change the inertial reference frame in which we view charge-current distributions – if the electric field and force (and possibly more) don't change at the same time, changing reference frames could produce distinct physical realities with *different forces, accelerations, and hence trajectories!*

Finally, we obtained with some hand-waving from the Biot-Savart Law a new equation we called *Ampere's Law* after its discoverer. Unfortunately it inherits a *flaw* from our sort-of-derivation – it is essentially a static result, good only for steady-state currents (like the Biot-Savart Law itself). We *did* find Ampere's Law to be remarkably useful for finding the *static* (i.e. at most "slowly" varying in time) magnetic field produced by suitably symmetric *static* current distributions, but we are, or should be, a bit worried about consistency because (hint hint) the "current through the closed curve C" that it explicitly references seems as though it can mean nothing but the flux of the current density through some open surface S *bounded* by that closed curve, but there are an infinite number of these surfaces and we (should) have the uncomfortable feeling that the current we obtain *depends on the surface chosen* where it really shouldn't.

An *invariant* form of the current – one that one could prove does *not* depend on the surface chosen – would be much better, especially if it still gives us the usual static result where it should, but what physical principles or insight might lead us to such an invariant form?

Ah, puzzles in abundance! Things are finally getting interesting! This is a *good* thing, as reality is undeniably rather complex and if the electric and magnetic force were *too* simple they could not sustain the complexity we see every time we, well, *see*. This seems like a good time to wrap up electro***statics*** and magneto***statics*** and move on to electric and magnetic field ***dynamics***.

We'll begin by trying to understand a puzzle that we haven't really faced until now. Magnetic forces are by definition always exerted at right angles to the direction of motion of a charged particle or moving current. This means that *magnetic forces do no work!* on isolated classical charged particles (with no intrinsic magnetic moment), because work requires a force component *in* the direction of motion. Next week we will study what at first glance then seems like a paradox – cases where magnetic fields clearly appear to do work – and then *resolve* the paradox by concluding instead that magnetic fields under some circumstances create *electric* fields, and electric fields have no difficult at all doing work on charged particles!

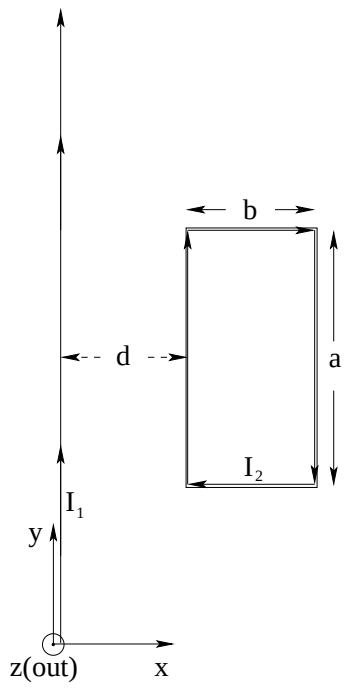
Homework for Week 7

Problem 1.

Physics Concepts

Make this week's physics concepts summary as you work all of the problems in this week's assignment. Be sure to cross-reference each concept in the summary to the problem(s) they were key to. Do the work carefully enough that you can (after it has been handed in and graded) punch it and add it to a three ring binder for review and study come finals!

Problem 2.



An infinitely long straight wire carries a current I_1 in the $+y$ direction. At $x = d$ there is a rectangular loop of current I_2 in the xy plane, with two sides of length a parallel to the long wire and two sides of length b perpendicular to the long wire. The current in the wire segment nearest the long wire is *parallel* to the current I_1 in the $+y$ direction as drawn.

Find the net force acting on the rectangular loop (in the provided coordinate frame).

Problem 3.

Using Ampere's Law, find the magnetic field in all space produced by:

a) A solid conducting cylinder carrying a total current I.

b) Two cylindrical conductings shells carrying opposite currents (each equal to I in magnitude). The inner one has radius a, the outer one b.

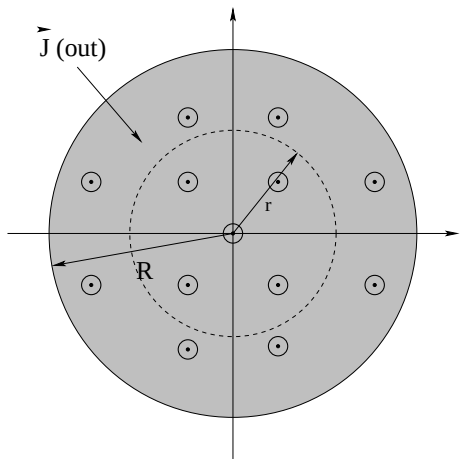
c) A solenoid with N turns and length L carrying current I in each turn (inside only, far from the ends).

d) A toroidal solenoid with N turns, inner radius a, outer radius b.

e) An infinite plane sheet of current into the paper (above and below the sheet).

This more or less *exhausts* the list of possible problem types where one can find the magnetic field using Ampere's Law. Most are examples presented in lecture and/or the textbook, so this forces you to recapitulate on your own what is presented there and ***make the solutions your own by practicing them*** instead of just trying to remember how *I* (or your current lecturer/instructor/professor) did them *for* you.

Problem 4.



A cylindrical long straight wire of radius R carries a current density of magnitude $J = J_0 \frac{R}{r}$ ***out*** of the page as drawn.

a) Find the magnetic field (magnitude and direction) for $r < R$ (inside the wire).

b) Find the magnetic field (magnitude and direction) for $r > R$ (outside the wire).

c) Sketch (graph) the ***magnitude*** of the magnetic field $B(r)$ from $r = 0$ to $r = 2R$. Where is the field maximum and what is its value there in terms of I?

Problem 5.

Based on the ***analogy*** between electric and magnetic dipoles, deduce the probable form of the magnetic field of a spherical ball of charge Q, mass M, and radius R centered on the origin that is rotating at angular velocity $\vec{\boldsymbol{\Omega}} = \Omega\hat{\boldsymbol{z}}$:

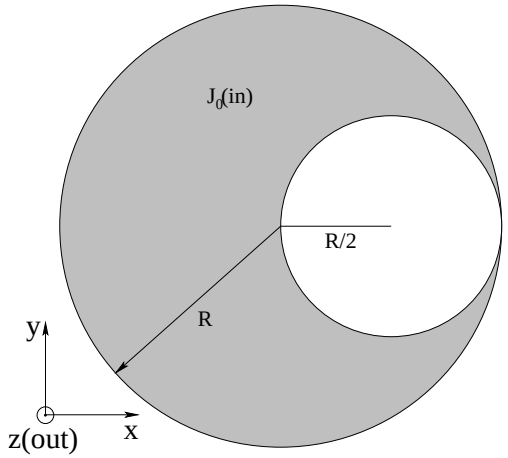
a) at a point on its axis of rotation;

b) at a point in the plane that passes through the ball perpendicular to the axis of rotation;

in both cases *far* from the ball of charge, that is, for $z \gg R$ and $x \gg R$ for the ball spinning around the z axis.

Note: that it is quite a bit of work to actually *derive* this result (though it can be and is done in more advanced electrodynamics courses). This is part of the *point* of multipolar expansions – once one knows the form of the field for any given multipolar moment, one merely has to compute that moment for a give charge-current density to discover the (far) field "for free"!

Problem 6.



A cylindrical conductor of radius R is aligned with the z direction (perpendicular to the page). It has a cylindrical *hole* of radius R centered at $x = R/2$ also aligned with the z direction as shown. The conductor carries a *uniform current density* $\vec{\boldsymbol{J}} = -J_0\hat{\boldsymbol{z}}$ into the page and obviously $\vec{\boldsymbol{J}} = 0$ in the hole where no conductor is present.

Using Ampere's Law, the superposition principle, and some cleverness, ***find the magnetic field at all points inside the hole***.

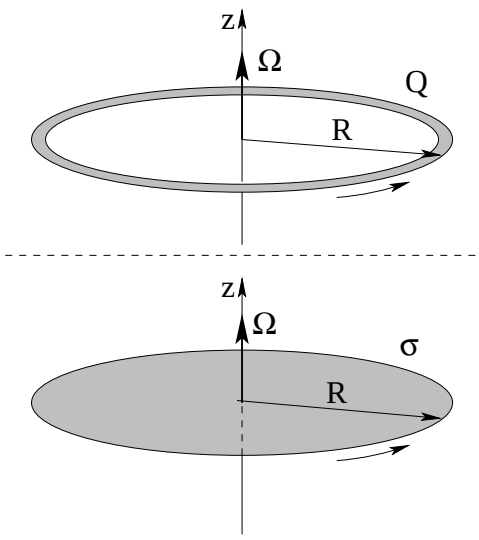
Problem 7.

Using the Biot-Savart law:

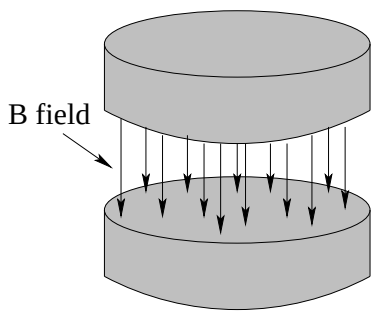
a) Find the $\vec{B}$-field on the z axis of a rotating ring of radius R of (uniformly distributed) charge Q located in the x-y plane and centered on the origin that is rotating with angular frequency Ω around the z axis. (See upper figure to the right.)

b) *Set up* the integral to be done to find the $\vec{B}$-field on the z axis of a flat *disk* of radius R of uniformly distributed charge density σ located in the x-y plane and centered on the origin that is rotating with angular frequency Ω around the z axis. (See lower figure to the right.)

c) **For Advanced Students:** *Do this integral!* This requires u-substitution and either *integration by parts* a couple of times or a clever second substitution but is otherwise straightforward.

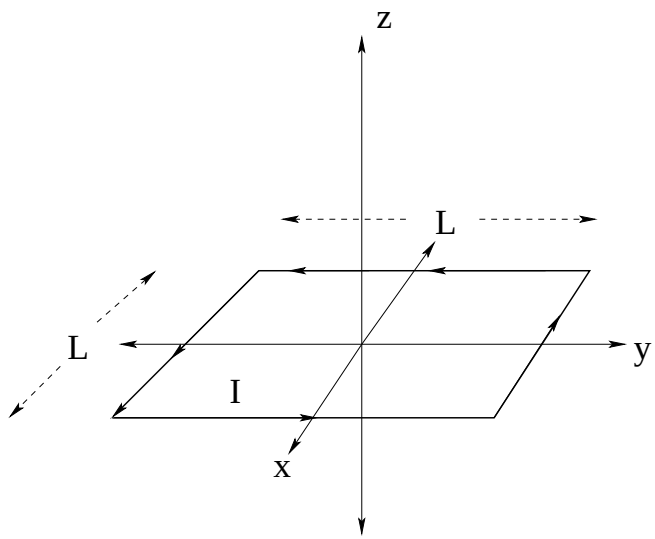


Problem 8.



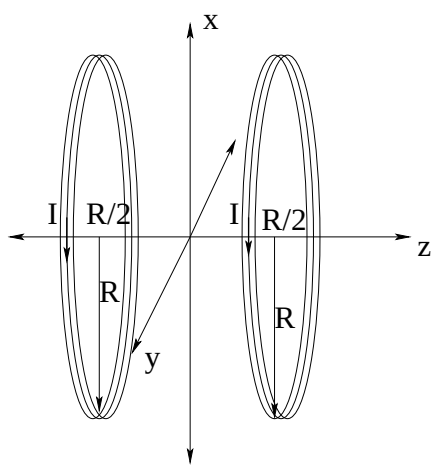
Using a rectangular closed curve C that lies partly inside and partly outside the region where the field between two magnetic "poles" or in the gap between two solenoids is not zero, show that a *uniform* magnetic field there that has *no fringing field* violates Ampere's law. Then explain why this does *not* apply to the uniform field inside a solenoid, which goes "sharply" to zero as one crosses the *current* in the solenoid loops inside to outside.

Problem 9.



A square loop of wire lies in the $x - y$ plane centered on the z axis and carries a current I. It has side length L. Find the magnetic field at an arbitrary point on the z axis, and show that in the limit $z \gg L$ it gives an expected result in terms of the magnetic moment m_z of the loop. Note that this problem is "simple" – just a repeated use of the field of a straight segment of wire – but visualizing the *geometry* in terms of the *givens* is not simple and is the object of the exercise. So draw a very good, very large picture! Or several! Visualize!

Advanced Problem 10.



A "Helmholtz coil" is made up of two loops of wire with N turns and radius R carrying a current I per turn. They both are concentric with the z axis with centers at $z = \pm R/2$. Show with suitable expansions that at $z = 0$:

a) $\dfrac{dB_z}{dz} = 0$; and

b) $\dfrac{d^2B_z}{dz^2} = 0$.

This result means that the magnetic field is quite "flat" (nearly constant and uniform) in the middle region of a Helmholtz coil, making it suitable for experiments that would otherwise have to be conducted inside e.g. an opaque solenoid.

IV: Electrodynamics

Week 8: Faraday's Law and Induction

- Suppose a conducting bar moves through a field at right angles to the field lines and the alignment of the bar. Magnetic forces quickly push charges to the two ends until an electric field is created that *balances* the electric force. The integral of this field is called a *motional* potential difference.

- Suppose now that a rectangular wire loop is pushed *into* (or pulled out of) a uniform field that terminates at an edge (perhaps generated by a solenoid with a slot in it). We note that the field now pushes charges around the loop in agreement with the motional potential difference and that the net magnetic force on the current carrying wire *resists* the push into (or pull out of) the field.

- We consider a conducting rod on rails as it slides through such a field. We can see that the induced/motional potential difference is equal to the time rate of change of the field times the area the field occupies within the rectangle.

- Time for our final Maxwell equation. If the magnetic field flux through an open surface S bounded by a closed curve C *varies in time* it *induces* an *electric field* dynamically around the closed curve according to *Faraday's Law*:

$$\oint_C \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{\ell}} = -\frac{d}{dt} \int_{S/C} \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA \tag{8.1}$$

  The integral on the left is the *induced voltage* around the curve C.

- In this equation the minus sign is called *Lenz's Law* and tells us that the induced voltage decreases around the loop in the direction such that a flow of positive charge in that direction (the *induced current* if the loop is a conducting pathway) will *oppose the change* in the varying flux. If the flux is decreasing it will generate a magnetic moment that points in the direction that will increase it. If it is increasing it will generate a magnetic moment that points in the direction that will decrease it. This causes the *opposition* to motion noted in the motional voltage problems above.

- The flux through a conducting loop is directly proportional to the current through the loop itself or to the current through nearby sources of magnetic field that produce the flux. The constant of proportionality in either case depends solely on the *geometry* of the loop and source(s). That is, given a bunch of loops:

$$\phi_i = \sum_{j \neq i} M_{ij} I_j + L_i I_i \tag{8.2}$$

where the M_{ij} are called the *mutual inductances* between the ith and jth loops and L_i is the *self inductance* of the ith loop.

- From this we can compute the *self*-induced (loop) voltages for simple current-carrying loops, in particular solenoids. To compute the self-inductance of a solenoid we begin with the result for the magnetic field inside an ideal solenoid from Ampere's Law:

$$B = \frac{\mu_0 N I}{L} \tag{8.3}$$

(parallel to the solenoid axis). The current I creates a flux *per turn* that is equal to:

$$\phi_t = BA = \frac{\mu_0 N A I}{L} \tag{8.4}$$

where A is the cross-sectional area of the solenoid. The total flux is thus:

$$\phi = NBA = \frac{\mu_0 N^2 A I}{L} = L_s I \tag{8.5}$$

where L_s is the self-inductance of the solenoid. Clearly:

$$L_s = \frac{\mu_0 N^2 A}{L} \tag{8.6}$$

which depends *only on the geometry of the solenoid* just as the capacitance of an arrangement of conductors depended only on *their* geometry.

- The self-inductance of solenoids can be altered by wrapping them around suitable *magnetic materials* that enhance (para) or reduce (dia) the magnetic fields inside. Solenoids so constructed are ubiquitous in circuit design, where they are known as *inductors*; they are labelled with their inductance L in *Henries*, the SI unit of inductance:

$$1 \text{ Henry } = \frac{1 \text{ Volt} - \text{Second}}{\text{Ampere}} = 1 \text{ Ohm} - \text{Second} \tag{8.7}$$

- In terms of inductance:

$$V_L = -L\frac{dI}{dt} \tag{8.8}$$

is a statement of the voltage across an inductor using Faraday's Law.

- Mutual inductance is the basis of a number of devices, in particular a center-tap full-wave rectifier commonly used in e.g. DC power supplies or AM radios and in *transformers*, an essential component of the power distribution grid. If one imagines *two* solenoids, one with N_1 turns and cross sectional area A and a second one with N_2 turns wrapped *around* the first (so all of the flux (per turn) in the first passes through the loops of the second:

$$\phi_t = \frac{\mu_0 N_1 A I_1}{L} \tag{8.9}$$

for the first solenoid, so:

$$\phi_2 = N_2 \frac{\mu_0 N_1 A I_1}{L} \tag{8.10}$$

is the total flux through the second solenoid due to the current in the first. Thus:

$$M_{21} = \frac{\phi_2}{I_1} = \frac{\mu_0 N_1 N_2 A}{L} = M_{12} = M \tag{8.11}$$

8.1: Magnetic Forces and Moving Conductors

Last week we saw that our study of the sources of the magnetic field, together with the Lorentz force law, are starting to raise red flags concerning the consistency of electromagnetic theory. One victim of the developments so far is *Newton's Third Law* [107], directly violated by magnetic forces between moving charged particles! We noted that there is some sort of hidden problem with Ampere's Law – that you may or may not have figured out on your own from the hints – but it seems as though it might have something to do with dynamics and using a current that is *invariant* when we make an arbitrary choice of "the surface S bounded by a closed curve C".

In addition, the magnetic fields that appear when we *change reference frames* in which we view even isolated charges did not appear to lead to Newton's *Second* Law being invariant under Galilean frame changes as well! Up to now, we have absolutely relied on Newton's Second Law and its frame-invariance as the basis for *all of our dynamics*, so this is *very disconcerting!* Worse, when we studied the Lorentz force law, we learned that magnetic forces, by their very nature and defining equation, *can do no work* on isolated charged classical (spinless) particles or, by extension, electrical currents!

This week we will *begin* by looking at a very simple scenario in which it will certainly *look* algebraically like magnetic fields do work and confront the puzzle: *How is this possible?*

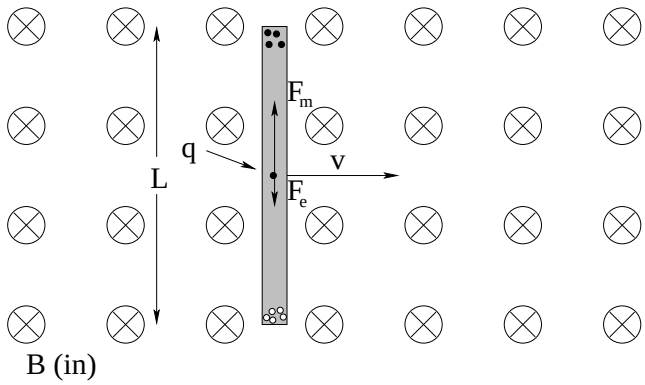


Figure 8.1: A conducting rod of length L moving through a uniform magnetic field into the page. The field *polarizes* the free charge in the rod until a region of crossed fields is produced.

To see the nature of the difficulty, consider a conducting rod of length L moving through a uniform magnetic field at right angles to the field as show in figure 8.1. The rod is, of course, made up of many, many microscopic point charges, and as the rod *moves to the right at velocity* $\vec{v}$ in the magnetic field, all of those charges experience a magnetic force (according to the Lorentz Force Law that we learned two weeks ago).

Because it is a conductor, it has an "inexhaustible" supply of free charge that can move within the conductor under the influence of this force while the equal and opposite charge of the presumed neutral conductor is pushed the other way. We will assume that the free charges have magnitude q (which might be positive or negative) – none of what we work out will depend

[107] One can "sort of" rescue it by insisting that it only holds for force pairs *directed along the line connecting two particles* and not forces with nasty right-angled cross-products in abundance inside, but that both begs the question – sure, it holds except where it doesn't – and prevents us from sensibly setting out to rescue it and thereby the *Law of Conservation of Momentum* by inventing a *field theory* where *fields* can carry energy and momentum.

on its sign or whether only one or both charge flavors are free to move.

The magnetic force on any given charge in the rod is, of course:

$$\vec{\boldsymbol{F}}_m = q(\vec{\boldsymbol{v}} \times \vec{\boldsymbol{B}}) \tag{8.12}$$

which is *up* for any e.g. positive conduction charge that ***appears*** to be moving to the right at velocity $\vec{\boldsymbol{v}}$. We therefore expect the magnetic field to push free charge up until it reaches the end of the rod, where a surface potential holds it in (the vacuum beyond is basically an insulator, if you like). Every charge that migrates to the upper end leaves behind a "hole" (ion of the opposite charge in the lattice) and following the exact same reasoning we used in our study of the Hall Effect, we conclude that these negatively charged "holes" will migrate (via backfilling) until they are located at the lower end, at which point there is no charge available to backfill them.

The charge in the rod therefore *polarizes*, creating a net negative charge at one end and a net positive charge at the other end that create an *electric field in between* pointing from the top end to the bottom one. Charge will move until the remaining free charge in the rod in between the ends experiences no net force when the electric and magnetic forces *balance* – figure 8.1 shows an intermediate state where they don't yet balance and charge is still moving around. Ultimately, the rod spontaneously forms a *region of crossed fields*, exactly the same way it spontaneously formed in the case of the Hall Effect, only now there is no current; the forces that balance are brought about solely by the motion of the rod through the stationary, uniform magnetic field!

With this insight, we can easily deduce the condition for force balance for the charges in the rod proper:

$$\vec{\boldsymbol{F}}_m + \vec{\boldsymbol{F}}_e = 0 \tag{8.13}$$

or (since they are in opposite directions and the motion is at right angles to the magnetic field)

$$qvB = qE \tag{8.14}$$

or the magnitude of the electric field that is generated in the polarized rod is given by $E = vB$. This field, in turn, creates an *electric potential difference* between the ends of the rod:

$$\Delta V = L \cdot E = (vL) \cdot B \tag{8.15}$$

If we were to somehow construct a conducting pathway between the ends of the rod, we would expect current to flow, and naively at least we would expect it to be driven by the magnetic force on the positive conduction charges that push them *up* in the rod and hence around in a loop of we build that conducting pathway, even though we know that the *magnetic force cannot be doing any work!* This, of course, is the paradox – if the magnetic field isn't doing any work, but work is being done, *what is?*

To make the problem we are confronting absolutely clear, please note that in the figure above, we are examining what happens in a frame of reference in which the rod moves through a static, uniform magnetic field. Let's imagine that we have changed reference frames – mentally jumped *onto* the rod so that we and the rod are now at rest and the uniform magnetic field is sweeping past us the *opposite way*. This is portrayed in figure

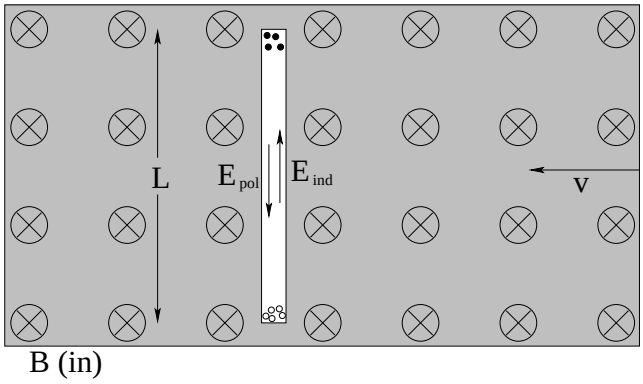


Figure 8.2: A *stationary* (white) conducting rod of length L sits inside a magnet producing a uniform magnetic field (shaded grey) moving in the *opposite direction* at speed v. We *must* observe the *same* charge polarization and electrostatic field in the rod observed in figure 8.1 as this is a trivial inertial frame change! But in this frame *there is no magnetic force* acting on the stationary charges in the rod! We are left with little choice but to imagine that an equal-but-opposite *electric* field has appeared inside the moving magnetic field.

In this case we have no reason to think that there should be a magnetic force on charges in the rod at all! *They are all at rest* in this frame of reference, and the magnetic field they are moving in isn't varying at the location of the rod, it is *constant in magnitude and direction!* Yet things like the observed distribution of charge (at least!) in the frame where the rod is stationary has to *agree* with the distribution of charge in the frame in which the rod moves. Physical reality *itself* cannot change along with our point of view; the charges are where they are (at the ends of the rods) no matter the frame we look at the rod in!

Even in this stationary frame, then, the charge in the rod has apparently polarized in such a way that that the electric field inside the rod is *zero* (as charges there are no longer moving and there is certainly no magnetic force acting on them). We know that the electric field generated by the polarization charges has not changed – it is still $E_{\rm pol} = vB$ pointing *down* in the stationary rod.

We can explain this – I'm tempted to say *only* explain this but that is likely an overreach of our logic – by asserting that an ***induced*** electric field $\vec{\boldsymbol{E}}_{\rm ind}$ *of exactly the same magnitude, pointing up* has appeared *in space* outside of the rod but inside of the moving magnet, in such a way that the vector sum of $\vec{\boldsymbol{E}}_{\rm ind}$ (up) and the *electrostatic* field $\vec{\boldsymbol{E}}_{\rm pol} = vB$ (down) produced by the polarization of the rod *cancel* inside the rod – as they must!

If there is no possible way for a magnetic force to be exerted on the stationary charges in the rest frame of the rod, the only remaining force that (as far as we know) acts on charge per se *at all* is an electric force. A consistent explanation, however odd it might seem at first, is that the motion of the rod through the magnetic field, when viewed in the frame of the stationary rod, has generated an *external electric field from the bottom of the rod towards the top!* This field has acted *exactly* like an external field always does, and created surface charge densities at the ends that polarize the rod until the internal field *cancels* the external field inside of the conductor!

Because our results for the reaction/polarization field have to agree in both frames (where electrostatic fields shouldn't depend on the slowly moving frame) the "induced" external field

$\vec{\boldsymbol{E}}_{\rm ind}$ must be equal in magnitude and opposite in direction to the polarization field:

$$E_{\rm ind} = vB = E_{\rm pol} \tag{8.16}$$

but pointing *up*, not down, when seen in the rest frame of the rod. This, believe it or not, is our first glimpse of a natural law that is one of the fundamental cornerstones of human civilization in disguise – without it our lives would be far, far poorer.

By once again using our imagination to change our point of view to a different inertial reference frame and using the expected invariance of the laws of physics when we perform such a change in frame, we have discovered *induction* – the creation of electric fields from magnetic fields subjected to a change in the frame of reference. We have a ways to go before we completely understand this and can write the result down as our fourth and final Maxwell equation, *Faraday's Law*, but we can already see that it must be so as the result beautifully resolves the paradox of "what does the work" on moving charges in a magnetic field (which can do no work, yet work as we shall see in a moment is clearly done).

In the next section we will reconsider this rod when we do indeed provide it with an idealized conducting pathway that allows current to flow. In the process, we will get a step close to a suitable general formulation of the underlying physical principle.

8.2: The Rod on Rails

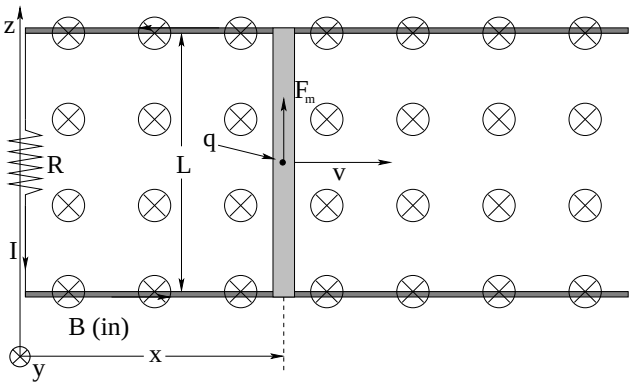


Figure 8.3: A conducting rod of mass M and length L moving through a uniform magnetic field into the page and sliding on *frictionless conducting rails* that are connected by a resistor R outside of the magnetic field. Current can flow around the loop thus formed.

In figure 8.3 we have added a pair of frictionless conducting rails connected by a wire outside of the magnetic field. The total resistance of the loop thus formed (including the rod) is R. We have added an x coordinate to show indicate the instantaneous position of the rod, which is still moving to the right at speed v.

In the previous section we decided that while in the lab it looked as though there was a magnetic force acting up on any given free charge q in the rod (which is now free to move all the way around the loop as part of a "continuous" current I formed in the usual coarse-grained limits we have now seen several times), in the frame of the rod itself there was an *external electric field* generated as the magnetic field moved across it in the opposite direction

of magnitude $E = vB$. In this frame, at least, this is what actually pushes the charges along, doing work as needed. Of course this electric field *now* has to exist in the entire conducting pathway as it has to push the charges along against the actual resistance R, and we know that to properly ensure that the work-energy theorem is satisfied, we should think not of the field, but of the *potential difference* produced by the field. The magnitude of the potential difference induced across the rod in its own frame is clearly:

$$|\Delta V_{\rm ind}| = E_{\rm ind} L = (vB)L \tag{8.17}$$

The next thing to consider is the *sign* of this potential difference and the direction of the induced current in the loop. This is *very confusing*; I will do my best to explain it, but you will have to work through it carefully to understand it.

We can see that both the magnetic force and the induced electric force in the rod, at least, are going to push an electrical current counterclockwise in the arrangement above and we will see shortly that any other direction would *badly* violate energy conservation. We drew the resistance R outside of the magnetic field for convenience in a frame where its leg and the rails and the magnetic field itself are stationary and the rod is moving, but we could have put the resistance *anywhere in the loop* – in the stationary rails inside of the magnetic field, or in the moving rod itself inside of the magnetic field – and of course we can also still "hop" mentally into the frame where the rod is stationary and it is the rails, the resistor, and the magnetic field that are all moving!

Consider this last case and imagine the resistance R to be inside the rod, in the frame where the rod is stationary. Counterclockwise current is then flowing from the bottom of the resistive rod to the top, so (from Ohm's Law) the bottom of the rod *must be at a higher potential than the top*. This makes sense from the point of view of the induced electric field in the frame of the rod, which is great!

Next, let's consider the resistance to be in the top rail, in any frame. The current is still counterclockwise, so the right hand side of the resistance is at higher potential than the left! When the resistance is outside of the field in the vertical left-hand leg, the potential at the top is greater than that at the bottom! When it is in the bottom leg, the potential is higher to its left than its right!

We see that no matter where we put the resistor in the loop, the potential has to *decrease* when we traverse the circuit loop *counterclockwise* across the resistor! That means that the "battery" produced by the induced voltage has to *increase* as we traverse the circuit *the entire circuit loop, going the other way* (clockwise) no matter where we imagine that "battery" to be.

Clearly the induced voltage doesn't appear only "in the rod", either in the frame where it is moving or the frame where it is stationary or any frame at all – it appears *in the entire loop all at once* because the fields that produce it appear and disappear in different places depending on the frame we choose, but *observers in all frames have to agree about the current in the wire, magnitude and direction* just as in our initial rod-only example, they had to agree on the rod polarization.

This is basically another appeal to physical invariance plus what we've learned about resistances in series – in both frames we know that the magnetic force or induced electric force respectively must somehow push the charges around the loop *counterclockwise* when v is to

the right. All we have to do to ensure this is make up a sign convention for the induced voltage so that this is true, and ensure that energy conservation is satisfied (that is, ensure that Kirchoff's Loop Rule is satisfied in both frames).

Let's take the direction of the magnetic field itself through the loop as the direction of a unit vector normal to the loop, $\hat{\boldsymbol{n}}$. We will then use the *Right Hand Rule* as usual to determine the positive direction around the loop by letting our thumb point in this direction and noting which way our fingers curl around the loop (clockwise in this example). If we traverse the loop clockwise, the potential across the *resistor* will *increase* by IR, no matter where in the loop we place it. This means that the induced electric field has to circulate around the loop *counterclockwise* (in agreement with what we concluded considering the frame where the rod is stationary).

We can therefore write Kirchoff's Loop Rule for the loop, going clockwise, and putting in an *explicit integral for the potential produced by the induced electric field*:

$$\Delta V_{\text{loop C}} = -\oint_C \vec{\boldsymbol{E}}_{\text{ind}} \cdot d\vec{\boldsymbol{\ell}} + IR = 0 \tag{8.18}$$

$\vec{\boldsymbol{E}}_{\text{ind}}$ goes around *counterclockwise*, so that the voltage integral around the loop is *positive*, and we already evaluated it in the frame where the rod is stationary as having magnitude BLv, so:

$$\Delta V_{\text{loop C}} = BLv + IR = 0 \quad \Rightarrow \quad I = -\frac{BLv}{R} \tag{8.19}$$

The minus sign in the current is relative to the positive direction around the loop – it tells us that the current is counterclockwise! This is the only possible sign that can correctly cause energy to be conserved as a charge is pushed around the loop without gaining or losing net energy in a circuit; the charge has to *gain* energy from the induced field and *lose* energy into Joule heating of the resistance, and the physically induced electric field has to be parallel to the current to be able to drive charge through the resistor (which could be anywhere in the circuit)!

This last bit is the final thing we have to clear up. Where, exactly, *is* this $\vec{\boldsymbol{E}}_{\text{ind}}$ electric field induced? What is it (in detail) inside of the conductor? At the moment, it appears to follow the resistor around as we change the location of the resistor and/or change frames. And in fact, this is *exactly* where it is, at least in this simple example.

The electric field inside the conducting loop depends on the resistivity and current density associated with the entire conductive pathway, since we know that Ohm's Law can be written as:

$$\vec{\boldsymbol{E}} = \vec{\boldsymbol{J}}\rho \tag{8.20}$$

at all points inside the current carrying conducting pathway. Where ρ is zero, there is no field at all. Where ρ is not zero, there must be a field pushing the charges through the resistive conductor there. The cumulative work done by that field equals the rate that work appears as heat in the resistor, and the only thing that can be creating that electric field is *something* associated with the changing magnetic field that does *not* depend in detail on the particular frame we watch the experiment in!

The best that can be said, then, is that the field appears in the *entire loop*, not "across the rod" or "across the resistor" (which isn't even moving) or "along the rails" (which might actually

be a part of the net resistance, as might be the rod). This *also* means that the induced electric field *forms a closed loop*. ***It is neither an electrostatic field nor what we have called so far a conservative field!***

This does not violate Gauss's Law for Electrostatics – we can add any electric field loops we like to the electrostatic field loops it describes and they will not contribute to the net electric flux through any closed surface S – but it does make one of our rules for visualizing electric field lines obsolete. Electro*static* fields begin and end on electric charges, but induced electro*dynamic* fields apparently can form closed loops, not beginning or ending on any charge!

This does have a significant impact on how we write the *electric potential* associated with the electric field. Recall that we defined a *conservative force* as one where:

$$\oint_C \vec{\boldsymbol{F}} \cdot d\vec{\boldsymbol{\ell}} = 0 \tag{8.21}$$

for all closed loops C one can draw in space. The electrostatic field was conservative – if we let $\vec{\boldsymbol{F}} = q\vec{\boldsymbol{E}}$ and factored and cancelled q, we got:

$$\oint_C \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{\ell}} = 0 \tag{8.22}$$

The *induced* electro*dynamic* field that appears in the loop, however, is *not conservative!* It has a nonzero integral around the loop:

$$\Delta V_{\rm ind} = \oint_C \vec{\boldsymbol{E}}_{\rm ind} \cdot d\vec{\boldsymbol{\ell}} = BLv \neq 0 \tag{8.23}$$

We recall that the whole point of a conservative field and its associated potential was that $\vec{\boldsymbol{E}} = -\vec{\boldsymbol{\nabla}} V$ (encapsulating Newton's Second Law) in cases where the work done going around a closed loop didn't depend on the path taken. This new result more or less means that the work done *does* depend on the path taken, but in a very special way. It also does indeed mean that $\vec{\boldsymbol{E}}$ is *no longer going to be equal to the negative gradient of the electrostatic potential!* We are going to get an *additional* piece that depends in some way on the magnetic field and the loop itself!

My goodness, things are getting complicated! Perhaps it is time to make just two more observations and then finish off this particular problem before coming back to the equation that it seems to imply. The first observation is that (given constant B and L in the picture above):

$$|\Delta V_{\rm ind}| = \oint_C \vec{\boldsymbol{E}}_{\rm ind} \cdot d\vec{\boldsymbol{\ell}} = BLv = \frac{d(BLx)}{dt} \tag{8.24}$$

(because $v = \frac{dx}{dt}$) and, noting that $A = LX$ is the area inside of the loop we can write this as

$$|\Delta V_{\rm ind}| = \oint_C \vec{\boldsymbol{E}}_{\rm ind} \cdot d\vec{\boldsymbol{\ell}} = \frac{d(BA)}{dt} \tag{8.25}$$

which is just begging to be turned into the *flux of the magnetic field through the loop* C:

$$|\Delta V_{\rm ind}| = \oint_C \vec{\boldsymbol{E}}_{\rm ind} \cdot d\vec{\boldsymbol{\ell}} = \frac{d\phi_m}{dt} \tag{8.26}$$

where:

$$\phi_m = \int_{S/C} \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA \tag{8.27}$$

is the *magnetic flux* through the surface S bounded by the closed loop C.

The second is that if energy isn't ultimately conserved, life is going to be *bad* for physics students because magic[108] and perpetual motion machines both become possible, and yet we never seem to actually observe either one in nature. Nature is *stable*, not unstable the way it would be if induced forces *increased* the very motion that induced those forces (to make them increase even faster, with no source for the energy associated with the ever-increasing force).

We've already seen that the potential around the loop has to *increase* when we go counterclockwise in order to balance the rate that energy is *removed* from the loop by the total resistance in Kirchoff's rule combined with Ohm's Law. Eventually we're going to need to formalize this as a rule for the *sign* of the change in potential we get going around the loop in any given direction. In order for us to be able to tell somebody far away about this rule, we ought to make sure that it is based on the use of our right hands to determine loop directions relative to something that uniquely orients the problem, such as the direction of the magnetic field through the loop.

8.2.1: Problem and Solution

In the next section we will, as promised, take all of these observations and combine them into a new physical law, and a very beautiful one it will turn out to be! But yeah, let's finish off *this* problem first. Of course you may be asking *what problem*, since I haven't stated one yet. How's this: Let's find *everything* about the rod sliding on rails in this system, assuming only that it starts at time $t = 0$ moving at initial velocity v_0 to the right. That is: $I(t)$, $v(t)$, $x(t)$ and so on, we'll *find it all!* Time to use Newton's Laws once again!

We begin with:

$$\begin{aligned} \Delta V_{\text{ind}} - IR &= 0 \\ \oint_C \vec{\boldsymbol{E}}_{\text{ind}} \cdot d\vec{\boldsymbol{\ell}} - IR &= 0 \\ BLv_x - IR &= 0 \end{aligned} \tag{8.28}$$

(where we have used the results of the first section to evaluate the total induced voltage in the loop and where we've added the x subscript to v to make it clear that we are dealing only with x-directed motion and force) or:

$$I = \frac{BLv_x}{R} \tag{8.29}$$

in the direction (counterclockwise) shown around the loop.

Next, compute the force acting on the rod. I flows *up* perpendicular to $\vec{\boldsymbol{B}}$ when v_x is positive, so Newton's Second Law becomes:

$$F_x = -ILB = m\frac{dv_x}{dt} \tag{8.30}$$

[108] You might, if you are a science fiction and fantasy reader (and writer) like myself, think that it would be great fun to live in a Universe where either one was possible. Think again. Life is unstable, chaotic, and whimsical enough as it is with the *negative* feedback associated with the laws of thermodynamics; with unbounded positive feedback loops possible at all, it seems rather likely that the Universe would simply explode instantly, much the same way that positive feedback in an amplifier leads to an ear-shattering screech and (if the gain is turned up enough) blown fuses. We wouldn't want to live in a Universe with a blown fuse now, would we?

or

$$\frac{dv_x}{dt} + \frac{B^2L^2v_x}{mR} = 0 \tag{8.31}$$

which is the usual first order, linear, homogeneous ordinary differential equation and is trivially integrable. Note that I illustrate using *definite* integrals to solve the problem to avoid issues with *dimensions*, keeping both sides of the equals sign *dimensionless* from the second line on down to the last line.

$$\begin{aligned}
\frac{dv_x}{dt} &= \left(-\frac{B^2L^2}{mR}\right)v_x \\
\frac{dv_x}{v_x} &= \left(-\frac{B^2L^2}{mR}\right)dt \\
\int_{v_{0x}}^{v_x(t)} \frac{dv_x}{v_x} &= \int_0^t \left(-\frac{B^2L^2}{mR}\right)dt \\
\ln(v_x(t)) - \ln(v_{0x}) &= \int_0^t \left(-\frac{B^2L^2}{mR}\right)dt \\
\ln\left(\frac{v_x(t)}{v_{0x}}\right) &= \left(-\frac{B^2L^2}{mR}\right)t \\
e^{\ln\left(\frac{v_x(t)}{v_{0x}}\right)} &= e^{\left(-\frac{B^2L^2}{mR}\right)t} \\
v_x(t) &= v_0e^{-t/\tau}
\end{aligned} \tag{8.32}$$

with $\tau = mR/B^2L^2$ the exponential time constant of the rod's velocity as it slows down. A plot of $v_x(t)$ is shown in figure 8.4.

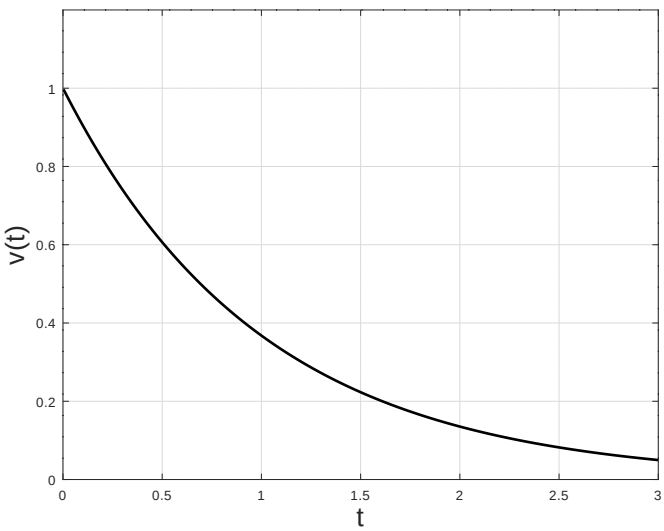


Figure 8.4: A plot of the exponential decay of the velocity of the rod as its initial kinetic energy is "burned" in heating the resistor with the induction-driven current that also slows it down. The units of the plot are v_{0x} (for v) and τ (for t) to make it "universal".

Given $v_x(t)$, we can easily find:

$$I(t) = \frac{BLv}{R} = \frac{BLv_{0x}}{R}e^{-t/\tau} = I_0e^{-t/\tau} \tag{8.33}$$

where $I_0 = BLv_{0x}/R$ is the initial current at $t = 0$. Similarly, we can directly integrate:

$$\begin{aligned} v_x(t) = \frac{dx}{dt} &= v_{0x}e^{-t/\tau} \\ dx &= v_{0x}e^{-t/\tau}\,dt \\ x(t) = \int_0^{x(t)} dx &= -v_{0x}\tau \int_0^t e^{-t/\tau}(-\tau\,dt) \\ x(t) &= v_{0x}\tau\left(1 - e^{-t/\tau}\right) \end{aligned} \tag{8.34}$$

This let's us see at a glance that the rod will (eventually, after "infinite" time) come to rest having moved down the rails a maximum distance $v_{0x}\tau = v_{0x}mR/B^2L^2$.

The magnetically induced electrical voltage produces a current that produces a force in the magnetic field that *slows the rod down*. If energy is indeed conserved, we would expect that the rate at which the kinetic energy of the rod decreases should *exactly match* the rate at which Joule heating from the current occurs in the resistor. That way the negative work done by the induction force is precisely balanced by the positive appearance of heat energy in the resistor throughout; energy isn't being created, it is just being changed from one form to another.

This is easy enough to test algebraically. The rate at which power appears in the resistor is (substututing in several results from above):

$$P_R(t) = I^2(t)R = \frac{B^2L^2v_x(t)^2}{R} = \frac{B^2L^2v_0^2}{R}\exp(\left(-\frac{2B^2L^2t}{mR}\right) \tag{8.35}$$

The rate at which work is done on the rod is:

$$P_F(t) = F_x(t)v_x(t) = (-BLI(t))\,v_x(t) = -\frac{B^2L^2v_0^2}{R}\exp(\left(-\frac{2B^2L^2t}{mR}\right) \tag{8.36}$$

which is *exactly the same* but which has, of course, the opposite sign because F is slowing the rod down! If we add the two, we see that:

$$P_R(t) + P_F(t) = 0 \tag{8.37}$$

at all times t and energy is indeed conserved! The kinetic energy removed from the rod by the induced force appears in the resistor as heat, precisely. Our "non-conservative" loop integral of the field is, in fact, conservative after all!

At this point we know pretty much everything about this loop and it is all consistent with the fundamental physical laws we have learned so far. If nothing else, the physics of the rod sliding in the magnetic field works *as if* an electric field is induced around the conducting loop which does indeed do work on the system that transforms its initial kinetic energy into heat energy in the resistor as it slows down the sliding rod.

But wait! Isn't the force involved a *magnetic* force? What about *magnetism*?

8.2.2: The Magnetic Field and Work

The astute student will have noted (from the hints I have liberally supplied) something puzzling about this problem and solution. When we looked at the rod moving through the field alone

(no rails) the rod spontaneously developed an internal region of crossed fields with an electric field that balanced the force exerted by the magnetic field on the moving charges. However, when we connected the ends of the rod with the rails, a current was established that runs in what appears to be the direction of the ***magnetic*** force and the potential decreases in the loop when it is traversed in the *opposite* direction of the static field set up in the rod moving alone!

There is a strong temptation to look at this and say "Wait a minute. It isn't an electric field pushing the charge around the loop, it is the magnetic field! What kind of swindle are you trying to pull, here?" This, in turn, might make you doubt that Faraday's Law (which we're working steadily towards) is correct at all – maybe it is just the magnetic field that is doing the work of pushing those charges around the loop!

Be careful! If you look back at the chapter on Magnetic Force, you will see that ***magnetic forces can never, ever, ever do work on spinless point charges!*** To remind you:

$$P = \frac{dW}{dt} = \vec{\boldsymbol{F}} \cdot \vec{\boldsymbol{v}} = q(\vec{\boldsymbol{v}} \times \vec{\boldsymbol{B}}) \cdot \vec{\boldsymbol{v}} = 0$$

is an *identity* of the cross product. The magnetic field itself is incapable of doing work of this sort as it can only exert forces at right angles to the direction of motion of a charged particle. We really have little choice but to believe that the electric field introduced in:

$$|V_{\rm ind}| = \oint \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{\ell}}$$

above in Kirchoff's Loop Rule for the rod on rails is "real", at least as real as the electric field we invented to describe the action-at-a-distance Coulomb force so many weeks ago.

It is, however, a worthwhile experience to run down this "work" thing, so let's look carefully at the problem and see just exactly where the work that appears as heat in the resistor does, in actual fact, come from, and in the process start to clear up just what the fields are and where they are. In figure 8.5 the *actual* motion of a (positive) point charge carrier is depicted as it moves in a rod that is *pulled by a hand* so that it moves at a constant speed v_0 to the right on the rails from the initial (dashed) position on the left (where the charge q starts to move up from the rail at the bottom) to the final (dashed) position on the right (where it reaches the top rail). I've invented and drawn some θ angles into the figure to make the important vector decompositions easier to see and algebraically evaluate.

As the charge q moves *across* the rod, it also moves *down the rails with the rod* so that its actual trajectory is the diagonal dashed line of length L, *not* the vertical line of length $\ell = L\cos\theta$. It's actual velocity $\vec{\boldsymbol{v}}$ is *also* in this direction, not straight up! The magnetic field is still into the page (I drew fewer "$\times$"s but it is still there) so the vector magnetic force $\vec{\boldsymbol{F}}_m$ acting on it is similarly not straight to the left (as it would be if the charge were really moving straight up) but at right angles to $\vec{\boldsymbol{v}}$, diagonally up and to the left as drawn. We see that the speed of the rod to the right $v_0 = v\sin\theta$ is just the *horizontal component* of the actual velocity of the charge!

Note well that I've drawn the "hand" that pulls on the rod hard enough to maintain it at a constant speed v_0 (doing work all of the while). When the test charge turns the corner to start moving up the rod, the *rod* has to exert a force on the charge – $\vec{\boldsymbol{F}}_{\rm rod}$ – that accelerates it in the x direction this speed or it would come out of the rod as it is left behind! While *moving* at the *constant* x-directed speed v_0, the force $\vec{\boldsymbol{F}}_{\rm rod}$ has to *precisely cancel* the x component of

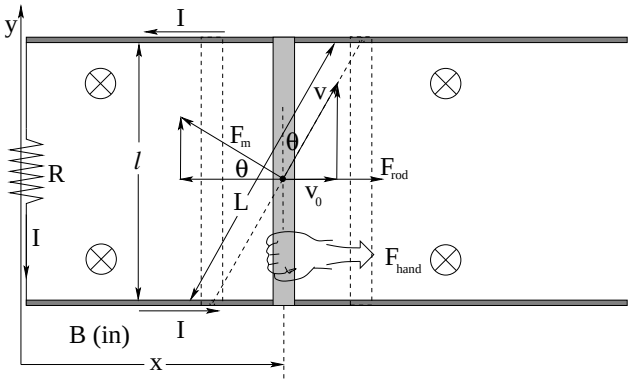


Figure 8.5: The *actual* motion of the charge q is along a diagonal of length L. A "hand" pulls the rod, which transfers some (electrostatic) force $F_{\rm rod}$ to the charge to overcome the component of the magnetic force pulling it back.

the magnetic force in the opposite direction, or the charge would come out of the rod and be left behind.

Of course, the charge does *not* come out of the rod, for the same reason that *all* of the conduction charges in the rod – however free they are to move *inside* the rod – do not come out. *Electrostatic* forces exerted at the surface of the rod prevent charge from coming out until pushed by a force great enough to cause dielectric breakdown, and internal *electrostatic* forces otherwise cancel the field *inside* the rod in all directions except that of current flow (where $\vec{\boldsymbol{E}} = \rho_r \vec{\boldsymbol{J}}$ for a rod with a non-zero resistivity)! It is even *electrostatic* forces which, with an assist from quantum mechanics and the Pauli exclusion principle[109], are ultimately responsible for the "normal" force between the hand and the rod that exerts the external force on the rod itself!

Now that we have this concept firmly in hand – errr – mind, we can do some geometry and some *algebra*. The magnitude of the magnetic force is $F_m = qvB$, but now v is the magnitude of $\vec{\boldsymbol{v}}$ along the diagonal, *not* v_0 (it's x component). As argued above, the *total* force in the horizontal direction on the charge:

$$F_{{\rm rod},x} + F_{m,x} = 0 \tag{8.38}$$

so it will remain inside the rod as it travels at the constant speed in the x direction $v_0 = v \sin\theta$. The horizontal component of $\vec{\boldsymbol{F}}_m$ is (from basic trig, using the angle θ defined in figure 8.5):

$$F_{m,x} = -F_m \cos\theta = -qvB\cos\theta \tag{8.39}$$

so the force exerted on the charge by the rod that cancels this must be:

$$F_{{\rm rod},x} = qvB\cos\theta \tag{8.40}$$

This electrostatic force does the net positive work:

$$W_q = F_{\rm rod} L \sin\theta = qvBL\cos\theta\sin\theta \tag{8.41}$$

[109]Beyond the scope of this course, but remember these words or pursue the topic further on e.g. Wikipedia if you are a physics major!

on the charge as it moves diagonally along the line L, starting at the bottom rail and ending at the top. You can think of this as either the component of the rod force in the L direction $F_{\rm rod}\sin\theta$ times L, or the component of the L displacement in the direction of the force $L\sin\theta$ times $F_{\rm rod}$; the net work done is the same either way.

We rearrange this as follows:

$$W_q = q(v\sin\theta)B(L\cos\theta) = q(v_0 B\ell) = qV_{\rm ind} \tag{8.42}$$

and see that the work done on the charge q (ultimately) ***by the hand***, transmitted through the *electrostatic field binding the rod and charges together* as it moves between the rails, is *exactly* what one gets if one (mistakenly) claims that it is the magnetic force acting on q to drive it vertically that does the work which in turn is *exactly* what you get if you consider the work to be done by an induced electric field/potential around the entire conducting loop!

Note well that the Newton's Third Law reaction force tells us that the force on the charge is transferred back to the *rod* through the same electrostatic binding force, so that the entire rod itself is pulled back by the magnetic force with exactly the overall force $F_{\rm hand}$, which is what really does all of the work. This is in perfect agreement with the result argued for somewhat differently above. The one last thing we might need to consider is why this induced field seems to follow nonzero resistivity around, but this is no different from the way that there is almost no field inside a *good* conductor – we treat wires in electric circuits as being equipotential even if they carry (moderate) currents – but recognize that charge piling up at the ends of a *poor* conductor results in an electric field inside that drives the charge on through against its resistance.

In the next section we will clearly state the conclusions of the first two sections of this chapter in the form of a single equation: Faraday's Law.

8.3: Faraday's Law

In the last section, we saw that for the rod sliding down the rails (at least) we could describe the voltage induced around the closed loop formed by the rails as the time rate of change of the magnetic flux through the loop. We left open the question of how to specify the *direction* of the induced E-field, although clearly we have to have just the right sign (direction) in order for energy to be conserved as it was for the rod and resistor together.

If we point our right hand's thumb in the direction of the magnetic field through the loop in the previous section and let its fingers curl around the loop, this is in some sense "the" natural direction to specify as the "positive" direction for $d\vec{\ell}$ in the loop (clockwise as drawn in figures 8.3 and 8.5). In this case an *increasing* loop area and *increasing* flux produced a *negative* directed electric field (counterclockwise as drawn in figure 8.5) and the induced current went in this direction (which is also the direction of the electric field we "imagined" appearing in the rod in a frame where the rod is *stationary* and it is the *magnetic field* that swept across the loop). This in turn made the force on the rod *negative* (in the opposite direction of its velocity $\vec{v}$) as it ***had to be***, it turned out, for ***energy to be correctly conserved***.

This suggests that we could have written the voltage that appears in the loop completely

consistently with respect to magnitude and direction using this "right hand rule" as:

$$V_{\text{induced in C}} = \oint_C \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{\ell}} = -\frac{d}{dt}\int_{S/C} \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA \tag{8.43}$$

This equation is known as **Faraday's Law** and is our first truly dynamical field equation for the electromagnetic field. It tells us that *changing magnetic flux* through an arbitrary loop creates an *electric field around the loop*.

The minus sign on the right hand side tells us the direction of this field – if we let the fingers of our right hand curling around the loop as our thumb points in the (predominant) direction of $\vec{\boldsymbol{B}}$ through the loop, then if the flux through the loop is increasing the E-field circulates the loop C in the negative direction, *opposite* to the direction our right-handed fingers curl around the $\vec{\boldsymbol{B}}$ field. Similarly, if the flux through the loop is *decreasing* the E-field circulates around C in the positive direction, that is to say, *in* the direction our right-handed fingers curl around the magnetic field through the loop.

Note that (to be sure!) magnetic fields can easily go "through" any given surface bounded by a loop first in one direction, then the other, but because this integral is linear in $\vec{\boldsymbol{E}}$ and $\vec{\boldsymbol{B}}$, we can actually pick either of the two directions possible to be right-handed positive, use superposition, and get the correct answer *relative* to this choice. The fact that we pick the obvious direction in cases where it *is* obvious doesn't mean the equation can't handle cases where it isn't obvious or even isn't known initially at all, forcing you to arbitrarily choose a direction to be "positive" for both $\hat{\boldsymbol{n}}$ in the flux expression and $d\vec{\boldsymbol{\ell}}$ around the loop!

The information encoded in this humble minus sign (which leads to energy conservation) is so important that it has a name of its own – it is called **Lenz's Law**. Lenz's Law can be stated a different way in words as well:

> The electric field induced in a loop by changing magnetic flux goes around the loop in the direction such that any current generated by the field will create a magnetic field of its own that *opposes the change* in the magnetic flux.

This is a very interesting result, and is worth studying for a moment all by itself before returning to the many *applications* of Faraday's Law.

First, though, we have to consider a serious question. Our discussion up to now has involved *current carrying loops of conducting material*. To put it another way, C (the closed loop that appears in our integral of $\vec{\boldsymbol{E}}$) is a conducting pathway. However, we also saw that – in our imaginations, at least – a field appeared when we jumped into a frame of reference where an *isolated* rod was stationary! What happens, then, if we *remove the rod entirely?* Then we are left with *just space*, with a magnetic field sweeping across it. Does this mean that there is an electric field there?

This is not a question theoretical physics can answer. In science, it is an *empirical* question, one that can only be answered by experiments! In particular, what happens if one applies equation 8.43 to a loop in empty space? Does an electric field appear when we change magnetic flux through the loop, or doesn't it? Also, does the answer to this question on their actually *being* a magnetic field at the point in space in question? Note that equation 8.43 ***does not require this*** – one could imagine a loop in space around a *toroidal solenoid* or *infinitely*

long solenoid such that there is no magnetic field at all (or a truly tiny one) at the point in question while the electric field there would be quite large!

As it turns out, equation 8.43 is indeed ***completely general***. At this point, uncountable experiments and a large mountain of evidence have demonstrated that it holds at points in space without any need for a charge to be there at all, without the need for any "conducting path" at all[110]. Therefore, from now on we will simply drop the very *idea* of a *conducting* pathway as a requirement for a discussion of:

$$\oint_C \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{\ell}} = -\frac{d}{dt}\int_{S/C} \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA \tag{8.44}$$

and call equation 8.44 ***Faraday's Law*** (with Lenz's Law embedded as the minus sign in the term on the right). This is our ***fourth Maxwell equation***, valid for *arbitrary closed curves C in space* whether or not there is a conductor, or nonzero charge, at any point on those curves.

The existence of the induced electric field in free space even where there are no charges or conductors is key to our later development of the dynamic electromagnetic field – it suggests that the induced E-field can *propagate* through empty space as long as there is a changing magnetic field present *somewhere* to produce it, even with no charges or conductors locally handy for the field to act on.

Faraday's Law is truly a sublime result. As we will see, this Maxwell Equation is directly responsible for our ability to generate and transmit electrical energy to run our homes, our businesses, our industries, our entertainments, our lives. If it were not for Faraday, I would at best be laboriously typing this textbook on a mechanical typewriter by candlelight and you would not be able to read it until a publisher (at great expense) typeset the entire book and printed it with a steam or water driven press to sell for a small fortune, making its contents available only to the fortunate and the wealthy.

Instead you are very likely reading a purely electronic version of the textbook that you got for free, or perhaps paid a pittance for as a gesture of courtesy to the author[111], all thanks to electricity generated via Faraday's Law and transmitted as electromagnetic wave energy and processed in countless ways inside your computer that also rely completely on Faraday's Law. Each and every one of these carefully engineered occurrences is an "experimental test" of Maxwell's Equations in general and Faraday in particular, so you can have a great deal of confidence that Maxwell's equations (and the associated Lorentz force law) are at the very least a very good approximation to some true underlying principle or law of nature.

In the next section, we will discuss Lenz's Law and give several examples of using it either algebraically or conceptually to determine the direction of the induced electric field around a loop, as promised.

[110] Fans of "causality" might wish to assert that there must be *some* sort of field that is nonzero at the point where the induced $\vec{\boldsymbol{E}}$-field appears. They will be pleased to learn that there is indeed a *vector potential* that is generally *not* zero at the points where the $\vec{\boldsymbol{E}}$ field appears, one of many reasons physicists prefer potentials to fields.

[111] Yes, that's me, and if you aren't a Duke student you should very much consider the virtue of such courtesy and how it enables high quality, cheap textbooks to be created and improved for your delight and edification...

8.4: Lenz's Law

Lenz's Law, as we have just seen, tells us in a general, mathematically consistent way, what the direction is of the induced E-field around a loop through which magnetic flux is *changing in time* regardless of the mechanism of that change in flux and whether or not there are charges or a conductor handy to produce or contain currents. However, if you think about the equation for the magnetic flux through some surface S bounded by a closed curve C:

$$\phi_m(t) = \int_{S/C} \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA \tag{8.45}$$

you will soon realize that the flux ϕ_m can vary in time for any or all of *four reasons*:

a) C can change in time (and hence so can S).

b) The magnitude of $\vec{\boldsymbol{B}}$ can change in time.

c) The angle between $\vec{\boldsymbol{B}}$ and $\hat{\boldsymbol{n}}$ can change in time because the direction of $\vec{\boldsymbol{B}}$ changes.

d) The angle between $\vec{\boldsymbol{B}}$ and $\hat{\boldsymbol{n}}$ can change in time because the direction of $\hat{\boldsymbol{n}}$ changes.

Yes, one can imagine a loop that is changing its size and its orientation inside a magnetic field that is changing its magnitude and *its* orientation, all four changes in time contributing to the overall change in magnetic flux through a surface S bounded by the loop! This multiplicity of ways the magnetic flux depends on geometry and field strength *can* make it difficult to figure out the direction of the induced field. In this section, we will endeavor to provide examples of each of these *separately* to help you see how it all goes. With a bit of meditation, you should then be able to figure out how to synthesize this knowledge and work out the direction when multiple things are changing at once.

8.4.1: Lenz's Law for changing C

We've already seen an example of this in our single meaningful example this far – the rod on rails. If a plane loop C in a fixed magnetic field is *increasing in size*, then the induced field points opposite to the right handed direction determined from the magnetic field through the loops. If it is decreasing, it points around the loop C in the *same* right handed sense.

In terms of the verbal statement (illustrated in figure 8.6), if a conductor of resistance R were placed along a path C *increasing* in area (in (a)), the current in the loop thus formed would have a magnetic moment that *opposes the increasing flux* through the loop. Incidentally, the magnetic force acting on this current would point *in* towards the center of the loop which is the direction that makes the loop try to *shrink*, not grow, opposing again the increase in flux.

If the conducting loop were decreasing in area (in (b)), the induced current would be in the direction that creates a magnetic moment for the loop *in* the same direction as the magnetic field through the loop, again *opposing the* (now decreasing) *change in flux*. This direction for the current also creates a general *outward* directed force on all parts of the loop, which would make the loop *grow* to oppose the decrease in flux, if e.g. the rails were not rigidly affixed.

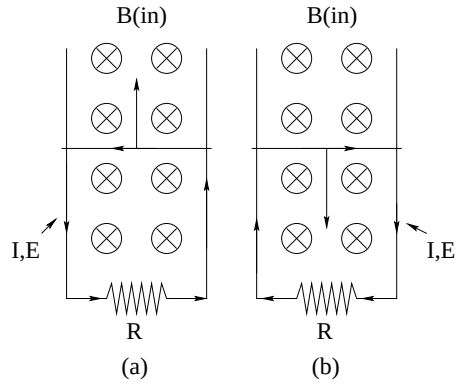


Figure 8.6: Illustration of $\vec{E}$-field direction for loops that change size. In (a) the loop is getting larger (tending to increase the magnetic flux) so the induced magnetic moment from a counterclockwise $\vec{E}$ field and current opposes the existing field through the loop. In (b) the loop is getting smaller (tending to decrease the flux) so the induced magnetic moment from the clockwise $\vec{E}$ field and current supports the existing field through the loop.

8.4.2: Lenz's Law for changing B (magnitude)

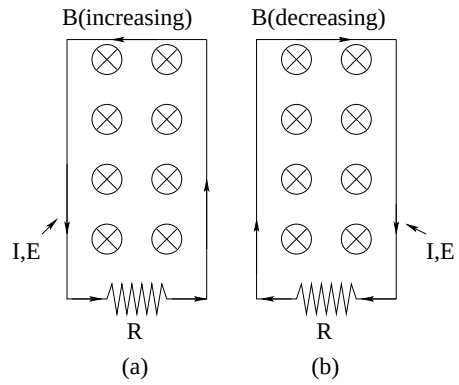


Figure 8.7: Illustration of $\vec{E}$-field direction when the magnitude of B through the loop changes. In (a) B is getting larger (tending to increase the magnetic flux) so the induced magnetic moment from a counterclockwise $\vec{E}$ field and current opposes the existing field through the loop. In (b) B is getting smaller (tending to decrease the flux) so the induced magnetic moment from the clockwise $\vec{E}$ field and current supports the existing field through the loop.

In figure 8.7 we illustrate what happens when the *magnitude* of the B-field changes. In (a), B is increasing in magnitude through a fixed loop while maintaining a fixed direction. Again if we imagine a conducting pathway around C the (counterclockwise as shown with $\vec{B}$ into the page) current induced in it would create a magnetic moment from the loop that is in the *opposite* direction as $\vec{B}$, opposing the change in flux. The forces acting on this current in each wire of the loop would point *inward*, trying to *shrink the loop* as an alternative way of reducing the flux.

In (b), B and the magnetic flux are decreasing in magnitude and the opposite happens – the induced moment would create an $\vec{E}$-field and associated current that circulate in the (clockwise) direction such that the induced magnetic moment *supports* the decreasing $\vec{B}$ field (opposing the change in flux). The magnetic forces on the loop wires would point *outward*,

trying to *expand the loop* as an alternative way of increasing the flux.

8.4.3: Lenz's Law for changing the $\vec{B}$ and/or $\hat{n}$ direction

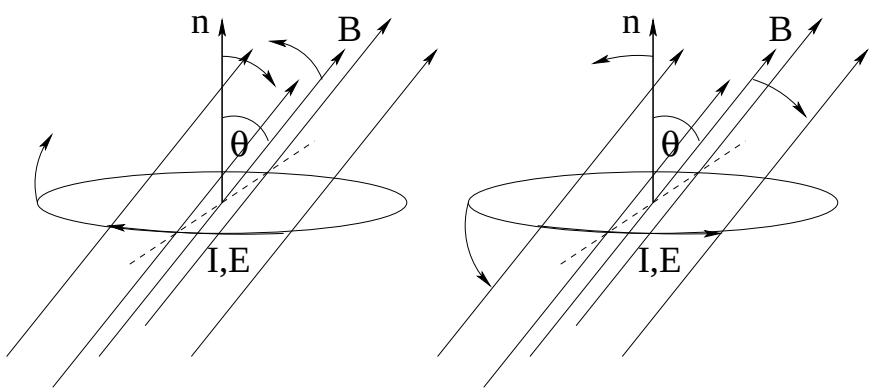


Figure 8.8: Illustration of $\vec{E}$-field direction when the direction of $\vec{B}$ *or* the direction of the normal to the loop $\hat{n}$ changes. In (a) $\cos(\theta)$ is getting larger (tending to increase the magnetic flux) so the induced magnetic moment from a counterclockwise $\vec{E}$ field and current opposes the existing field through the loop. In (b) $\cos(\theta)$ is getting smaller (tending to decrease the flux) so the induced magnetic moment from the clockwise $\vec{E}$ field and current supports the existing field through the loop.

Now we imagine the shape of the loop C doesn't change, the (uniform) magnetic field is constant in magnitude, but the loop's *orientation* in the magnetic field is changing or the *direction* of the magnetic field is changing (or both). Note that both changes have the same effect: they alter the *angle* between the field and the normal to the plane of the loop, and hence the flux through the loop. This is actually a very common situation – it describes an electrical generator or electrical motor rather well.

If $\vec{B}$ and/or $\hat{n}$ are rotating *into* relative alignment about the dashed line axis shown (that is, decreasing θ and hence increasing $\cos\theta$ and the flux) as shown in (a) of figure 8.8, the field direction and induced current are *clockwise* when viewed from above the loop to make the induced magnetic moment opposite to $\vec{B}$. If they are rotating *out* of alignment as shown in (b), $\cos\theta$ is getting more negative and the flux is decreasing, so the induced moment will support the B-field, resulting in a counterclockwise current viewed from above the loop.

8.4.4: Lenz's Law Summary

Note that it is entirely possible for all four of these contributions to the total flux to be changing at once. The loop and field could both be rotating, the loop could be shrinking or growing, and the field could be turning on or turning off *all at the same time!* Problems where all of this is going on at once are a bit excessive, perhaps, largely because it is such a pain to specify all of the possibly competing parameters, but *in principle* you know what you need to know to determine the $\vec{E}$-field/current direction from Lenz's Law. It will always point in the direction such that a magnetic moment associated with a current in the induced $\vec{E}$-field direction (whether or not one actually exists) would *oppose the change in magnetic flux through the loop*.

This is precisely the right direction for energy conservation to always hold for the system. We can breathe a sigh of relief!

There is one more observation we can make that can help you solve direction problems involving Lenz's Law – a whole new way of stating it that should make physical sense to you. When C was changing, the forces acting on the loop all opposed the change – if it was expanding the induced forces tried to keep it from expanding. If it was contracting, the induced forces tried to keep it from contracting. When B's magnitude was increasing, the forces tried to make the loop smaller, opposing the increase in flux by decreasing its area. When it was decreasing, the forces tried to make the loop larger for the same reason! In the final case, where the angle θ between the field and the normal was decreasing (increasing the flux), the induced forces on the loop itself would be trying to make it decrease its area *and* the *induced magnetic torque* on the loop would be acting in the direction to make θ *larger*, *opposing* the external torque that is causing θ to shrink. If the angle θ is increasing, the forces will try to *increase* the area of the loop and the induced magnetic torque will again opppose the torque increasing the angle θ, trying to make it smaller.

In summary, ***every aspect of physics*** resulting from the change in flux ***opposes the change***. The magnetically induced current creates a field that augments or opposes a decreasing or increasing magnetic flux through the loop, respectively. The magnetic force on the induced current in the loop will act to increase or decrease the area as the flux through it decreases or increases, respectively (opposing the change). The induced magnetic torque on the loop will tend to rotate it so its normal is more parallel or at right angles to the field through it as the flus through the loop decreases or increases, respectively, (opposing the change). If the loop is in an non-uniform field, the *total* force acting on the loop will point in the direction where the field is weaker or stronger as the flux increases or decreases, respectively, ***opposing the change***. We can rewrite our verbal statement of Lenz's Law as:

> **The direction of the induced electric field/current in Faraday's Law will be the one that causes *all physical reactions caused by the induced current to oppose the change!***

If the flux through a current loop is increasing, the loop will *simultaneously try to shrink in size, rotate to an angle at right angles to the loop, move overall in the direction the field strength decreases if any, and generate its own magnetic moment in the opposite direction to the increasing field.* If it is decreasing, it will do the exact opposite: increase its size, move to stronger field, torque its normal until it is parallel to the field, and augment the decreasing field with its own induced magnetic moment. This means that we will always need to be doing external work *against* these Lenz's Law reactions to bring about the changes, causing work and energy to remain in perfect balance.

Example 8.4.1: Wire and Rectangular Loop – Direction Only

In figure 8.9 above, a long straight wire is carrying a current I. It sits a distance d away from a rectangular loop with side lengths of a and b (all wires in the plane of the page) as shown. I can be increased or decreased at will.

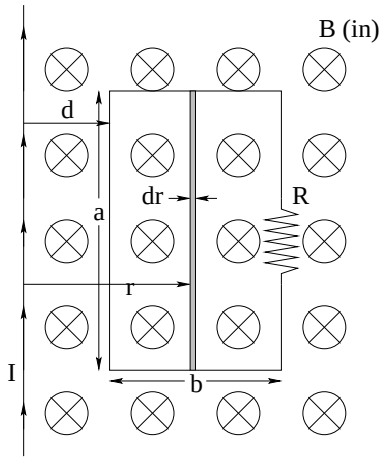


Figure 8.9: A long straight wire sits next to a rectangular loop of wire and carries a current I *up* as shown. The current in the long straight wire can be increased or decreased.

Here's the physics of this picture. The current I creates a magnetic field through the loop. We can easily compute that field using Ampere's Law (so we don't have to remember things like the magnetic field of long straight wires). On the other hand, we've worked enough with the magnetic field of long straight wires that perhaps you do remember that it is $\frac{\mu_0 I}{2\pi r}$, *into* the page on the right for a current up as drawn – I've helped you out a bit with lots of "dressing" on this figure that on a quiz or exam you'd have to provide for yourself.

If I is varied, the field it generates varies as well. This changes the magnetic flux through the rectangular loop. Mr. Faraday then tells us that there must be a voltage induced in the loop that will create a current!

You can actually *completely calculate* the induced voltage in the rectangular loop using Faraday's Law (and will, in a homework problem) and from the voltage compute the current in the loop, and from the current the force on the loop. But here our goal is more humble. We simply want to figure out the *direction* of the induced current, and the *direction* of the induced force, using Lenz's Law.

Suppose the current I is *increasing*. Then we expect the magnetic field into the page – and the magnetic flux through the loop – to be increasing as well, and we can tell the following (highly anthropomorphized) story:

The increasing flux makes the loop *sad*, because it is a very conservative loop. It hates change, and is happy with things just the way that they are. It says to itself "Gosh, I'd really rather the magnetic flux through me *not* change, what can I do?" It then has the brilliant idea: Create an electric field to drive a current around itself so that its own magnetic moment opposes the change in flux! Perhaps it won't keep the flux from changing altogether, but it will ensure that at least the flux will change *more slowly* than it would without the induced current.

But which way is that? Well, a clockwise current would make the moment of the loop point *into* the page, which would make the field through the loop even stronger, so that won't work. Instead the reactionary little loop makes the current counterclockwise. Now its own magnetic field *opposes* the field due to the wire, and slows the rate of change of magnetic flux through itself. Eventually, of course, the field might reach a new constant value as the current in the

long straight wire stops changing and the loop becomes happy again with constant flux through it and no current at all.

The induced current in the counterclockwise direction has an additional bonus for the loop. It makes the net force on the loop point *away* from the wire (as you can verify when you solve the problem completely). If the loop is free to move, moving away from the wire moves it from a strong field near the wire to a weaker field farther away from the wire! This, too, helps to keep the flux through the loop from increasing, and is a part of the responses predicted by Lenz's Law. If the loop is even *slightly* tipped relative to the field, then there will be a nonzero *torque* on it as well, trying to twist it towards right angles relative to the field, reducing the flus through it in that way as well! Again, *every* reaction of the loop to the increasing flux will physically tend to oppose the increase in flux!

When you do this problem for homework, you will have to compute the net magnetic flux through the loop (in order to differentiate it to find the induced voltage). I've helped you out here by shading a strip of length a and width dr, a distanced r from the main wire. It should be pretty easy to compute the flux $d\phi_m$ through this strip, and then to sum up the total flux using integration between suitable limits. Give it a try.

Example 8.4.2: Rectangular Loop Pulled from Field

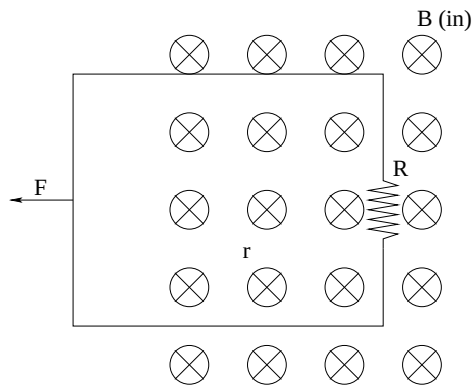


Figure 8.10: A rectangular loop of wire is pulled out of a region of uniform magnetic field as shown.

In figure 8.10 you can see a wire loop (rectangular, although this makes no real difference) being pulled from the field. A typical short answer question might show this picture, or a similar picture, of a loop of any shape you like being pushed into or pulled out of a magnetic field and ask you the following questions:

- What is the *direction* of the induced $\vec{E}$-field/current in the wire as it is being pulled out (or pushed in)?
- What is the direction of the *magnetic force* acting on the loop while this is going on (in either direction)?
- A trick question might show you the loop *completely inside* the uniform field (so it isn't actually coming out!) and ask the same questions.

What are the answers?

- When the loop is being pulled out, the flux through the loop is *decreasing*. The sad little loop doesn't want the flux to go away, so it generates a clockwise current whose magnetic field sustains the disappearing flux.
- The net force on this current *resists the motion of the loop out of the field*. Check it yourself!
- If the loop were entirely in the field, the flux *wouldn't be changing as it moved* and there would be *no current and no net force*.

This example is *almost* identical to a rod on rails problem, is it not? For a specified geometry and mass m of wire loop and speed v, you might well be able to *compute* the current, the force, the acceleration, the trajectory.

8.5: More Rod on Rails Problems

Example 8.5.1: Rod on Rails with Battery

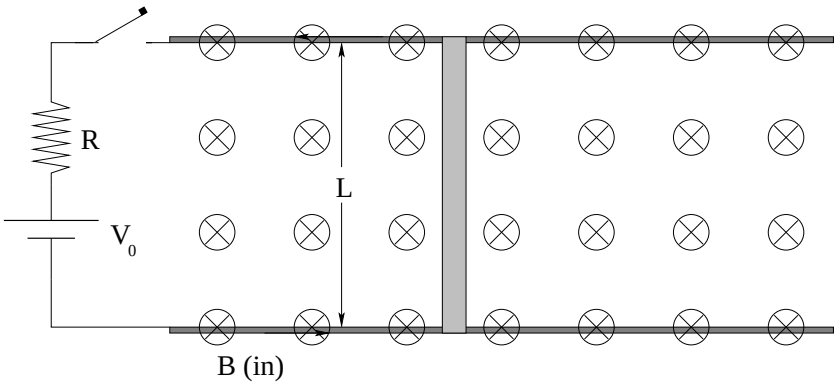


Figure 8.11: A conducting rod sits on conducting, frictionless rails and a switch is closed at $t = 0$ to send current through the loop thus formed. A magnetic field (into the page) exerts a force on the rod.

In figure 8.11 above, the switch is closed at time $t = 0$ with the rod (of mass M and length L) sitting at rest on a pair of frictionless conducting rails that are on the other end connected by a resistor R and battery with potential difference V_0. A uniform magnetic field of magnitude B points into the page as shown.

We would like to find a number of things in this problem:

a) The voltage in the loop as a function of v, the velocity of the rod (at some instant in time t).

b) The current in the loop as a function of this voltage.

c) The force on the rod as a function of this current.

d) The *terminal* velocity of the rod, after the switch has been closed for a long time.

e) The equation of motion of the rod as determined by the force.

f) The velocity of the rod as a function of time.

This list lays out a very nice solution strategy. Using Faraday's Law

$$V_{\text{ind}} = -\frac{d\phi_m}{dt} = -\frac{dBLx}{dt} = -BLv \tag{8.46}$$

(where the minus sign is Lenz's Law and must be interpreted accordingly). Note that the induced voltage is zero until the rod is moving, then decreases in the direction that will cause currents that experience forces that oppose the motion.

Using Kirchoff's rule for the loop:

$$V_0 - BLv - IR = 0 \tag{8.47}$$

We can then solve for the current in the loop:

$$I = \frac{V_0 - BLv}{R} \tag{8.48}$$

and will circulate *clockwise* (positive loop direction) in the loop initially when v is small.

This lets us easily compute the force on the loop:

$$F = BLI = \frac{BLV_0 - B^2L^2v}{R} \tag{8.49}$$

At this point we can see that the force that results as v increases will produce an asymptotic approach to a *terminal* velocity such that that the net force on the loop (and hence current in the loop) will be zero. Using either the force or the current equation above we can easily see that:

$$v_{\text{terminal}} = \frac{V_0}{BL} \tag{8.50}$$

Alternatively, using the force equation we can write Newton's second Law and turn it into an equation of motion:

$$F = \frac{BLV_0 - B^2L^2v}{R} = Ma = M\frac{dv}{dt} \tag{8.51}$$

which we can rearrange into a *first order, linear, inhomogeneous, ordinary differential equation*:

$$\frac{dv}{dt} + \frac{B^2L^2}{MR}v = \frac{BLV_0}{MR} \tag{8.52}$$

As usual, this equation is simple enough to directly integrate:

$$\begin{aligned}
\frac{dv}{dt} &= \frac{BLV_0 - B^2L^2v}{MR} \\
&= -\frac{B^2L^2}{MR}\left(v - \frac{V_0}{BL}\right) \\
\frac{dv}{\left(v - \frac{V_0}{BL}\right)} &= -\frac{B^2L^2}{MR}dt \\
\int \frac{dv}{\left(v - \frac{V_0}{BL}\right)} &= -\int \frac{B^2L^2}{MR}dt \\
\ln\left(v - \frac{V_0}{BL}\right) &= -\frac{B^2L^2}{MR}t + C \\
v - \frac{V_0}{BL} &= e^{-\frac{B^2L^2}{MR}t} * e^C \\
v(t) &= \frac{V_0}{BL}\left(1 - e^{-\frac{B^2L^2}{MR}t}\right) \qquad (8.53)
\end{aligned}$$

where we've used our initial condition, $v(0) = 0$, to set the constant of integration. Note well that this curve represents an *exponential approach to the terminal velocity*:

$$v(t) = v_{\text{term}}\left(1 - e^{-t/\tau}\right) \qquad (8.54)$$

where $\tau = MR/B^2L^2$ is the same as it was for our original rod on rails problem.

With this in hand we can easily integrate over time again to get $x(t)$, differentiate it to get $a(t)$, substitute it to get $I(t)$ or $F(t)$. We can compute the power being delivered to the circuit by the voltage and show that it equals the rate at which energy is burned in the resistor plus the rate that work is being done on the rod. We can answer *anything* asked about the rod – the motion is now *completely known* subject to the usual idealizations in the problem (no friction or drag and so forth).

8.6: Inductance

We have seen that changing the current in *one* wire causes the magnetic field associated with that current to change in time. That, in turn, will usually cause the magnetic flux through other nearby conducting loops to change in time. This, according to Faraday's Law, will induce a voltage around those loops and, assuming they have some resistance, cause current to flow in the direction predicted by Lenz's Law.

For *loops of fixed size and orientation*, the field produced by them at any given point in space is *directly proportional* to the current they carry (from the Biot-Savart Law, which contains the current in the wire on top and constant so it can be pulled out of the integral over the geometry of the wire). The magnetic flux both through the loop itself and through all *other* loops that its field passes through is thus *also* proportional to the current.

This general state of affairs is pictured in figure 8.12. In this figure, loop 1 (we suppose) carries a current I_1. At the instant shown, this current produces a magnetic that swirls up through loop 1 in field line loops that go around the current in the right-handed direction. These

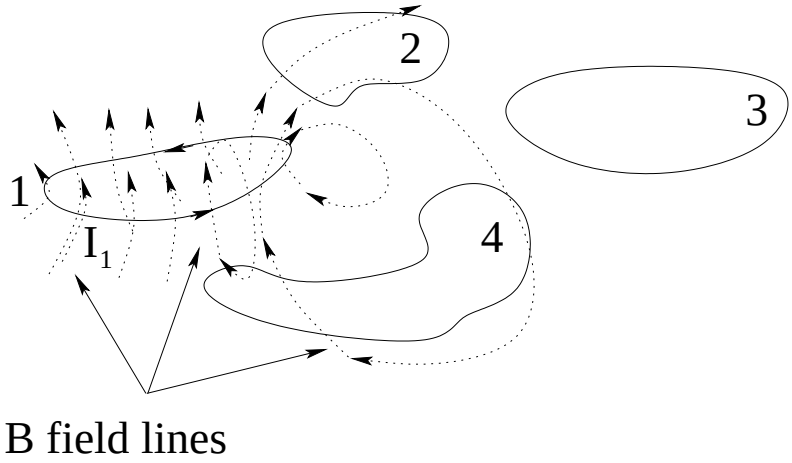


Figure 8.12: A set of current loops indexed by $i = 1, 2, 3...$, fixed in space and carrying currents I_i. The B-field produced by (say) current I_1 swirls around the current and passes through both loop 1 and the other loops in the figure, creating both *self inductance* and *mutual inductance*.

field lines pass both through any surface S_1 we might draw that is bounded by the curve C_1 and through the surfaces $S_{i\neq 1}$ bounded by the other curves C_i. These fields create *magnetic flux* that is proportional to I_1 in all of the loops.

We can write this in an algebraic form. The flux through the ith loop caused by the current in the jth loop is:

$$\begin{aligned}
\phi_{ij} &= \int_{S_i/C_i} \vec{\boldsymbol{B}}_j \cdot \hat{\boldsymbol{n}}_i dA_i \\
&= \frac{\mu_0}{4\pi} \int_{S_i/C_i} \left(\int_{C_j} \frac{I_j d\vec{\boldsymbol{l}}_j \times (\vec{\boldsymbol{r}}_i - \vec{\boldsymbol{r}}_j)}{|\vec{\boldsymbol{r}}_i - \vec{\boldsymbol{r}}_j|^3} \right) \cdot \hat{\boldsymbol{n}}_i dA_i \\
&= \frac{\mu_0}{4\pi} \left(\int_{S_i/C_i} \int_{C_j} \frac{d\vec{\boldsymbol{l}}_j \times (\vec{\boldsymbol{r}}_i - \vec{\boldsymbol{r}}_j)}{|\vec{\boldsymbol{r}}_i - \vec{\boldsymbol{r}}_j|^3} \cdot \hat{\boldsymbol{n}}_i dA_i \right) I_j \\
&= M_{ij} I_j
\end{aligned} \tag{8.55}$$

where I've take some pains to label the coordinates with the object: $\hat{\boldsymbol{n}}_i$ normal to the surface S_i bounded by the curve C_i, where dA_i is the area element of this surface and $\vec{\boldsymbol{r}}_i$ the vector coordinate of a point on its surface; coordinates $d\vec{\boldsymbol{l}}_j$ and $\vec{\boldsymbol{r}}_j$ on the curve C_j.

There are a few very interesting things to observe about this pair of integrals. One is that the integral over the surface S_i *cannot depend on the particular surface chosen* out of the *infinite* number of surfaces S_i bounded by any particular curve C_i. Understanding how integrals like this can be invariant as one selects different surfaces will be a key aspect of our addition of the Maxwell Displacement Current in two more weeks, so consider this a hint.

Ultimately, it can therefore only depend on C_i itself, so both integrals *can* be represented as integrals around the closed loops C_i and C_j using theorems from multivariate calculus that

you do probably do not yet know[112] . The result is (eventually):

$$\begin{aligned} M_{ij} &= \frac{\mu_0}{4\pi}\left(\int_{S_i/C_i}\int_{C_j}\frac{d\vec{\boldsymbol{l}}_j\times(\vec{\boldsymbol{r}}_i-\vec{\boldsymbol{r}}_j)}{|\vec{\boldsymbol{r}}_i-\vec{\boldsymbol{r}}_j|^3}\cdot\hat{\boldsymbol{n}}_i dA_i\right) \\ &= \frac{\mu_0}{4\pi}\oint_{C_1}\oint_{C_2}\frac{d\vec{\boldsymbol{l}}_i\cdot d\vec{\boldsymbol{l}}_j}{|\vec{\boldsymbol{r}}_i-\vec{\boldsymbol{r}}_j|} \end{aligned} \tag{8.56}$$

which is obviously symmetric under interchange of i and j:

$$M_{ij} = M_{ji} \tag{8.57}$$

for any two loops C_i and C_j carrying currents I_i and I_j respectively.

Of course we've formulated this result in a completely general way, but for *arbitrary* conducting pathways M_{ij} hides a whole lot of integration evil that we just won't be able to manage. In simple cases, however, we *can* evaluate it analytically (and we will, in examples and for homework), and in others we can evaluate it numerically, and when both of these fail we can at the very least *measure* it in a lab, so this is a useful decomposition. We call the M_{ij} the *mutual inductance* of the ith and jth circuit and give it a set of SI units all its, own, *Henries*. We will specify Henries more precisely shortly, as they are still obscure.

Note that there is no real reason for $i \neq j$ in this expression. There is a magnetic field through the loop C_i due to the current I_i *in* C_i; this current creates a flux through the loop due to its own current:

$$\begin{aligned} \phi_{ii} &= \int_{S_i/C_i}\vec{\boldsymbol{B}}_i\cdot\hat{\boldsymbol{n}}_i dA_i \\ &= \frac{\mu_0}{4\pi}\int_{S_i/C_i}\left(\int_{C_i}\frac{\mu_0 I_i d\vec{\boldsymbol{l}}_i{}'\times(\vec{\boldsymbol{r}}_i-\vec{\boldsymbol{r}}_i{}')}{|\vec{\boldsymbol{r}}_i-\vec{\boldsymbol{r}}_i{}'|^3}\right)\cdot\hat{\boldsymbol{n}}_i dA_i \\ &= \frac{\mu_0}{4\pi}\oint_{C_1}\oint_{C_1}\frac{d\vec{\boldsymbol{l}}_i\cdot d\vec{\boldsymbol{l}}_i{}'}{|\vec{\boldsymbol{r}}_i-\vec{\boldsymbol{r}}_i{}'|}I_i \\ &= M_{ii}I_i \\ &= L_i I_i \end{aligned} \tag{8.58}$$

where we define the *self-inductance* of the ith loop to be the symbol L_i. Note that I had to add primes to the "j" coordinates in the previous expression to differentiate between the integral over the current loop and the integral over the area.

In practical terms, the self-inductance will be very important to us as design elements in electronic circuits designed to process information and as an important aspect of any piece of electrical equipment based on coils of wire with many turns, e.g. electrical motors and generators.

Inductance is the magnetic equivalent of capacitance. Inductances can (as we will see) store energy, generate voltages, and do many useful things for us. Before we move on to see how by actually computing inductances and the potentials they can generate, we should

[112]Wikipedia: http://www.wikipedia.org/wiki/derivation of self inductance. It uses Stoke's Theorem and the definition of the magnetic field in terms of the vector potential, both things that are beyond the scope of this course, but it actually isn't terribly difficult. I link the wikipedia page so that interested students (or students in a more advanced course trying to connect back to simpler concepts by reading this book) can take a look.

complete the formal work we have begun by introducing the L_i and M_{ij} symbols. In terms of these, we can now write the total magnetic flux through the ith circuit loop due to the currents in *all* of the loops:

$$\phi_i = L_i I_i + \sum_{j \neq i} M_{ij} I_j \tag{8.59}$$

If we then differentiate this with respect to time and use Faraday's Law, we get the following expression for the induced voltage in the ith loop:

$$V_i = -L_i \frac{dI_i}{dt} + \sum_{j \neq i} M_{ij} \frac{dI_j}{dt} \tag{8.60}$$

Finally, in many, if not most, cases of interest, we can neglect mutual inductance because the magnetic field dies off rapidly with distance. For that reason we will often speak of the self-inductance *only* of specific circuit elements, especially "inductors", the magnetic equivalent of capacitors in a circuit, labelled with a plain L with or without an index. The key equation for a single self-inductance will be:

$$V_L = -L \frac{dI}{dt} \tag{8.61}$$

where V_L is the voltage drop or rise across the inductor and I is the current through the inductor. This expression finally gives us a good way of specifying the SI units for inductance. One Henry is a Volt-Second/Ampere, or a Volt-Second^2/Coulomb, or (since a Volt is a Joule/Coulomb) a Joule/Ampere^2.

Henries can, of course, also be expressed in terms of Webers – you *do* remembers what Webers are, don't you? It should be fairly obvious that 1 Henry is 1 Weber/Second, but nobody cares much about Webers, while everybody cares about Henries.

Example 8.6.1: The Mutual Inductance of a Wire and Rectangular Current Loop

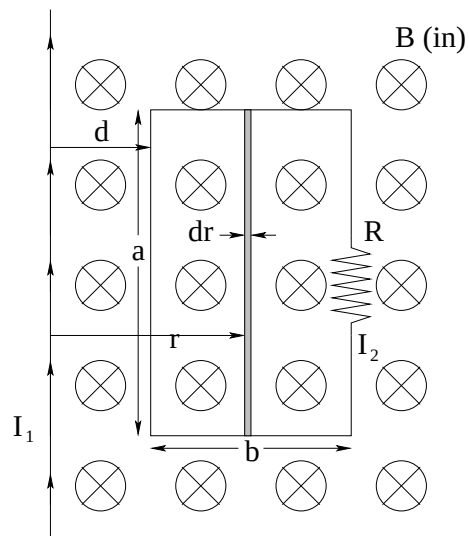


Figure 8.13: A long straight wire carrying a time-varying current $I_i(t)$ near a rectangular current loop induces a voltage V_2 in that loop, which in turn creates a current I_2 in that wire and a force $\vec{F}_2$ on the wire loop.

In figure 8.13 we return to the long straight wire and adjacent rectangular loop of wire (all in a common plane) that we examined above in the limited context of Lenz's Law and the direction of the induced current. This time, we want to answer *all* of the questions we might ask, such as:

- What is the magnetic field due to I_1 (Ampere's Law, of course).
- Given this field, what is the *magnetic flux* through the rectangular loop ϕ_2?
- Given this flux, what is the *mutual inductance* M_{21}?
- Given this flux and a current $I_1(t)$ that is *increasing*, what is the voltage V_2 induced in the rectangular loop?
- Given this voltage, what is the current $I_2(t)$ in the rectangular loop (magnitude and "direction", that is clockwise or counterclockwise in the arrangement shown?
- Finally, given this current, what is the net *force* on the loop, and is it attractive (back towards the long straight wire) or repulsive?

That's a lot of questions, but I laid it out in this way so you can see the very simple flow of reason. In a *quiz or exam* problem I'd be much more likely to just give the picture (without any "dressing") and say $I_1(t) = \frac{I_0}{T}t$, what is $\vec{\boldsymbol{F}}_2(t)$? So practice thinking about how this chain works so that each answer is a trivial step away from the previous one, but put together the answer isn't "simple" at all!

At this point you should really all be able to answer each and every step on your own, so I'll provide the most cursory review of each step and let you fill in the details (completely, of course!) for homework.

- From Ampere's Law (show!):
$$B_1 = \frac{\mu_0 I_1}{2\pi r} \tag{8.62}$$
into the paper on the side of the loop.
- To find the flux through the obvious plane surface S bounded by the rectangle, we have to start by finding the flux in the differentially thin strip shaded in the figure. The magnetic field is known and approximately constant in the strip in the limit that it is differentially thin. Thus:
$$d\phi_2 = \left(\frac{\mu_0 I_1}{2\pi r}\right) a dr \tag{8.63}$$
and
$$\begin{aligned} \phi_2 &= \left(\frac{\mu_0 I_1 a}{2\pi}\right) \int_d^{d+b} \frac{dr}{r} \\ &= \left(\frac{\mu_0 I_1 a}{2\pi}\right) \ln\left(\frac{d+b}{d}\right) \end{aligned} \tag{8.64}$$
- We can find the mutual inductance by dividing the flux by I_1:
$$M_{21} = \frac{\phi_2}{I_1} = \left(\frac{\mu_0 a}{2\pi}\right) \ln\left(\frac{d+b}{d}\right) \tag{8.65}$$

(This doesn't really help us find the force, but it is certainly something you should be able to do.)

- From Faraday's Law (show!)

$$V_2 = -\frac{d\phi_2}{dt} = -\left(\frac{\mu_0 a}{2\pi}\right)\ln\left(\frac{d+b}{d}\right)\frac{dI_1}{dt} \tag{8.66}$$

  and since I_1 is *increasing*, we expect the voltage to decrease (and drive a current) *counterclockwise* from Lenz's Law (see above).

- From Kirchoff's Rule and Ohm's Law (show!)

$$V_2 - I_2 R = 0 \tag{8.67}$$

  or

$$I_2(t) = \left(\frac{\mu_0 a}{2\pi R}\right)\ln\left(\frac{d+b}{d}\right)\frac{dI_1}{dt} \tag{8.68}$$

  (counterclockwise for $\frac{dI_1}{dt} > 0$).

- Finally, the force on each wire is – naaaah, I'm too lazy to help you out any more. Besides, I think you already found it in a previous homework assignment. The force on the side wires is a bit tricky, mind you, but not *that* tricky and the final answer is now very simple to obtain. What direction does the net force have to point even *before* you work it out?

As noted, this is pretty much your first homework problem, given down below. While it is OK to skim this part of the chapter before starting it, once you start it *do not look back* at this example; try very hard to work through the reason on your own. This means, of course (if you are reading these words right before you start the homework, maybe you'd better skim through this example *aqain* before you start...

There are a few other examples of "simple" geometries where one can compute the mutual inductance, and you will do at least one other one on your homework. The place where mutual inductance is a *critical feature*, the whole *point* instead of an annoyance is in the design and construction of transformers and inductively coupled rectifiers and the like. There are some places where one can make very clever use of mutual induction to accomplish some astounding things, such as in a *Tesla Coil*[113]

8.7: Self-Induction

Now we get to one of the most important parts of this chapter: computing the self-inductance of various simple current loops. We will have even fewer cases of geometries (and idealizations!) where we can even *think* of doing the integrals in a course at this level, and I will pretty much present all of them here. Interested students can, and should, visit wikipedia here: Wikipedia:

[113]Wikipedia: http://www.wikipedia.org/wiki/Tesla Coil. A Tesla Coil is basically a big resonant transformer that makes Big Sparks. In fact, it pretty much makes *lightning*. As such, it is a great favorite for students to make for an extra-credit project, because taming the lightning is what physics is all about, isn't it...?

http://www.wikipedia.org/wiki/Inductance both to read more about inductance itself and to see its lovely table of the self-inductance of a number of circuit shapes with *less* idealization. Nevertheless, our idealized answers herein will be more than sufficient to help us fully understand both the essential concepts and the general algebra required to do a better job.

Our general solution strategy here will be:

a) Find the magnetic field produced by the current I in the loop in question. Usually we will use Ampere's Law for this simply because integrating the Biot-Savart Law for arbitrary points in space is usually too difficult.

b) Write an expression for the flux produced by that field through the loop(s) that produce(s) it. This may be a simple product of field times area (for constant field perpendicular to the surface bounded by the loop) or an integral not unlike the one we did for rectangular loops near a long straight wire.

c) In cases where there are many "turns" (loops of wire) contributing to the overall flux, multiply by N, the number of turns.

d) Divide out the current. Voila! The self-inductance L!

Let's start with the simplest and most important example, the moral equivalent of the parallel plate capacitor for magnetic fields. The Self-Inductance of the (ideal) Solenoid:

Example 8.7.1: The Self-Inductance of the Solenoid

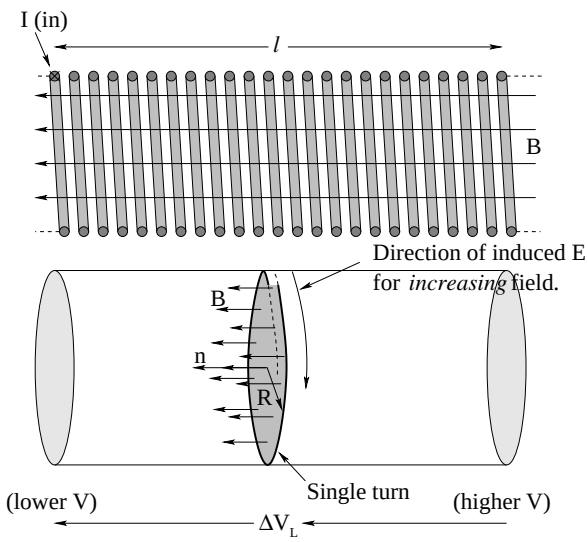


Figure 8.14: An ordinary (ideal) solenoid with N turns each carrying a current $I(t)$ is drawn above. The total flux through the solenoid is N times the flux through a single turn.

In figure 8.14 I've drawn an "ideal" circular cross-section solenoid, one with N (tightly wound) turns, a radius R, and a length $\ell \gg R$. Obviously I've had to exaggerate some of these features in the drawing – the radius of the wire itself is really very small compared to the other length dimensions, there is very little space between turns, and it should be longer compared to its illustrative radius.

Following the rubric given above, we first find the field inside of the solenoid using Ampere's Law (see week 7 if you cannot remember the correct Amperian path to use as the curve C):

$$\begin{aligned}\oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{l}} &= \mu_0 I_{\text{thru C}} \\ Bb &= \mu_0 \frac{N}{\ell} Ib \\ B &= \mu_0 \frac{N}{\ell} I \end{aligned} \tag{8.69}$$

where the direction is determined from the right hand rule, in the figure above to the left through the solenoid.

This field is *uniform* within an *infinite* solenoid and *vanishes outside of it* and we will idealize it as being uniform in this one and vanishing very rapidly at the ends (neglecting "fringing fields" outside of the volume of the solenoid, basically, much as we did for electric fields outside of the volume of an indealized parallel plate capacitor). This idealization will be valid as long as $\ell \gg R$ and the solenoid is tightly wound as noted.

Next, we find the self-induced flux through a *single* turn of the solenoid. Again we idealize the turn as being a circle in a plane instead of a segment of a helix, with area πR^2, so that:

$$\begin{aligned}\phi_{\text{turn}} &= \int_S \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA \\ &= B\pi R^2 \\ &= \mu_0 \frac{N}{\ell} I \pi R^2 \end{aligned} \tag{8.70}$$

The solenoid has N turns, *each* with this flux. Yes, they all count, as *each* of them contributes a piece ΔV_{turn} to the total potential difference as the current changes, so the total will be N times that of just one turn:

$$\phi_{\text{total}} = \left(\frac{\mu_0 N^2 \pi R^2}{\ell} \right) I \tag{8.71}$$

Finally, we find the self-inductance by noting that $\phi_{\text{total}} = LI$ so that:

$$L = \frac{\mu_0 N^2 \pi R^2}{\ell} \tag{8.72}$$

Note that we generally make L positive by convention and figure out any signs using Lenz's Law and a bit of common sense, so inductors don't come with a polarity or sign.

Nothing to it! Now suppose that $I(t) = I_0 \sin(\omega t)$ (a reasonable assumption for harmonic alternative voltages such as those we will shortly study). We can easily find:

$$\Delta V_L = -L \frac{dI}{dt} = I_0 (\omega L) \cos(\omega t) \tag{8.73}$$

where the field of the induced voltage *opposes* the increasing current during that part of its harmonic oscillation and *reinforces* the decreasing current during that part of its oscillation. As we indicate on the figure, if I, directed into the page at the top of the coils and out at the bottom, is increasing, then the induced E-field points out of the page at the top and in at the

bottom and the induced potential decreases right-to-left, opposing the increasing left-to-right current.

This may be tricky for you to see! The direction of the potential difference ultimately depends on *which way the coil was wound* – if the helix spirals from left to right (in at the top) as drawn then the net current transport is left to right and the induced voltage from an increasing current decreases from right to left. If it is wound right to left (in at the top) so that the net current transport is right to left as well, then the induced voltage for an increasing current will be left to right. It all makes perfect sense in terms of Lenz's Law either way – the voltage decreases in the direction that *opposes* the flow of the increasing current either way, and reverses to *support* it if and when the current decreases instead.

Before we move on, it is indeed worth pointing out that ωL in the expression for ΔV_L above has units of *resistance* (since $I_0 \omega L$ has units of volts). Next week we will name ωL *inductive reactance* as it will be a very important quantity in AC circuits.

Example 8.7.2: Toroidal Solenoid

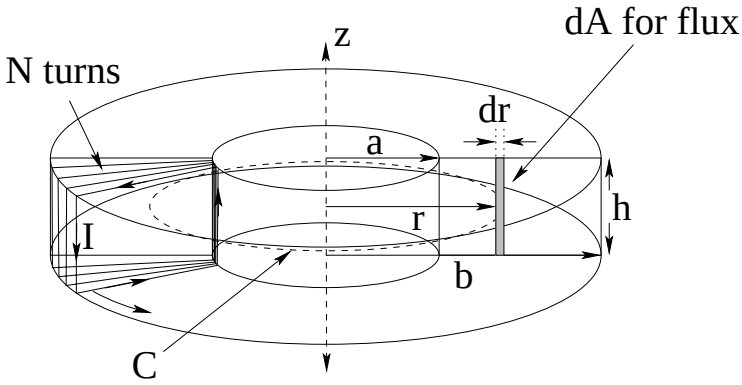


Figure 8.15: A tightly-wrapped toroidal solenoid with N turns produces a magnetic field inside that varies with r, but is approximately constant everywhere in a narrow strip of height h and width dr. The field is, of course, in the direction determined by the right hand rule, meaning that it points *in* to the page through the shaded strip we need to use to find the flux.

In figure 8.15 we see the same toroidal solenoid that we saw in week 7, where we evaluated the magnetic field inside using Ampere's Law. We will follow exactly the same rubric as before, except that this time I won't actually do the steps for you; they are part of this week's homework. Remember:

a) Evaluate the field (magnitude) $B(r)$ using Ampere's Law. Only refer back to week 7 if you must, as by now you *should* be able to do this on your own without looking!

b) Evaluate the flux through a single turn of the toroidal solenoid. This will involve setting up an integral that is almost *exactly* the same as the integral in the example of finding the mutual inductance of a long straight wire and a rectangular loop above. Again, try *not* to have to go back and look, as the picture should remind you of what you need to do, and the integral itself is pretty trivial.

c) Multiply the flux for a single turn by N, the number of turns in the solenoid (as once again *each* turn contributes to the overall potential difference) to find the total flux.

d) Divide the flux by the current to find the self-inductance of the solenoid.

e) Think a minute. Suppose the current $I(t)$ in the direction shown in the figure is *increasing*. What is the direction of the induced electric field around a loop? Suppose it is decreasing, ditto? Either way, of course, the induced voltage across the two wires leading to/from the solenoid will *oppose* the change in the current!

f) If desired, find e.g. the voltage $V_L = -L\frac{dI}{dt}$ or any other quantities of interest.

Example 8.7.3: Coaxial Cable

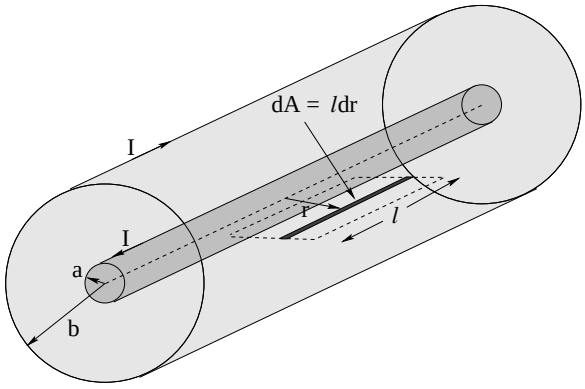


Figure 8.16: Coaxial cables have a self-inductance measured *per unit length*. At high frequencies the inductance only depends on the outer radius of the inner conductor a and the inner radius of the outer conductor b. A strip of area $dA = \ell dr$ is shown that may be of use in computing L/ℓ, the self-inductance per unit length.

This sets up another homework problem, as I'm feeling even lazier than before and *you* need to do the work in order to learn how! In figure 8.16 a current $I(t)$ flows e.g. up the (long, straight cylindrical shell) inner conductor and back on the outer (long, straight cylindrical shell) outer conductor. From Ampere's Law you can easily find the magnetic field where it is confined in between the inner and outer conducting shells.

With the magnetic field in hand, it should be easy to find the flux through the dark shaded strip shown (with the parameter ℓ in it, so this will yield the *flux per unit length* once the ℓ is divided out) and integrate from a to b, an integral that should by now be boringly familiar to you. Divide by the current *and* ℓ to find the *self-inductance per unit length of the cable*.

That isn't quite *all* of the cases where one can compute the self-inductance of something without needing to do absurdly difficult integrals or deal with even more heavily approximated fields – for example, you might think about what the self-inductance per unit length is for a thick cylindrical *wire* of radius R and resistivity ρ_r – but it is pretty close.

8.8: LR Circuits

From here on out, with rare exceptions we will work with *inductors* as (self-inductive) circuit elements just like capacitors and resistors. We will use "The Solenoid" (idealized) as our architypical inductor, and we will often pretend that they are made with superconducting wire (as a further idealization) so that they have no resistance to worry about. Real inductors, of course, are made with many turns of relatively thin wire and can have substantial (non-negligible) resistance as well as self-inductance. However, their "resistive" properties can always be considered to be a resistor in series with a pure zero resistance inductor, so nothing is lost by the idealization as long as we remember to include their resistance in our circuits.

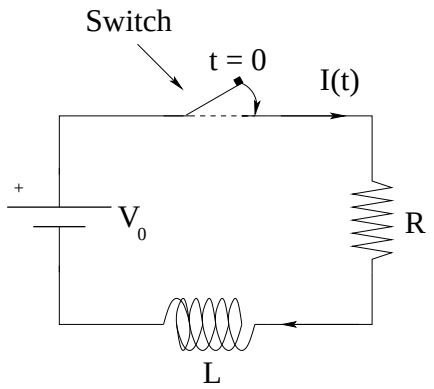


Figure 8.17: The architypical direct current LR circuit. We generally assume the switch closes at time $t = 0$ with the current in the circuit $I(0) = 0$.

Let us, then, figure out a simple DC LR circuit, given in figure 8.17: an inductor in series with a resistance R, which *could* be the natural resistance of the inductor itself, or an external resistor, or the combined resistance of an external resistor and the resistance of the inductor. Note well that we have generated a *symbol* for an inductor in an electrical circuit, the squiggly thing that looks like a coil/solenoid with many turns of wire. We don't care much about how many turns it has, or how long it is, or what its cross-sectional area is, or whether or not it contains a magnetic material (discussed later). All we care about is the *combined* effect of all of this, the (self) inductance L (and possibly its contribution to the total resistance R of any branch of a circuit it is in).

Obviously no current flows while the switch is open. We imagine closing the switch at time $t = 0$. The battery will drive current through the wire. The resistor will oppose this current (Ohm's Law), and the inductor will *also* oppose this current as long as it is *increasing* (Faraday's Law). At some finite time t later, we expect to find some non-zero current in the circuit, one that is changing in time, and will use this assumption in analyzing the circuit algebraically.

First, however, let's see what we can figure out using nothing but *verbal reason* and *dimensional analysis* instead of algebra and calculus. We begin, as we see, at $I(0) = 0$. After a *very* long time, we rather expect that the current will arrive at some constant value, at which point the back-voltage generated by the inductor will be zero The voltage gain from the battery will all drop across the resistor, suggesting that the current will be $I_\infty = V_0/R$. We therefore expect a current $I(t)$ that starts at zero and approaches V_0/R *before beginning the problem*, and we might guess that it will approach this current exponentially. All that is left is guessing

the exponential time constant.

Well, we have two parameters to play with: R and L. Ohms are Volts/Ampere. Henries are Volt-Seconds/Ampere. We want a time constant in seconds, so it looks like:

$$\tau = \frac{L}{R} \tag{8.74}$$

will have units of seconds and is the simplest way of getting such a time out of the three quantities that *could* appear in the answer, V_0, L and R. If our life depended on just writing down an expression for $I(t)$ that is at least approximately correct, we would then guess:

$$I(t) = \frac{V_0}{R}\left(1 - e^{-\frac{t}{\tau}}\right) = \frac{V_0}{R}\left(1 - e^{-\frac{R}{L}t}\right) \tag{8.75}$$

before starting the problem!

Although perhaps it will be a bit anticlimactic, let's solve it the more difficult but formally correct way. We start, as usual, with Kirchoff's Loop Rule, some arbitrary time after the switch is closed:

$$V_0 - IR - L\frac{dI}{dt} = 0 \tag{8.76}$$

We rearrange this to put it in the standard form of a first order, linear, inhomogeneous ordinary differential equation:

$$\frac{dI}{dt} + \frac{R}{L}I = \frac{V_0}{L} \tag{8.77}$$

At this point I shouldn't have to help you. We've now solved this equation several times over two semesters[114] – it is directly integrable after some rearrangement and is clearly an important equation to be able to effortlessly solve if you want to understand Nature, not only in the context of physics but in biology and chemistry and medicine as well. If you remember how, stop reading here, get out a piece of paper, and do so, verifying that you get the solution I already deduced above without using algebra or calculus. Work neatly, as this is a straight up homework problem so your efforts won't be wasted.

But what the heck, you're learning, you've forgotten, so I'll solve it here again. But *pay attention* this time – really *learn to recognize this kind of equation and solve it when you see it!* Practice it a bit, then *wait a day* and try working through this section again, this time solving the FOLIODE above without looking.

[114]Approach to terminal velocity with a linear drag force, approach to a terminal velocity for a rod on rails with a battery or gravity, charging a capacitor in a DC RC circuit, for example.

So here we go:

$$\begin{aligned}
\frac{dI}{dt} + \frac{R}{L}I &= \frac{V_0}{L} \\
\frac{dI}{dt} &= \frac{V_0}{L} - \frac{R}{L}I \\
\frac{dI}{dt} &= -\frac{R}{L}\left(I - \frac{V_0}{R}\right) \\
\frac{dI}{(I - \frac{V_0}{R})} &= -\frac{R}{L}dt \\
\int \frac{dI}{(I - \frac{V_0}{R})} &= \int \left(-\frac{R}{L}\right) dt \\
\ln\left(I - \frac{V_0}{R}\right) &= -\left(\frac{R}{L}\right)t + C \\
\exp\left\{\ln(\left(I - \frac{V_0}{R}\right)\right\} &= \exp\left\{-\left(\frac{R}{L}\right)t + C\right\} \\
I - \frac{V_0}{R} &= e^{-(\frac{R}{L})t}e^{C} \\
I &= \frac{V_0}{R} + Ae^{-(\frac{R}{L})t} \\
I(t) &= \frac{V_0}{R}\left(1 - e^{-(\frac{R}{L})t}\right) \qquad (8.78)
\end{aligned}$$

where we've used the fact that the natural log and exponential are inverse functions of one another and where we set the (exponential of) the constant of integration from the indefinite integrals A to $-V_0/R$ in order that $I(0) = 0$ (the initial condition, recall).

8.8.1: Power

Let's track the flow of energy in this circuit. Remember, the power delivered to/used by any given circuit element is $P = VI$ where V is the voltage gain/drop across the element and I is the current through it (which we now know).

The power provided by the battery (positive):

$$P_V = V_0 I(t) = \frac{V_0^2}{R}\left(1 - e^{-(\frac{R}{L})t}\right) \qquad (8.79)$$

Wow, that was easy!

The power burned in the resistor (negative – remember, this is energy that is all turned into (joule) heat(ing):

$$\begin{aligned}
P_R &= V_R I(t) = (-I(t)R)I(t) = -I(t)^2 R \\
&= -\frac{V_0^2}{R}\left(1 - e^{-(\frac{R}{L})t}\right)^2 \\
&= -\frac{V_0^2}{R}\left(1 - 2e^{-(\frac{R}{L})t} + e^{-(2\frac{R}{L})t}\right) \qquad (8.80)
\end{aligned}$$

which is a bit more complicated, but still not terrible. Note that I stuck a minus sign in front because this is power being *removed* from the system by the voltage *drop* across the resistor. With this sign choice, we are guaranteed to have energy conserved, as we will see below.

The power delivered to the inductor (negative, but where does this energy go? See the next topic...):

$$\begin{aligned} P_L &= V_L I(t) = (-L\frac{dI}{dt})I(t) \\ &= -\left\{L\frac{V_0}{R}\left(\frac{R}{L}\right)e^{-(\frac{R}{L})t}\right\}\frac{V_0}{R}\left(1-e^{-(\frac{R}{L})t}\right) \\ &= -\frac{V_0^2}{R}\left(e^{-(\frac{R}{L})t} - e^{-2(\frac{R}{L})t}\right) \end{aligned} \tag{8.81}$$

Note that we used the fact that

$$V_L(t) = -L\frac{dI}{dt} = -V_0 e^{-(\frac{R}{L})t} \tag{8.82}$$

is the voltage drop across the inductor just as:

$$V_R(t) = -IR = -V_0\left(1-e^{-(\frac{R}{L})t}\right) \tag{8.83}$$

is the voltage drop across the resistor.

You can easily verify that these three add up to zero, so energy is conserved, but of course how could it *not* be conserved? Take Kirchoff's rule for this circuit above and multiply it by $I(t)$:

$$\begin{aligned} V_0 - IR - L\frac{dI}{dt} &= 0 \\ (V_0 - IR - L\frac{dI}{dt})I(t) &= 0 \\ V_0 I(t) - I(t)^2 R - L\frac{dI}{dt}I(t) &= 0 \\ P_V + P_R + P_L &= 0 \end{aligned} \tag{8.84}$$

(where the signs all hopefully make sense to you). The *whole point* of Kirchoff's Loop Rule is that it guarantees energy conservation around circuit loops, so we shouldn't really be surprised when it works, but it is useful to *show how* it works in an actual context from time to time to reinforce the idea.

But *is* all of that power being delivered to the inductor going? It isn't being burned and released as heat – that part of the tally is accounted for in the resistance! Maybe – could it be – is it possible – that the energy is going into the *magnetic field*?

It is.

8.9: Magnetic Energy

Let's imagine that the power delivered to the inductor is is somehow being *stored* in the inductor in the magnetic field. Then:

$$P_L = \frac{dU_L}{dt} = -LI\frac{dI}{dt} \tag{8.85}$$

or (multiplying by dt):

$$\begin{aligned} dU_L &= -LI\,dI \\ \int_0^{U_{\rm tot}} dU_L &= \int_0^{I_0} -LI\,dI \\ U_{\rm tot} &= \frac{1}{2}LI_0^2 \end{aligned} \tag{8.86}$$

This is the moral equivalent of the $U = \frac{1}{2}CV^2$ that we similarly derived for a capacitor, but this is a *dynamic* quantity as it depends on the current *flowing* in the inductor.

Let is imagine that our inductor is an ideal solenoid with N turns, length ℓ, and cross-sectional area A, one where the magnetic field inside the solenoid is constant and equal in magnitude to:

$$B = \frac{\mu_0 N I_0}{\ell} \tag{8.87}$$

and that vanishes at the ends of the solenoid (neglecting fringing fields). We showed above that the self-inductance of this ideal solenoid is:

$$L = \frac{\mu_0 N^2 A}{\ell} \tag{8.88}$$

Let's do an algebra-morph of the energy stored on the inductor:

$$\begin{aligned} U &= \frac{1}{2}LI^2 \\ &= \frac{\mu_0 N^2 A}{2\ell} I^2 \\ &= \frac{\mu_0^2 N^2 A\ell}{2\ell^2\mu_0} I^2 \\ &= \frac{1}{2\mu_0}\frac{\mu_0^2 N^2 I^2}{\ell^2} A\ell \\ \Delta U &= \frac{B^2}{2\mu_0}\Delta\mathcal{V} \\ \frac{\Delta U}{\Delta\mathcal{V}} &= \frac{B^2}{2\mu_0} \end{aligned} \tag{8.89}$$

where we have used the fact that $A\ell = \Delta\mathcal{V}$, the *volume* of the solenoid (the only region where our idealized field is not zero).

Note that I stuck delta's in so that I could relate the amount of energy per amount of volume or *energy density in the magnetic field* to help us make the *ansatz*[115]:

$$\eta_m = \frac{dU_m}{d\mathcal{V}} = \frac{B^2}{2\mu_0} \tag{8.90}$$

which strangely matches our similar equation (deduced from very similar considerations for the energy density in the *electric* field:

$$\eta_e = \frac{dU_e}{d\mathcal{V}} = \frac{1}{2}\epsilon_0 E^2 \tag{8.91}$$

There is something really sort of spooky about this – it is redolent[116] of as-yet undiscovered relationships between the electric and magnetic fields. Soon, my child, soon we will understand this and a great burst of *illumination* will occur. Literally.

As was the case for capacitors, it isn't enough to just make the *ansatz*. We need to verify that it works for at least one other geometry of inductor, ideally one with a varying field and inductance we can compute. Our only real choice here is the toroidal solenoid.

[115] Physicsspeak for "inspired guess"...

[116] Politespeak for "it stinks"...

Example 8.9.1: Energy in a Toroidal Solenoid

Suppose you have the very toroidal solenoid we study above, carrying a current I. We can use Ampere's Law to find the magnetic field strength $B(r)$ inside the solenoid, of course. We can then use it to find:

$$\frac{dU}{d\mathcal{V}} = \frac{B(r)^2}{2\mu_0} \tag{8.92}$$

if we multiply this out:

$$dU = \frac{B(r)^2}{2\mu_0} d\mathcal{V} \tag{8.93}$$

and *integrate both sides*, we should get U_m, the total energy stored in the magnetic field (according to our ansatz).

Show that this is exactly equal to:

$$U = \frac{1}{2} L I^2 \tag{8.94}$$

using the L you found above.

Note that I'm not actually doing this for you, but I will help you one teensy bit. the volume element $d\mathcal{V}$ you should use is the one of thickness dr at radius r with height h, or

$$d\mathcal{V} = 2\pi r h dr \tag{8.95}$$

Give it a shot, for homework. You can do it!

8.10: Eddy Currents

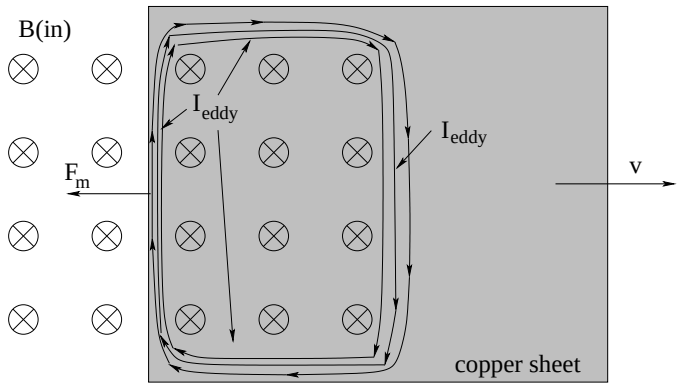


Figure 8.18: A sheet of copper being pulled rapidly out of a field has induced *eddy currents*. The forces from these currents, according to Lenz's Law, *resist* the motion, causing a magnetic "drag force" similar to that observed in the rod on rails problem. The kinetic energy of the object is transformed into heat by these currents (resistive Joule heating).

We have seen up above that a current loop resists being pulled from *or* pushed into a magnetic field because the field induces currents that exert forces that act against any change in flux. Just as this is true for actual e.g. loops of wire, it is also true for *bulk conductors!* Any conducting material such as a sheet of copper will resist being pushed into or pulled out

of a magnetic field, because the changing field causes currents to loop through the entire conductor as if it were many, many parallel wires. We call these currents "eddy currents".

Eddy currents are remarkably important, as they are a source of *energy loss* whenever we attempt to e.g. alter a magnetic field in the vicinity of *any conductor*. Eddy currents produce *Joule heating* of the conducting material very readily – one can actually cook food on stoves that use a rapidly varying magnetic field to directly heat metal pots placed in the field[117] . Transformers (covered later) rely on rapidly varying, ferromagnetically enhanced magnetic fields to step up or step down voltage, and unless care is taken to prevent eddy currents in the design of the magnetic cores, much of the energy being transmitted through the transformer will be lost to heating the cores. Eddy currents cancel electromagnetic radiation at the surfaces of conductors, both heating the conductors slightly and causing the electromagnetic field to *reflect* from the surface rather than be transmitted. It seems worthwhile to spend a moment trying to understand them.

In figure 8.18 above, a sheet of copper being pulled rapidly out of a strong magnetic field is illustrated. It is moving at some speed v to the right. As it is pulled out, the magnetic flux through the *entire sheet* is reduced. This creates an induced field in the conductor and its associated induced voltage that (because it is a *good conductor*) can and does drive a large current in the copper. This current is not isolated or confined in the conductor – the conducting sheet is like an entire field of parallel resistance pathways and the current spreads out to use them.

Note well, however, that *like the rod on rails problem* (which this greatly resembles!) the net *force* on the induced current is in a direction that *opposes* v (whichever direction the sheet is moving, in or out of the field). The current flow *in* the field produces this force , while the current flowing in the opposite direction through the part of the sheet that is out of the field does not. One expects that the velocity of this sheet, like the velocity of the rod, will be exponentially damped, or, if the sheet is being pulled, will reach a terminal velocity.

The current itself is like a "whirlpool" or eddy of charge swirling around in the material, hence the name eddy current. There are several simple demonstrations of eddy currents – swinging a sheet of copper down between the poles of a powerful magnet with or without slits that break up the conductive pathways and reduce the effect, swinging a magnet above a conducting sheet, or (my favorite) dropping a powerful magnet down through a copper pipe and a PVC pipe at the same time.

Magnetic brakes can use this same principle to stop a car, although (as a homework problem will demonstrate) one can avoid wasting the energy by turning the wheel rotors into "generators" that can store the energy in a battery as they remove it.

We will return to the notion of eddy currents when we treat transformers because the iron cores of transformers are usually *laminated* – made of thin sheets or wires of iron coated with and separated by an insulating resin – precisely to prevent eddy currents from the rapidly changing magnetic fields they help support from heating the iron and hence wasting the *energy*

[117] Wikipedia: http://www.wikipedia.org/wiki/Induction Cooking. This is actually a lovely article, and will introduce you via a link to the notion of *skin depth*, as induction stovetops only tend to work on ferromagnetic pans (such as cast iron) because they have a high magnetic permeability (discussed shortly), a small skin depth, and hence concentrate the induced current in a thin layer of the iron with a much higher electrical resistance than is obtained with an otherwise identical copper or aluminum pot.

in the time varying magnetic field.

8.11: Magnetic Materials

We have postponed discussing the magnetic properties of materials until here because we had to wait until we understood the basic idea of Faraday's and Lenz's Laws. As we will see, the *diamagnetic* property of some materials that corresponds to the *dielectric* properties we've already studied comes about as a result of Faraday's Law.

However, another good reason to wait until now is that *magnetic properties of materials are much more complicated* than electrical properties were. Back in electrostatics, dielectric polarization was about it. Well, not really – a very *few* materials exhibit e.g. ferroelectric properties, and further study also reveals that dielectric polarization and electrical conductivity are two aspects of a single complex quantity and not really independent – but close enough. If you put nearly any material in a static or slowly varying electrical field, the field inside that material will be *reduced*.

If you put that *same* material in a static or slowly varying electrical field, you might find:

- The magnetic field inside is *reduced*. We call this **diamagnetism**.
- The magnetic field inside is *increased*. We call this **paramagnetism**.
- The magnetic field is altered by the addition of another vector magnetic field produced by the material itself, a field that persists even if there is *no* external field. We call this **ferromagnetism**.

These are all bulk descriptions, and fail to capture the wide variety of magnetic structure one can discover on the microscopic scale of the material. They also are all properties that depend on the *temperature* of the material. In fact, a single material can, at different temperatures, be ferromagnetic, paramagnetic, and diamagnetic!

Thus far, we have been pretty successful in understanding things classically, but certain aspects of the magnetic properties of matter rely heavily on quantum mechanics, in particular the fact that electrons have *spin* (and hence an intrinsic magnetic dipole moment) and *orbit the atomic nucleus in non-radiating, non-resistive orbits*. We will have to draw at least on these "cartoon" ideas as we seek to grasp the general concepts and ideas underlying magnetic behavior of materials.

Diamagnetism

This is a course on classical physics, but magnetism in particular is very difficult to understand on purely classical grounds. For example, we've seen above how conductors will at least transiently *reduce* magnetic fields that attempt to penetrate them, as eddy currents are induced around their perimeter. We can imagine that a superconductor with *zero* resistance would reduce those fields to zero (and indeed that is the oversimplified case, with some limitations) but superconductivity is a purely quantum phenonmenon.

We don't have to go to the extreme case of superconductivity to require a bit of quantum theory in our explanation, however. Basically all three of the primary ways ordinary matter modifies magnetic fields are at least partially quantum mechanical in their explanation.

Atoms can be thought of as more or less spherically symmetric balls of electrons surrounding heavy pointlike nuclei. The electrons are in "orbits" around these nuclei, but the orbits are not classical orbits like the Moon orbiting the Earth, they are non-radiating, zero resistance flows of electronic current around the nucleus.

If a magnetic field is increased in the vicinity of an atom, Faraday's Law suggests that all electronic currents around an axis parallel to the magnetic field through the nucleus will be increased or decreased as needed in order to *reduce* that field. This alteration in the currents can be accompanied by an increase or decrease in the average radius of the orbits in question, and by small changes in the energy of those orbits.

If the currents were *classical* currents moving against some form of resistance, the decrease in magnetic field strength due to the induced current would be small, transient and difficult to detect. However, quantum atomic orbitals have *no resistance*. As long as the external magnetic field isn't varied *too* rapidly by *too* great an amount, so that the atom has time to "smoothly" adjust its orbitals, the induced current variation doesn't involve dissipation and the field reduction dynamically tracks the applied field and is "permanent".

To see what happens inside a block of dense matter, we need to consider how all of these reactive currents combine. In figure 8.19 an external magnetic field into the page is applied to

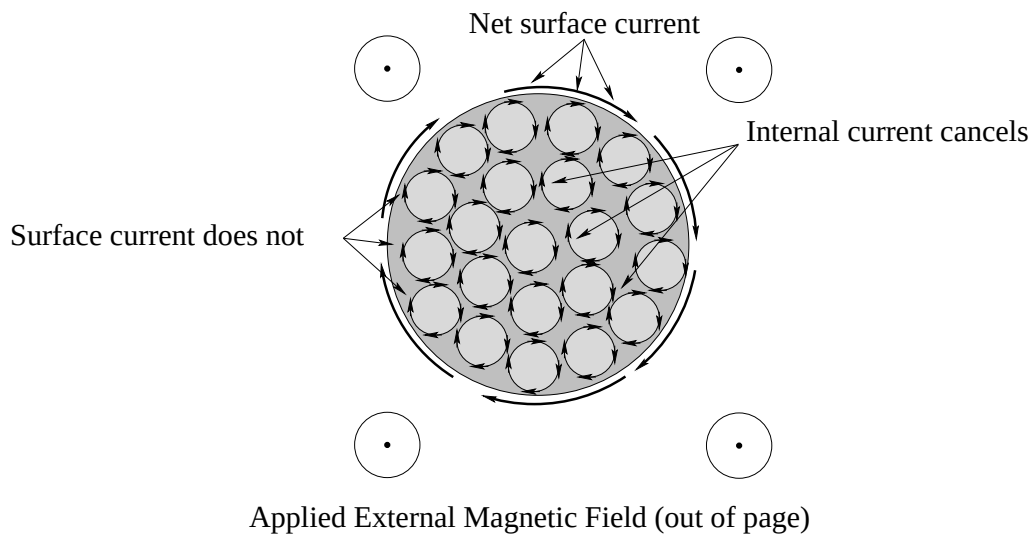


Figure 8.19: Wherever "atomic" magnetic current loops adjoin one another, the average current is zero. On the surface, however, there are no neighboring atoms, and the current loops there are not cancelled. They add (on average) into a *continuous surface current* not unlike that of a solenoid, so that the field everywhere in the interior is *reduced*.

a (highly magnified) block of material. This field induces non-dissipating atomic currents in the atoms that create magnetic dipoles pointing *into* the page.

Inside the bulk of the material, the current circulating around one atom approximately cancels the current circulating around the atoms next to it, where they are in contact. If one does a coarse grained average of the current, it is nearly zero in any small volume of the material

containing many atoms.

This is not true on the surface. The currents of the atoms on the surface have no neighboring atoms with currents running the opposite way on the outside, so there the currents all *combine*, on average, to produce a net current running around the perimeter of the object. This current is almost identical to that of a *solenoid*, and, like a solenoid, there is a uniform field inside the material that directly opposes the applied external field and hence reduces it inside of the material[118].

We will call this reactive response *diamagnetism*, the exact analog of the dielectric response of most insulators and conductors. Nearly all materials have a diamagnetic response to applied magnetic fields (especially at higher temperatures), but many materials have this response overridden by one or both of the following kinds of bulk magnetization, which have very different explanations.

8.11.1: Superconductors

Certain materials, when cooled to extremely low absolute temperatures, become *superconductors*. Superconductivity is a more or less purely quantum mechanical phenomenon and hence is beyond the scope of this book – basically a fraction of the electronic charge starts to behave collectively like a macroscopic quantum "orbital" that can transport electronic charge without resistance.

Superconductors can be thought of as being "diamagnetic" – indeed *perfectly* diamagnetic (as well as being perfectly dielectric) as they tolerate no magnetic or electric field inside at all, but it isn't exactly the same mechanism as merely opposing an applied field via induction; a superconductor actively *ejects* any existing magnetic field as it is cooled across the transition temperature where superconductivity appears, even if that field is not changing. One visible sign of this ejection is that superconductors placed above a permanent magnet *float*, suspended by its perfectly opposed magnetic field. This is called the Wikipedia: http://www.wikipedia.org/wiki/Meissner EffectMeissner Effect.

Superconductors, of course, are potentially very useful – a long term search continues for finding specially engineered materials that are superconducting at e.g. room temperature. A room temperature superconductor would have enormous positive implications for our civilization – levitating trains that require no energy to levitate, loss-free transmission of electrical energy over long distances, and much more – but so far they have eluded our search. As of the time of this writing, the highest temperature superconductors thus far found have critical temperatures in the range of 100-150 degrees Kelvin, over 100 degrees Kelvin short of even the freezing point of water.

Still, enormous progress has been made in recent decades. We can certainly at least hope that high(er) temperature superconductors eventually have a significant impact on our lives.

[118]This follows from Ampere's Law applied to e.g. paths parallel to the applied field on the inside of the material that contain a piece of the surface current, similar to the "infinite plane sheet of current" we considered earlier.

Paramagnetism

Some molecules have permanent electric dipole moments. *Many* atoms or molecules have permanent *magnetic* dipole moments. This is a purely quantum mechanical phenomenon. Charged electrons and protons have *spin* and hence *are* permanent magnetic dipoles. As atoms and nuclei are "built" out of many protons, neutrons, and electrons these spins are paired when possible in such a way that no net moment results, but all across the periodic table are elements with unpaired electrons or protons, and at least potential net spin and magnetic moment. This angular momentum combines with orbital angular momentum to produce many atoms with magnetic dipole moments[119] .

We know that magnetic dipoles have a potential energy in an applied magnetic field that is a *minimum* when the dipoles are aligned with the field. Although (as we have seen) magnetic dipoles associated with angular momentum on the scale of elementary particles or atoms experience a torque due to an applied magnetic field that causes their angular momentum to *precess around the magnetic field*, they *also* experience many small "random" torques due to thermal (heat) fluctuations in their environment. These torques caused by e.g. collisions between atoms or vibrations in a lattice constantly more or less randomly reorient the magnetic moments at high temperatures so that the system has no net average magnetic dipole moment. A lattice of "spins" at high temperature is pictured in figure 8.20.

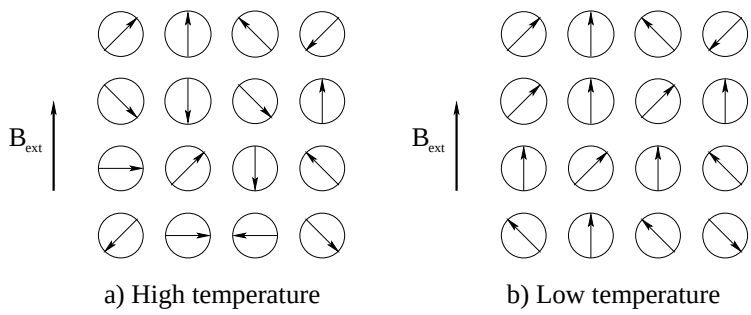


Figure 8.20: A lattice of "spins" at high temperature (a) and low temperature (b) is portrayed as a two dimensional cartoon. The direction of the arrows can be thought of as the directions of the angular momentum and hence magnetic moment of each atom, in a side view that reveals their rough degree of alignment with the field. At high temperature the spins are more or less randomly aligned with the field, but at low temperature there is less free energy and the spins are much more likely to be in a lower energy state, partially or completely aligned with the external field.

At low temperatures there is less (free) energy to share among all of the spins – recall that the *equipartition theorem* (for example) relates the total kinetic plus potential energy in all of the degrees of freedom of an atom to its temperature. It is therefore a lot more likely to find the atoms in states that have "less" magnetic potential energy in the field than those that have more, and atoms have the least magnetic potential energy when they are in alignment with the field! Consequently, at low enough temperatures we are likely to find the "permanent" magnetic moments of the atoms or molecules (if any) *aligned with the applied external field!*

[119]Wikipedia: http://www.wikipedia.org/wiki/Magnetic_moment#Magnetic_moment_of_an_atom. In fact there is a dizzying array of ways these moments can arise, too many to exhaustively and correctly cover here.

This alignment causes the *exact opposite response* of the material to the field. Since all of the magnetic moments are lined up *with* the field, and can be much larger than induced magnetic moments that oppose it that are being created *at the same time*, the net field produced by the "current loops" still cancels on the interior and adds up on the surface, but this time to *enhance* or *augment* the applied field. The total magnetic field inside the material is *larger* than the original external magnetic field. This is portrayed in figure 8.21.

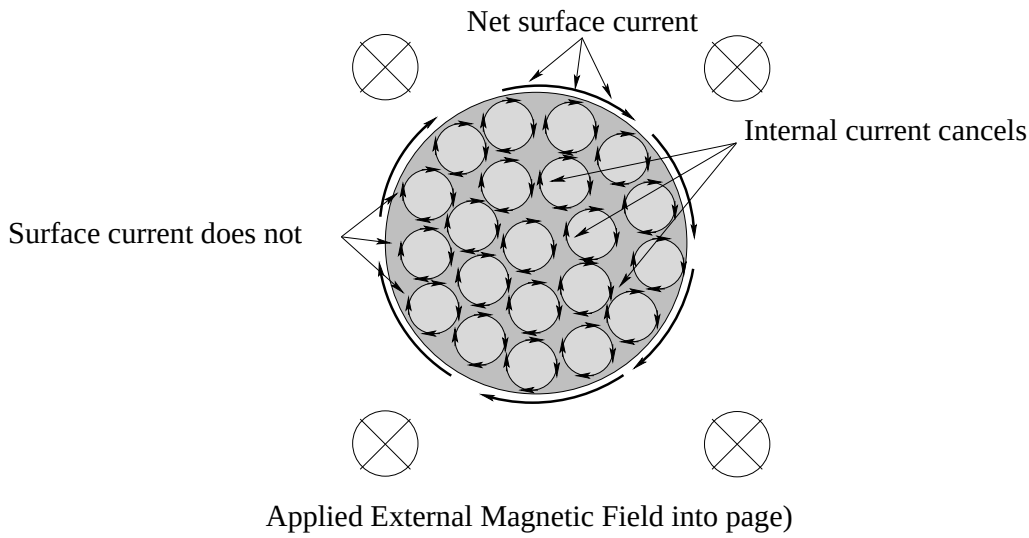


Figure 8.21: Just as was the case for a diamagnet, the internal currents of aligned magnetic moments cancel (on average) in the bulk of the material, but the surface currents *add*. The surface currents behave like the wires of a solenoid or sheet of current wrapped around the object to *increase* the total field inside.

This kind of response is called *paramagnetism*. A paramagnet increases the strength of the magnetic field inside. Since this (in turn) increases the magnetic *flux* through the material, putting a paramagnetic material inside a solenoid increases its self-inductance the same way a dielectric material increases the capacitance of a capacitor. Most solenoids in electronics use some sort of paramagnetic material (or ferromagnetic material, read on) to enhance the inductance of their inductors, getting the same inductance with fewer turns, material, and resistance.

Ferromagnetism and Antiferromagnetism

One can *barely* appreciate paramagnetism classically. Spinning electrons and orbits with both angular momentum and a magnetic moment are classically accessible, even though their properties (such as quantization of the angular momentum) are partly determined by quantum theory. Not so for the next two kinds of magnetic behavior of materials. They are purely quantum mechanical; one has the opposite sign altogether to anything you would expect classically.

Let us suppose that the permanent magnetic moments on two neighboring atoms can themselves interact. This alone isn't inconceivable – one creates a (weak) magnetic field at the location of the other, although the actual direction of that field is determined by the *relative* orientation of the source dipole and the target location and hence not easy to imagine. We will

further suppose that the interaction is bilinear in the magnetic moments themselves, and since energy is a scalar, we'll make the bilinear product the scalar product for simplicity.

That is, let us suppose that the potential energy of interaction between two neighboring atoms (labelled with i and j respectively) has the general form:

$$U_{ij} = -J_{ij}\vec{\boldsymbol{m}}_i \cdot \vec{\boldsymbol{m}}_j \tag{8.96}$$

where J_{ij} is the *energy coupling* between the two moments. Note well that this form is by no means unique or necessarily correct – it is more or less a hypothesis that we'd need to test against observed materials.

If $J_{ij} > 0$, the two moments will have minimum energy when they are *aligned* (ferromagnetism). If $J_{ij} < 0$, the two moments will have minimum energy when they point in *opposite directions* (antiferromagnetism). As before, when the temperature goes down, the energy removed has to come from somewhere, so low temperatures will favor a "paramagnetic" alignment or antialignment of the moments. The interesting thing is that this alignment will occur *even in the absence of an external field!*

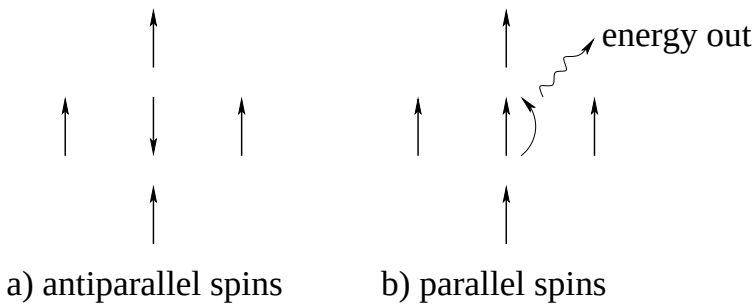


Figure 8.22: A cluster of five magnetic moments (spins) is illustrated with the central spins in two possible configurations. When the central spin is antiparallel to the four surrounding spins, it has potential energy $U_a = +4Jm^2$ in a suitable system of units. When it lines up parallel to the four surrounding spins, it's energy is $U_p = -4Jm^2$.

The energetics of this are illustrated in figure 8.22. This is yet another cartoon representation in two spatial dimensions, this time of "spins" in one dimension (each spin is associated with a magnetic dipole moment more or less as usual by a relation such as:

$$\vec{\boldsymbol{m}}_e = \frac{e}{2m_e}\vec{\boldsymbol{s}} \tag{8.97}$$

in a suitable system of quantized angular momentum units). In this kind of toy model, we only let the spins point in one of two directions: up or down, to study only their tendency to align or antialign at different temperatures. This is a "real" model of some importance in physics in the study of *magnetic phase transitions* between paramagnetic and ferromagnetic states (the latter with permanent magnetic dipole moments) and is called the *Two Dimensional Ising Model*[120] .

[120]Wikipedia: http://www.wikipedia.org/wiki/Ising Model. Note well the other links at the end of this article to an (as promised!) dizzying array of magnetic models and theories. Magnetism in matter is *interesting and important* and a simple Ising model computation/simulation is well within the reach of a student looking for a project who knows a programming language or how to use e.g. Matlab or Mathematica.

In this figure two spin configurations are presented – the first with four neighboring spins (all in the same direction) surrounding a spin that points in the opposite direction. The energy of the central *antiparallel* spin in this case is $U_a = +4Jm^2$. In the second, the central spin is parallel to the surrounding spins and the energy is now *negative*: $U_p = -4Jm^2$. The energy difference between these two configurations is hence $\Delta U = 8Jm^2$.

At high temperatures, both configurations are nearly equally probable in a given lattice of spins, with the parallel configuration only slightly favored, and the system would behave like a paramagnet or even a diamagnet if the diamagnetic response was larger than the paramagnetic alignment to an external field (this is controlled with a different coupling constant in the case of the Ising model between the spins and an external field).

As one cools the system, one removes heat energy from it. That energy comes from (among other places) the *magnetic potential energy of interaction between the spins*. Note well that this is a place where ***magnetic fields indeed do work!*** It doesn't violate our earlier theorem because ***the magnetic moment due to quantum spin is not generated by rotating bulk charged matter!*** Our "no work" theorem specified work done on *particles* (or coarse grained chunks of moving spinless charge) but there *is* no charge moving with a velocity $\vec{v}$ associated with the magnetic moment connected to the *intrinsic* angular momentum of an elementary particle.

In very rough terms, here is the thermodynamics of it. As soon as the internal "thermal" energy that scales like k_BT (where k_B is Boltzmann's constant) is smaller than the energy difference between parallel and antiparallel configurations, the parallel configuration starts being much more likely to be found in the lattice and the spins in the lattice start to "order" in small clumps of locally parallel spins that grow (and compete) as the system further cools. At a *critical temperature*, the size of one of the clumps spans the lattice and the system develops a macroscopic magnetization characterized by a permanent magnetic dipole moment. Not *all* of the spins point in the same direction (until one reaches absolute zero in temperature, at any rate, which is impossible in any macroscopic sample) but the *majority* do, with a fraction that increases to unity as one approaches zero temperature.

One last time we resort to our magnetization picture, this time (in 8.23) to illustrate the *permanent* macroscopic magnetization of a bar magnet in the *absence* of an external field.

8.11.2: The Curie Temperature and Neel Temperature

The critical temperature for the paramagnetic-ferromagnetic transition is called the *Curie Temperature* after Pierre Curie (the husband of the perhaps better-known Marie Curie), who showed that ferromagnetism was lost at this temperature. The critical temperature for the related antiferromagnetic transition is called the *Neel Temperature* for similar reasons.

Physicists find the classic ferromagnetic phase transition to be very interesting because it is an excellent example of the (sudden) emergence of *long range order* in a system that is disordered at high temperatures. The magnetic susceptibility of the system, the heat capacity of the system, and other thermodynamic descriptors of the system all do unusual things at the critical temperature of the phase transition, often exhibiting divergent or non-continuous behavior. Considerable effort has been expended on deriving a theory that accurately describes

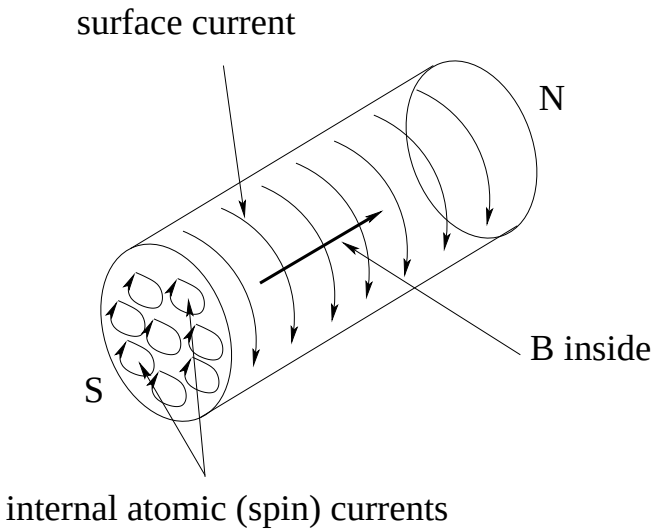


Figure 8.23: In a ferromagnet, the magnetic dipoles spontaneously align when cooled below a critical temperature. The resulting surface current transforms them into small "solenoids" with a non-dissipative surface current surrounding their interior volume and trapping magnetic flux that emerges from their north pole and flows to their south pole.

things like the particular value of the critical temperature and certain exponents that describe the divergences that occur there. These theories haven't been without some successes, but only a very few simple models have been solved *exactly*, notably the 2 dimensional Ising model mentioned and portrayed in cartoon form above.

However, we can now use powerful computers to simulate the behavior of "ideal" magnetic systems and compute their critical parameters with systematically improvable accuracy. These computations in turn can be used to check the theoretical predictions (since we lack "perfect" exemplars of the theoretical models in messy old nature).

Magnetism, Concluded

With this we'll wrap up our treatment of straight-up magnetic phenomena. As you can see, it is considerably more complicated than electrostatics even before the dynamical behavior associated with Faraday's Law is introduced.

Magnetic forces are right-hand twisty. They appear to violate Newton's Third Law, which *should* make you very worried about the consistency of physics and the laws of Conservation of Momentum and Angular Momentum. They appear or disappear, seeming to turn somehow into the electric force as we change inertial reference frames (transforming into a frame where a charge is at rest, for example).

The sources of magnetic fields are no less right-handed twisty. Fields circulate around moving *electric* charges, and although we might expect to find free magnetic charges, so far nobody has managed to salt the tail of one[121].

[121] Sorry, this is an ancient metaphor, associated with the idea that you can catch a bird by putting salt on its tail. It is used by bored parents to torment their four year old children who want to catch the pretty birdies. As in: "Oh,

Finally (and best of all), it looks like changing magnetic fields are somehow able to create electric fields! Magnetic induction is wonderfully complicated, with right hands twisting this way and that trying to simultaneously track the directions of currents, magnetic fields, electric fields produced by the magnetic fields, new currents created by the electric fields, and forces between all of these currents and the magnetic fields the sit in? And did I mention Lenz's Law, that makes all of the induced responses work backwards?

Furthermore, if we look at Maxwell's equations (so far) we have now seen the full set – two Gauss Laws, Ampere's Law, and Faraday's Law – and there is no sign yet of Maxwell. We *do* notice that the equations are getting more symmetric. Magnetic fields actually behave *almost* like electric fields and vice versa and it looks strangely like one can turn into the other if we merely look at it differently (changing reference frames, for example). However, they aren't quite right, somehow – Ampere's and Faraday's Law look like they *ought* to be more consistent, but we can't quite see how.

In a week, we're going to look at Maxwell's Equations again and make a startling discovery – the one due to Maxwell – that makes the set of equations *perfectly symmetric* except for the lack of magnetic charges, a problem that experimentalists might resolve tomorrow by finding one. Maxwell's addition will throw considerable *light*[122] on several puzzles in physics, and in the process give us plenty of stuff to study and learn for the rest of the semester.

But first, let's look at a complete different topic. Let's look at *harmonically alternating voltages* applied to electrical circuits containing inductances (L), resistors (R), and capacitors (C) as well as generators or other voltage sources that produce harmonically oscillating voltages. Along the way we will see how all of the things we have learned so far form pretty much *the basis for modern civilization*, given that modern civilization would regress to a form not seen for over a century overnight if our modern electrical power grid were to fail. You are finally knowledgeable enough to be able to *understand* the power grid – how electricity is generated, how it is transmitted long distances without significant losses, how it is used when it gets there in all kinds of work saving and life saving devices. You can also understand how electrical circuits can be combined to make *information processing devices* – radios, televisions, computers, cell phones, music players, networks – as well as a vast array of devices useful in medicine, business, industry, or the home.

Electricity helps make our cars and boats and planes and trains work, it cools our food to keep it fresh and cooks our food to make it safe and savory to eat, it cleans our dishes afterwards, it entertains us in all of the well-lit time we have to spare in the evenings in our electrically heated or cooled houses, a time when our ancestors only a hundred and fifty years ago either had to work or sleep for the lack of cheap light, huddling to keep warm in houses heated (if at all) with costly wood or coal. Electricity saws the wood that *builds* our houses, it weaves and sews the cloth we wear on our backs. Electricity enables us to grow far more food than we could without it, transport that food for vast distances, and store it safely until it is needed – cities would die almost overnight without it.

Nothing in human civilization is more important than maintaining and increasing the flow of

---

you want to catch that sparrow? All you have to do is put salt on its tail!" The child, of course, spends days in the field with a box of salt, trying to get close to birds. Birds, not being *that* stupid, fly away anytime the child and salt come near. Finally a great truth dawns on the child – you can't salt the tail of a bird you haven't already caught...

[122] Heh, heh. This is a pun, actually. If you don't get it now, in Yodaspeak: "You will. You will."

inexpensive electrical energy. With it, the poorest of our poor are wealthier than the wealthiest of the kings, emperors, and nobles of yesteryear. Without it, billions of humans would starve, our urban civilization would collapse, wars would erupt over access to food and other resources that electricity makes cheap and plentiful.

Yet – to get up, just a bit, on a political soapbox – our elected leadership and the population that elects them seem somehow to be blind to all of this. *Nothing* in human civilization is more important than ensuring an inexhaustible source of electrical energy to enable that civilization to continue, and yet we do almost *nothing* with our collective resources to construct an electrical grid that does not rely on scarce and exhaustible fuels, fuels that there are far better uses for than *burning* them.

There is plenty of non-scarce energy available on Earth to run a high level of civilization not just for the few, but for every person on the planet. The Sun, the wind and the water can provide us with power for as long as the Sun shines (some five billion more years), the wind blows (as long as the Sun shines), the water flows (ditto). If we must burn fuels, thermonuclear fuels such as deuterium are so abundant that they, too, are virtually inexhaustible – even if the Earth runs out in a billion years or so, there is all of the rest of the solar system to mine. Burning oil and coal, however, is simply inexcusible, except as a short time stopgap to keep civilization from collapsing while we change over to renewable or inexhaustible resources.

But to make this changeover, we require *political will*. We have to *invest* in the changeover, we have to *mandate* the changeover as a matter of social *will*. Until we have converted to renewable energy, human civilization will hang by an ever eroding thread over an abyss of misery. On the other hand, once we have converted *energy scarcity will never again be an important social or economic issue* and indeed, the world economies can actually stabilize by using the more or less fixed value of energy as a standard of monetary value. Nearly all scarcities in human affairs – water, food, living space, clothing, commodities – can be provided cheaply given only enough, cheap enough, electricity.

It is my hope that my students over the years, reading these words, will be inspired to take action and bring about the next great age of man, the unlimited energy age. But for you to have much hope of being effective, you have to *understand* electricity in a bit more detail than most people do. Hopefully the next chapter will help you accomplish that understanding.

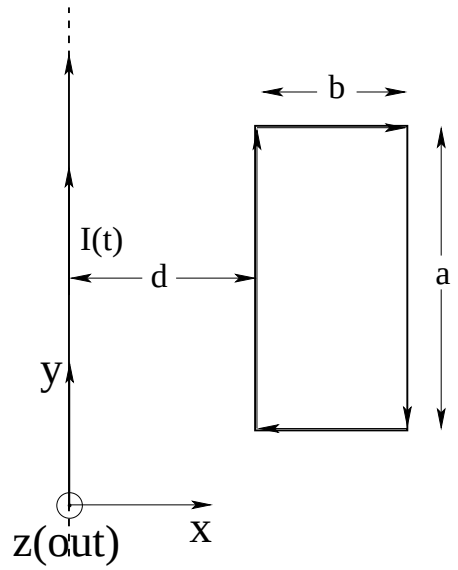
Homework for Week 8

Problem 1.

Physics Concepts

Make this week's physics concepts summary as you work all of the problems in this week's assignment. Be sure to cross-reference each concept in the summary to the problem(s) they were key to. Do the work carefully enough that you can (after it has been handed in and graded) punch it and add it to a three ring binder for review and study come finals!

Problem 2.



A a long straight wire carries a current $I(t) = I_0 \sin(\omega t)$. A rectangular loop of wire with resistance R and dimensions $a \times b$ is a distance d away as shown.

Find (as functions of time, in order):

a) The flux through the loop due to the wire;

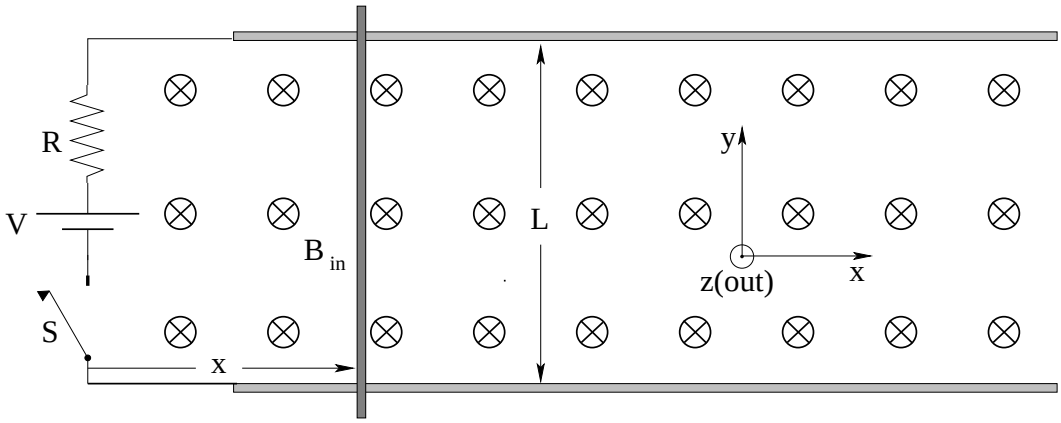
b) the mutual inductance M of the wire and loop;

c) the induced voltage in the loop;

d) the induced current in the loop;

e) the force between the loop and the wire.

Problem 3.



A rod of length L and mass m initially sits ***at rest*** on two frictionless conducting rails that sit in a plane perpendicular to a magnetic field as shown. At time $t = 0$ a switch S is closed connecting a voltage V that goes through a resistance R and the rod. Assume that the rod is initially at $x(0) = 0$.

In order:

a) Using *heuristic* reasoning (that is, not solving the equation of motion), what do you expect the *terminal velocity* of the rod to be after the switch has been closed for a long time? How do you expect it to approach this velocity? (Don't forget to give the direction and explain your reasoning!)

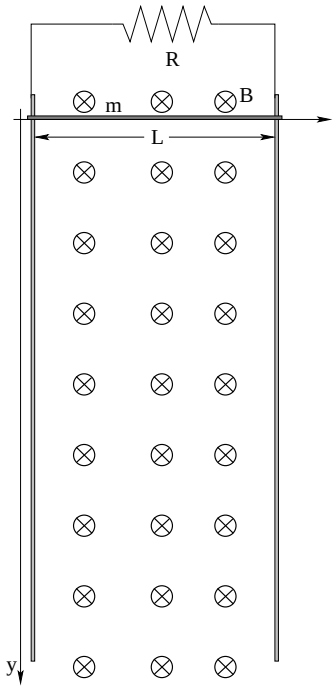
b) Write Kirchoff's Loop Rule for the circuit, including *both* the voltage of the battery *and* the induced voltage as the rod moves with a (presumed) speed v.

c) Find the current in the rod as a function of v.

d) Find the force acting on the rod as a function of v.

e) Write Newton's Second Law for the rod and solve the equation of motion for $v(t)$. Does it approach the terminal velocity the way you would expect?

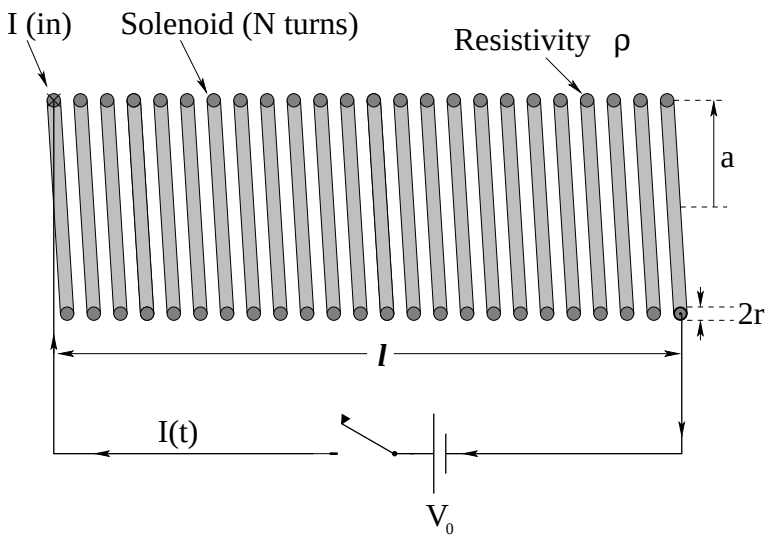
Problem 4.



A rod of length L and mass m rides on frictionless *vertical* conducting rails that sit in a plane perpendicular to a magnetic field as shown. A resistance R at the top completes a circuit. At time $t = 0$ the rod is released ***from rest*** and falls.

Using the methodology from previous rod-on-rails problems from the homework, lecture and the text repeat their steps to find $v(t)$ and $y(t)$, using down as positive and measuring $y(t)$ from its initial position.

Problem 5.

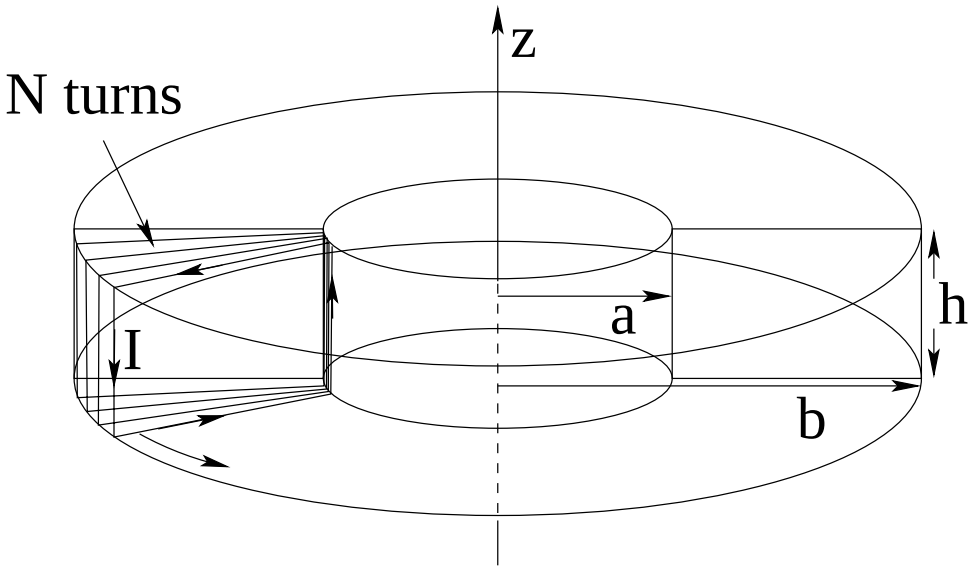


a) Find the self-inductance of the solenoid to the left that has N turns, length ℓ, and circular radius a.

b) The conducting wire used in making the solenoid has radius r (diameter $2r$ as shown, where $r \ll a$) and resistivity ρ. Find its resistance R.

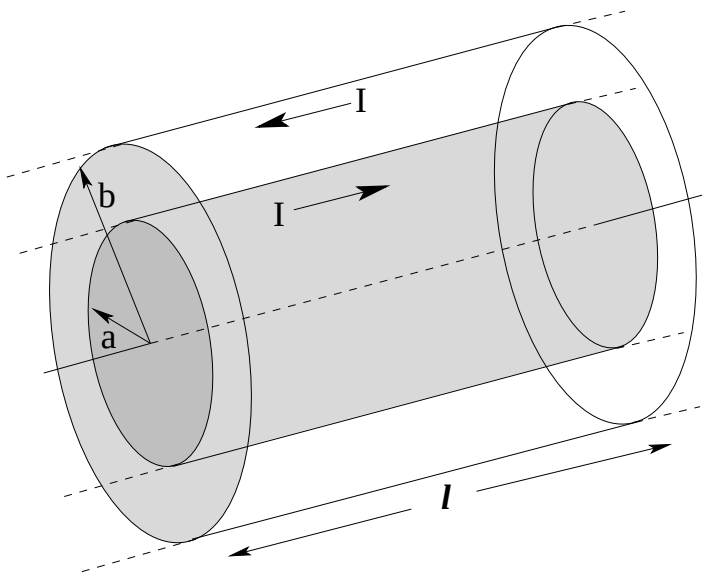
c) Find the current $I(t)$ in the circuit assuming that the switch is closed at time $t = 0$. Treat the wire used in the circuit itself (*not* including the solenoid) as having zero resistance, as usual.

Problem 6.



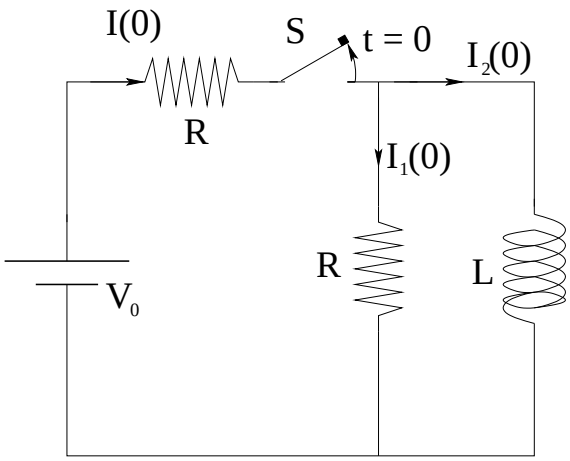
Find the self-inductance L of a toroidal solenoid of N turns that has inner radius a, outer radius b, and height h. Remember that this problem was started for you as a textbook example (and may have been worked for you in lecture as well).

Problem 7.



Find the self high-frequency self-inductance ***per unit length*** of a coaxial cable with inner conductor radius a, outer conductor radius b. This problem is set up for you in the textbook above and illuminated in the "self-guided learning problem" slides for this course if you have access to them.

Problem 8.



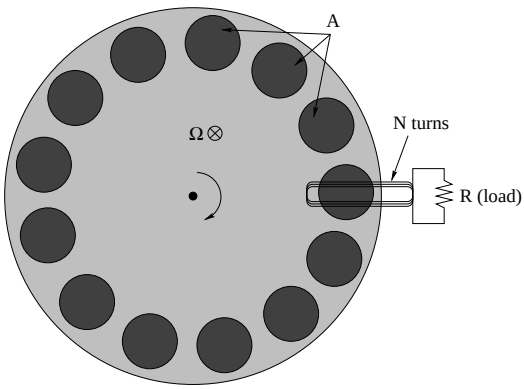
In the circuit above, switch S has been ***closed for a very long time***. At time $t = 0$ the switch is *opened*. Find:

a) The currents $I(0)$, $I_1(0)$, and $I_2(0)$ at $t = 0$ at the instant ***before*** the switch is opened.

b) Using Kirchhoff's voltage rule, find (derive) and solve the differential equation for $I_2(t)$. Draw a qualitative plot of this function.

c) Write an expression for the energy stored on the inductor as a function of time, using your answer to b). Draw a qualitative plot of this function.

Problem 9.



A magnetic braking system is drawn above. A wheel has M powerful permanent magnets mounted around the rim. Each magnet produces a uniform field B across a cross-sectional area A. As the wheel spins at angular velocity Ω, the magnets pass in front of a fixed coil with N turns in a circuit with a resistance R.

Estimate the braking power of the system as follows. Assume that each magnet produces a total *peak* flux $\phi = BA$ through a single turn of the loop. Also assume that the flux of each magnet ramps up ***linearly*** from zero to ϕ and then back down to zero in the time T_l required for the magnet to swing past a loop (so the flux is a "sawtooth" pattern as a function of time). Determine T_l as a function of Ω and M. Then:

a) Estimate the induced voltage and current during the ramp up and ramp down phases. Plot them as a function of time over several periods T_l, assuming constant Ω.

b) Compute the (effectively average) ***power as a function of*** $\boldsymbol{\Omega}$ and plot it as well for several T_l.

Advanced Problem 10.

In the previous homework problem, you should have gotten an average power (slowing the car down!) as a function of the angular velocity Ω of a single wheel of the of the general form:

$$P = -C\Omega^2 = \frac{dK}{dt}$$

for a constant C that depends on M, N, etc (the constant is not given here as you are supposed to have derived this in the previous problem).

This is the rate the kinetic energy of the car is being *reduced* while e.g. recharging the electronic or hybrid car's battery by this wheel. As the kinetic energy is reduced (probably by simultaneously braking at least two, maybe all four tires), the car will *slow down*. Note well that:

$$\Omega = \frac{v}{r_t}$$

where v is the speed of the car and r_t is the radius of the tire. The kinetic energy of the car is thus:

$$K = \frac{1}{2}m_c v^2 = \frac{1}{2}m_c \Omega^2 r_t^2$$

So here's the challenge: Use the *form* for the rate that energy is removed from the car by the coil to find $\Omega(t)$, $v(t)$ and $K(t)$, assuming that it begins at speed $v_0 = r_t\Omega_0$ and/or kinetic energy $K_0 = \frac{1}{2}m_c v_0^2$ at time $t = 0$ when the brakes are applied. **Big hint:** use ***calculus*** to work out the appropriate (first order, simple) differential equations and solve/integrate them to find the solutions requested.

Week 9: Maxwell's Equations and Light

I have also a paper afloat, with an electromagnetic theory of light, which, till I am convinced to the contrary, I hold to be great guns.

James Clerk Maxwell (1831-1879) Scottish physicist. In a letter to C. H. Cay, 5 January 1865.

- Ampere's Law has a bit of a problem. The current *through C* is not consistently defined so that it gives the same value for *all* surfaces S that are bounded by the closed curve C (through which we evaluate the flux of the current density to find the current "through C"). This means that two people can evaluate the integral to find the current through C and get *different answers* without either of them making a mistake. One can prove anything from a theory with an inconsistency, so this is a *bad thing*.

- James Clerk Maxwell noted this problem, and sat down to *invent* the mathematical tools and concepts to resolve it. We will proceed far more elegantly than he was able to, using the gift of hindsight. Either way, we will all arrive at the following *consistent* form for Ampere's Law, one to which we have added *Maxwell's Displacement Current*:

$$\oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} = \mu_0 \left(\int_{S/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA + \frac{d}{dt} \epsilon_0 \int_{S/C} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA \right)$$

  Both of these latter two integrals must be evaluated with the *same* surface S, but given this they sum together to give the same invariant current for *all* the surfaces S that are bounded by the closed curve C.

- In this new, *correct* version of Ampere's Law, you can see Maxwell's contribution: the *Maxwell Displacement Current* produced by a *time varying electric field*:

$$I_{MDC} = \frac{d}{dt} \epsilon_0 \int_{S/C} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA$$

- It is worth writing down the complete set of trading cards, suitable for engraving:

$$\oint_S \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA = \frac{1}{\epsilon_0} \int_{V/S} \rho_e dV \tag{9.1}$$

$$\oint_S \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA = \mu_0 \int_{V/S} \rho_m dV = 0 \tag{9.2}$$

$$\oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} = \mu_0 \left(\int_{S/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA + \frac{d}{dt} \epsilon_0 \int_{S/C} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA \right) \tag{9.3}$$

$$\oint_C \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{\ell}} = -\frac{d}{dt} \int_{S/C} \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA \tag{9.4}$$

- Physicists usually rearrange them to make the equations connecting fields to *sources* stand out from the equations that have no source terms (because we have yet to see a magnetic monopole):

$$\oint_S \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA = \frac{1}{\epsilon_0} \int_{V/S} \rho_e dV \tag{9.5}$$

$$\oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} - \frac{d}{dt} \mu_0 \epsilon_0 \int_{S/C} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA = \mu_0 \int_{S/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA \tag{9.6}$$

$$\oint_S \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA = 0 \tag{9.7}$$

$$\oint_C \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{\ell}} + \frac{d}{dt} \int_{S/C} \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA = 0 \tag{9.8}$$

This way, the symmetry *is compelling!* Two inhomogeneous equations have source terms connected to electric charge, two homogeneous equations have the *same form* but lack the source terms, at least until monopoles are discovered.

- If one applies these equations to a *source-free volume of space* where electric and magnetic fields are varying, one can show that they lead to the following *wave equations* for the *electromagnetic field* propagating in (say) the z-direction:

$$\frac{\partial^2 \vec{\boldsymbol{E}}}{\partial z^2} - \frac{1}{c^2} \frac{\partial^2 \vec{\boldsymbol{E}}}{\partial t^2} = 0 \tag{9.9}$$

$$\frac{\partial^2 \vec{\boldsymbol{B}}}{\partial z^2} - \frac{1}{c^2} \frac{\partial^2 \vec{\boldsymbol{B}}}{\partial t^2} = 0 \tag{9.10}$$

The $\frac{\partial^2}{\partial z^2}$ symbol in this expression, let me remind you, just means to take the derivative of the functions $\vec{\boldsymbol{E}}(\vec{\boldsymbol{x}}, t)$ and $\vec{\boldsymbol{B}}(\vec{\boldsymbol{x}}, t)$ with respect to the z-coordinate only, pretending that the other coordinates are constants. In this equation,

$$c = \sqrt{\frac{k_e}{k_m}} = \frac{1}{\sqrt{\epsilon_0 \mu_0}} = 3 \times 10^8 \text{meters per second} \tag{9.11}$$

is the *speed of light in a vacuum*, which we can see is *completely determined* from Maxwell's equations.

Since Maxwell's equations are laws of nature and expected to hold in all inertial reference frames, it is entirely *reasonable* to expect the speed of light to be constant in all reference frames! This postulate, together with some very simple assumptions about coordinate transformations, suffices to derive the theory of relativity!

- We will study the details of at least certain simple solutions to these wave equations over the next few weeks. For the moment, the most important solution for you to learn is:

$$E_x(z,t) = E_{0x}\sin(kz-\omega t) \tag{9.12}$$
$$B_y(z,t) = B_{0y}\sin(kz-\omega t) \tag{9.13}$$

known as a *harmonic plane wave* travelling in the z-direction. Note that E_x and B_y are *in phase* and do not have independent amplitudes – their amplitudes are connected by Maxwell's equations (Faraday or Ampere's law) and $E_x = cB_y$. There is an identical pair of solutions with a different *polarization*:

$$E_y(z,t) = E_{0y}\sin(kz-\omega t) \tag{9.14}$$
$$B_x(z,t) = -B_{0x}\sin(kz-\omega t) \tag{9.15}$$

that also propagate in the z-direction, as determined from the derivation of the wave equations above.

In these equations, note well that:

$$k = \frac{2\pi}{\lambda} \tag{9.16}$$

is the *wave number* of the wave, where λ is the *wavelength* of the harmonic wave, while:

$$\omega = \frac{2\pi}{T} \tag{9.17}$$

is the *angular frequency* of the wave. The wavelength is thus the "spatial period" of the wave, where T is the "temporal period" of the wave that harmonically oscillates in space and time. This wave propagates in the *positive* z-direction as can be seen by considering $kz - \omega t = k(z - \frac{\omega}{k}t) = k(z - ct)$. Note well that this uses the result that:

$$c = \frac{\lambda}{T} = \frac{\omega}{k} \tag{9.18}$$

for a harmonic wave.

- The flow of energy in an electromagnetic wave (and field in general) can be determined from the *Poynting vector*:

$$\vec{\boldsymbol{S}} = \frac{1}{\mu_0}(\vec{\boldsymbol{E}} \times \vec{\boldsymbol{B}}) \tag{9.19}$$

The magnitude of the Poynting vector is called the *intensity* of the electromagnetic wave – the energy per unit area per unit time or power per unit area being transported by the wave in the direction of its motion:

$$I = \frac{dP}{dA} = \frac{d}{dA}\frac{dU}{dt} = |S| \tag{9.20}$$

where U is the energy in the wave. To speak more mathematically precisely to communicate the transport of *power* (energy per unit time, in watts) across some given surface A, one evaluates the *flux of the Poynting vector through the surface*:

$$P_A = \int_A \vec{\boldsymbol{S}} \cdot \hat{\boldsymbol{n}}\, dA \tag{9.21}$$

As you can see one just cannot get away from flux integrals as a way of representing the "flow" of energy, current, fluid, or $\vec{\boldsymbol{E}}$ or $\vec{\boldsymbol{B}}$ field through a surface! As such, it is a very important idea to conceptually master.

- The Poynting vector can be understood and *almost* derived by adding up the total energy in the electric and magnetic fields in a volume of space being transported perpendicular to a surface A. In a time Δt, all of the energy in a volume $\Delta V = A\ c\Delta t$ goes through the surface at the end. This is:

$$\Delta U = (\frac{1}{2}\epsilon_0 E_x^2 + \frac{1}{2\mu_0}B_y^2)A\ c\Delta t \tag{9.22}$$

  If we use $|E_x| = c|B_y|$ (see above) for a wave travelling in the z-direction and do a bit of algebra, we can see that:

$$\frac{\Delta U}{A\Delta t} = \frac{1}{\mu_0}|\vec{\boldsymbol{E}}_x||\vec{\boldsymbol{B}}_y| \tag{9.23}$$

  which is just the Poynting vector magnitude in the z-direction for these two field components.

- The electromagnetic field also carries *momentum*, solving the dilemma of the "missing momentum" left over from our consideration of the magnetic force and the failure of Newton's third law. The field momentum is rather difficult to derive in a *simple* way, but it can *somewhat* be understood by assuming that the field *electrically* polarizes atoms that it sweeps over in such a way that it exerts a *magnetic* force along the direction of motion of the electromagnetic wave. We'll explore this with a problem later. The momentum density of the electromagnetic field is:

$$|p_f| = \frac{U}{c} \tag{9.24}$$

  and we can consider the net momentum transported per unit area per unit time by the electromagnetic field perpendicular to a surface A to be:

$$P_r = \frac{I_{\text{thru A}}}{c} \tag{9.25}$$

  This quantity is called the *radiation pressure* and it is partially responsible for the *solar wind*, created as sunlight pushes gas molecules away from the sun. Light "sails" have also been proposed as a propulsion for getting around inside the solar system without rocket fuel. We will explore both of these ideas with homework problems.

  To use radiation pressure properly, one has to compute the force it exerts on a surface. This force will depend on certain things, such as whether or not the radiation is perfectly absorbed or perfectly reflected and (eventually) the relative velocity of source and target (as the incident and reflected waves can be doppler shifted, affecting the momentum transfer). In the simplest cases (perfect absorption or reflection) the force is best computed by using an expression such as:

$$F_S = \frac{1}{c}\int_A \vec{\boldsymbol{S}} \cdot \hat{\boldsymbol{n}}\ dA \tag{9.26}$$

  that is, the flux of the Poynting vector yields the power transferred to a (perfectly absorbing) surface, and $1/c$ of the power is the effective force exerted along the line of the original Poynting vector. If the radiation is reflected, one has to construct a such quantity evaluated (with the same power) with respect to the direction of the angle of reflection, and vector sum the forces. In the simplest case of normal absorption or reflection:

$$F_S = \frac{SA}{c} \tag{9.27}$$

or

$$F_S = \frac{2SA}{c} \tag{9.28}$$

respectively.

- Electromagnetic radiation is produced when electrical charges *accelerate* (this follows from construction the *inhomogeneous wave equations* for the electromagnetic fields directly from Maxwell's equations, where moving charge and current terms become the sources of the time varying fields). In fact, if one works very hard in a graduate Electrodynamics class (as shown in my online book, for example, or in J. D. Jackson's *Classical Electrodynamics*) one can show that the *power cross-section* of a single charge q moving along the (say) z-axis is:

$$\frac{dP}{d\Omega} = \frac{q^2}{16\pi^2\epsilon_0}\frac{1}{c^3}\left|\frac{d^2z}{dt^2}\right|^2 \sin^2(\theta) \tag{9.29}$$

  The power cross section is the amount of power per unit solid angle ($d\Omega$) radiated away from the *accelerating charge*. The actual power then drops off like $1/r^2$ in this direction.

  A direct consequence of this result is the death of classical physics. Classically, we expect an electron to *orbit* a proton in a hydrogen atom, much the way the moon orbits the earth. After all, the forces of attraction between them have a more or less identical form! But if an actual hydrogen atom were bound in this way, the electron (like the moon) would be more or less perpetually *accelerating*. It would therefore be more or less perpetually *radiating away energy* and dropping into a lower orbit to provide it. If one considers how long it would take before an electron in a circular orbit around a proton with an initial radius around 10^{-10} meters (one Angstrom, roughly the size of almost any atom) to spiral in to the proton, it is a very, very short time (as the further in it gets the more strongly it mus accelerate and the faster it radiates to a still lower energy orbit with a still smaller radius). In a tiny fraction of a second, the classical "atom" would collapse!

  The fact that this manifestly does *not* occur, when it *must* occur if both Newton and Maxwell are correct, is one of several factors that led to the invention of quantum mechanics and modern physics (including relativity theory). This, then, is the next course in physics that students beginning a serious study of physics should undertake, as soon as they complete this one and solidify their understanding of classical electricity and magnetism and light. Things are getting interesting!

- When one considers a point charge oscillating around is oppositely charged mate (a dynamical version of our Lorentz model for an atom that helped us understand dielectric polarization earlier) one can either convert this expression into or derive directly from the Poynting vector the following expression for the power cross-section:

$$\frac{dP}{d\Omega} = \frac{c^2}{32\pi^2}\sqrt{\frac{\mu_0}{\epsilon_0}}k^4\left|p_z\right|^2\sin^2(\theta) \tag{9.30}$$

  The $k^4 = (2\pi/\lambda)^2$ is very important, as it is why the sky is blue! Remember it for later – shorter wavelength/higher frequency light waves have a much larger power cross-section, all things being equal, than longer ones, because the fields are related to the *time derivatives* of the dipole moments which increase with the frequency. Again, the actual power radiated away in any direction drops off like $1/r^2$.

- Finally, one can (as usual) consider the *collective* radiation from *many* charged particles oscillating against a neutral background in, for example, an *antenna*. An antenna is basically a wire that has a current in it such that it forms a macroscopic dipole moment (in say the z-direction) that oscillates at some frequency ω. This antenna will then radiate away energy in the form of electromagnetic radiation!. The power cross section is basically the same as that just given (but for a much larger dipole moment p_z), so that the *intensity* of the radiation field of a z-oriented dipole antenna located at the origin of a spherical polar coordinate system is usually given by:

$$I(\theta) = \frac{P_0}{r^2}\sin^2(\theta) \tag{9.31}$$

(and is azimuthally symmetric about the z-axis). P_0 has the units of power, and intensity has units of power per unit area, so this works. It is often given as:

$$P_0 = I_{\rm rms}^2 R_{\rm rad} \tag{9.32}$$

where $I_{\rm rms}$ is the root-mean-square current in the antenna and $R_{\rm rad}$ is the *radiation resistance* of the antenna, which can heuristically be thought of as resulting from the *reaction force* exerted on the radiating charges due to their own radiated field! Deriving these results is beyond the scope of this course, but it is nevertheless useful to understand and use the terminology when we consider radios (as we saw last week). Note well that the radiation is most strongly emitted *perpendicular to* the dipole moment, and that no energy at all is radiated *along* the dipole moment.

9.1: Ampere's Law and the Maxwell Displacement Current

As discussed at the end of week 8, Maxwell's Equations – so far – don't seem quite right. Let's write them out as we have them at this point:

$$\oint_S \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA = \frac{1}{\epsilon_0} \int_{V/S} \rho_e dV \tag{9.33}$$

$$\oint_S \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA = \mu_0 \int_{V/S} \rho_m dV = 0 \tag{9.34}$$

$$\oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} = \mu_0 \int_{S/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA \tag{9.35}$$

$$\oint_C \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{\ell}} = -\frac{d}{dt} \int_{S/C} \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA \tag{9.36}$$

The asymmetry will be a bit more apparent if I put all of the terms involving *charges* as *sources* of the fields on the right and all of the terms involving the fields themselves on the left:

$$\oint_S \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA = \frac{1}{\epsilon_0} \int_{V/S} \rho_e dV \tag{9.37}$$

$$\oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} = \mu_0 \int_{S/C} \vec{\boldsymbol{J}}_e \cdot \hat{\boldsymbol{n}} dA \tag{9.38}$$

$$\oint_S \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA = 0 \tag{9.39}$$

$$\oint_C \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{\ell}} + \frac{d}{dt} \int_{S/C} \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA = 0 \tag{9.40}$$

I put a tiny $_e$ subscript on the $\vec{\boldsymbol{J}}$ and reordered them *with a big hole in Ampere's Law* to emphasize the point. The top two equations are connected to *electrical charge* – either stationary or moving – to produce the fields. The bottom two are zero on the right, where the zero just means "there ain't no stinkin' magnetic monopoles been seen (yet)" but we can *imagine* that if there were, Gauss's Law for Magnetism would get a source term on the right that looked just like that for Gauss's Law for Electricity, and Faraday's Law would get a term on the right involving the *current density* of *moving* magnetic charge, just like Ampere's Law.

But what about poor Ampere's Law, in that case? Faraday's Law mixes electric and magnetic fields, so that time varying magnetic fields make electric fields.

Shouldn't Ampere's Law have a term such that time varying electric fields make magnetic fields? I left the gap just in case...

This is as good a thing as any to motivate a closer look at Ampere's Law. Maxwell's Equations are starting to look rather *beautiful*[123] but that big hole is *ugly*, as are (really) the big ugly zeros where magnetic monopoles should live. Natural philosophers have from time immemorial considered "beauty" – a certain appealing symmetry, as it were – to be an essential component of probable truth. Sometimes this belief is followed to a fault, of course, especially when the beautiful idea in question is *our* idea and we ignore the fact that it doesn't, actually,

[123]Seriously. If there is such a thing in this Universe as beautiful mathematics, Maxwell's Equations are It. This course won't cover the half of just how gorgeous they really are...

agree with nature particularly well when we look[124]. Ultimately nature itself must be the arbiter of truth in natural law, but still, at the very least, things that are *almost* beautifully symmetric demand a closer examination to see if we are missing something because symmetry *is* – empirically – often observed in nature! Experimentalists today continue the search for magnetic monopoles; we ourselves will follow in Maxwell's footsteps and search for the "problem" – that will turn out to be a missing term as we might guess from symmetry – in Ampere's Law.

9.1.1: The Problem with Ampere's Law – So Far

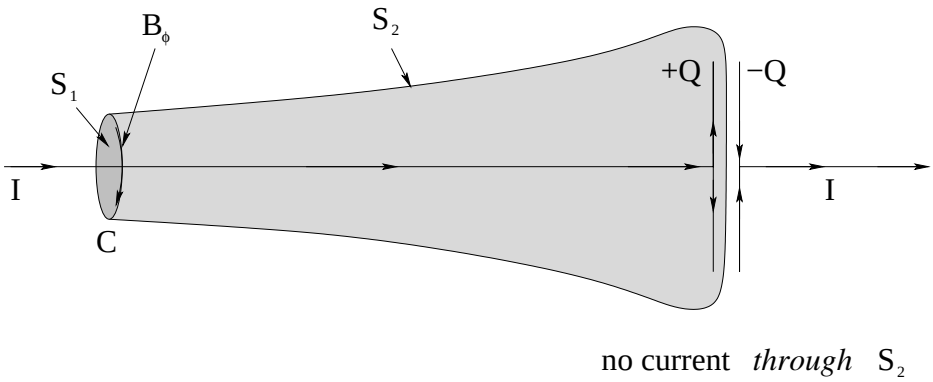


Figure 9.1: A simple circuit and pair of surfaces that illustrate how Ampere's Law is (so far) *wrong*, with two completely different currents for the two surfaces S_1 and S_2.

We have learned enough at this point to be able to see that Ampere's Law is ***obviously*** wrong – or at least, not mathematically consistent! – from a few simple, specific, examples that illustrate the problem. In figure 9.1 I've drawn a side view of a humble parallel plate capacitor. At this particular instant, a current $I(t)$ is flowing along the wire on the left, charging up the capacitor so that a charge $+Q(t)$ is increasing on the left plate.

To this innocuous looking problem we'll apply *Ampere's Law* – specifically to the nice circular loop C drawn around the supply wire. This loop is quite far away from the capacitor, and the electric field the capacitor is making is more or less confined to live between its plates, and the current I quite obviously goes *through* the surface S_1 stretched across C (and hence goes "through C"), so we should be quite justified in deducing the usual:

$$\oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} = B_\phi 2\pi r = \mu_0 \int_{S_1/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA = \mu_0 I \tag{9.41}$$

(where recall that S/C should be read as "the open surface S bounded by the closed curve C") so that

$$B_\phi = \frac{\mu_0 I}{2\pi r} \tag{9.42}$$

around the circle in the right handed sense. No problem, the field of an infinitely long straight wire carrying current, the simplest possible situation. How could this be wrong?

But wait. When I wrote the right-hand side of Ampere's Law, I *happened* to choose the "easy" surface S_1 that stretches straight across the curve C (and an easy curve C that lies in

[124] A serious problem with pre-Enlightenment philosophy...

a plane). However, there is nothing in the *mathematics* of Ampere's Law that requires me to use that particular surface.

Indeed, I *could* choose to use (say) surface S_2/C instead! S_2 is just as "bounded by the closed curve C" as S_1 is. They are topologically equivalent – S_1 is like the film of soap stretched across a bubble blowing loop, and S_2 is like the bubble as it has been blown out but is still attached to the loop. The only problem with this is that the current:

$$I = \int_{S_2/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA = 0(!) \tag{9.43}$$

because the surface S_2 goes in *between* the plates of the capacitor, where *no charge flows*!

This is a disaster! Ampere's Law seems to give us two possible answers. In fact, since there are an infinite number of surfaces S I could draw bounded by C that intercept different parts of the capacitor and wire supplying it, there are an infinite number of possible answers! But the two answers $B_\phi = \mu_0 I/2\pi r \neq 0$ and $B_\phi = 0$ are more than enough for us to see that we have a serious problem to deal with. The *current on the right hand side of Ampere's Law* (correctly evaluated as the flux of the current density through a surface bounded by the curve C) is not *invariant* when we vary the surface S in perfectly reasonable ways.

Now, in this particular example, based on the specific curve and geometry illustrated in figure 9.1, one *could* argue that using S_2 is silly – that we should "obviously" use the surface S_1 that lies in the same plane as C (or otherwise choose a "special" surface) so that we'll get the right answer. You should be deeply suspicious of this argument, of course – it sounds rather like choosing the surface on the basis of the fact that it gives the right answer instead of finding the right answer from the equation no matter *how* we choose the surface.

In fact, it is easy (and educational) to construct a simple counterexample to this assertion, one where there *is no possible way to a priori which of two surfaces S/C to use*. Both of the two "obvious" surfaces that stretch across C in the way closest to the way the plane surface S_1 stretches across the circular curve C in the example above turn out to be *identical* – simple rotations of one another, in fact – and (for what it's worth) empirically *neither* of them will give the right answer for the broken version of Ampere's Law!

Consider figure 9.2, the "potato chip" case[125]. In this case we imagine the curve C to be a circle that is bent over in just the right way so that there are *two* surfaces that are bounded by it that are *identical* – so much so that S_2 is a simple rotation of S_1 that has *exactly the same boundary curve C, in exactly the same orientation!*

If you put them together just right, the two surfaces thus joined make the *closed* surface $S = S_1 + S_2$, and we can clearly push a current through (say) S_2 that *does not exit the volume through* S_1 as illustrated to accumulate a nonzero total charge density ρ (and hence total charge Q) in the volume bounded by S. Obviously, there is now *no possible way* to decide which of these two surfaces to use as "the" correct surface to use so that the broken version of Ampere's Law will yield the correct answer[126].

[125] If you live in a country where "Pringles" potato chips – the ones sold in a simple cylindrical package where they are neatly and perfectly stacked – are available, the "saddle" shapes in question are nearly identical to the shape of a Pringles potato chip. If you take two such chips and rotate them just right and put them rim to rim, you can *almost* perfectly enclose a volume with a closed surface made up of potato chips. Sadly, I doubt that Pringles will send me any money for so shamelessly plugging their topologically useful product...

[126] As noted, *neither* of them.

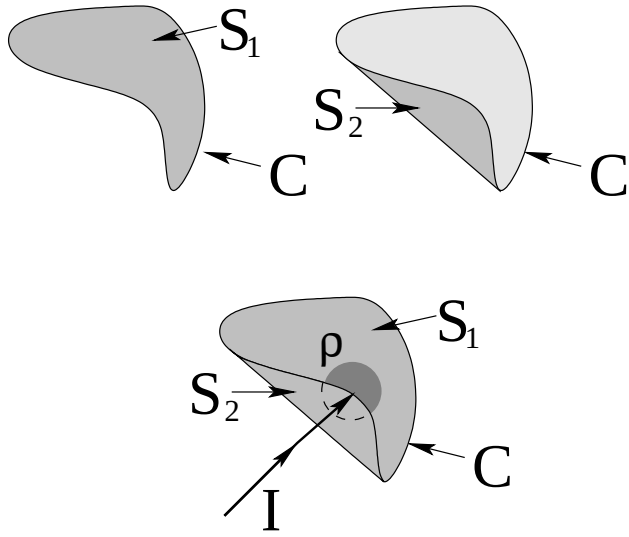


Figure 9.2: The "potato chip" case. Surfaces S_1 and S_2 are bounded by the same curve C and together form a closed surface that *completely encloses a charge density* ρ*!*

You can try to argue even further, that we must be sure to always use *both* "nice" curves C (ones in a plane, for example) *and* "nice" surfaces S (ones in that same plane) but a) that isn't very satisfactory, mathematically; and b) ***empirically, it doesn't work!*** Moving *point* charges make magnetic fields, and those fields pretty much *always* are going to violate the broken version of Ampere's Law on almost *all* curves C and surfaces S simply because the "current" through any given S is always going to be zero except for the tiny instant that the point charge crosses it, but there is going to be a magnetic field around C at least *some of the time* as the charge moves along a trajectory that would *eventually* pass through it, maybe, if it didn't stop first or curve away.

Basically, these explicit examples demonstrate that that so far, Ampere's Law is really just ***Ampere's Sort of OK Rule That Works, Sometimes, For A Subset Of All Possible Cases, If We Cheat***. This simply won't do. We want a natural *law* to *always* work – it has to be "unbreakable", especially by as simple a thing as bending C into a 3D twisted loop (like a crumpled coat hanger) or choosing the "wrong" C-bounded surface S for some perfectly reasonable plane loops C. Wrong by what standard? How can we decide that *any* of the variations are "wrong" without knowing the answer some other way, if the law itself isn't invariant across all possible choices?

Mathematicians and physicists get very anal about this sort of thing. If they don't, the bugaboo of all human efforts to reason, *inconsistency*, creeps into our set of beliefs, and mathematicians all well know that you can prove *anything* from a contradiction (and hence know *nothing* on the basis of your proofs)[127].

[127] In fact, by insisting that Maxwell's Equations *as* natural laws ought to be invariant under changes of inertial reference frame, Einstein threw out more or less *all of classical non-relatistic physics* – and was backed up by numerous experients that showed that he was *right* to do so! Kind of scary, that...

9.1.2: The Invariant Current through S/C

Our job, it appears, is to try to make the current in Ampere's Law invariant:

$$\oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} = \mu_0 I_{\rm inv} \text{ through } S/C \tag{9.44}$$

so that it gives us the exact same current for *any* surface S/C we might happen to choose to solve a problem. Obviously, we also want it to give the known/observable *right* answer for (say) the simple capacitor example illustrated above. If it works for a few cases like this where we know the answer from experiment and can compute it more than one way, we can even hope that the answer thus obtained from our new, improved version of Ampere's Law will *always* agree with experiment and is *indeed* a natural law! After all, that's the only real game in town!

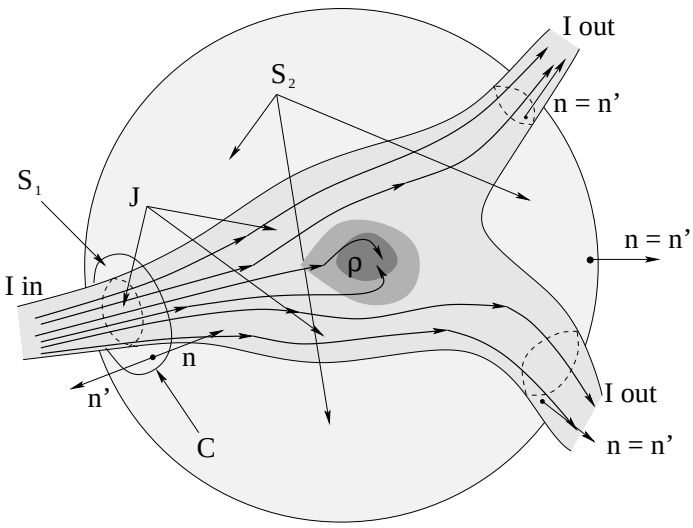


Figure 9.3: A very general current density flows through space. Some current flows in from the left and exits on the right, but some builds up in the current density ρ in the volume between the two surfaces S_1 and S_2. The point is that the *difference* between the flux (current) in through S_1 and out through S_2 must be equal to the rate that charge builds up in between, because charge is *conserved*.

The picture that will best help us find the invariant current is drawn in figure 9.3. We are going to take this picture and think about it in the light of another physical law that we really believe in, the *Law of Charge Conservation*. You might recall that we used figure 5.2, more or less equivalent to 9.3, in an alternate form of our *derivation* of the integral form of the law of charge conservation way back in chapter 5. As you should now be able to see from the examples above, Ampere's Law fails to consistently *account* for charge conservation in any case other than steady state (time independent or very slowly varying) current flow. When we analyze what happens for the two surfaces above and include Gauss's Law for Electrostatics consistently, the *correct* invariant current will more or less fall out of our analysis at our feet, ready to be plugged into Ampere's Law to make it correct.

In figure 9.3, I've chosen two simple surfaces S_1 and S_2 bounded by C. In fact, they are both parts of a sphere, and together they make a *closed* spherical surface, one that encloses the volume V inside. As was the case in the advanced discussion of charge conservation, the

current density $\vec{\boldsymbol{J}}$ flows in through surface S_1, but not *all* of it flows out through S_2. Some of it is building up in a charge distribution ρ inside the sphere. So the total current I flowing *in* to the sphere is larger than the total current flowing *out*. None of this – the choice of a sphere, the particular curve C or surfaces S_1 or S_2 – is important; we just choose them to make the result easy to see.

Since charge is conserved (empirical law!), the *rate* that charges builds up inside the closed surface $S = S_1 + S_2$ will equal the *difference* the the flux of the current densities:

$$\int_{S_1/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA - \int_{S_2/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA = \frac{d}{dt} \int_{V/S} \rho dV \tag{9.45}$$

which is exactly what we arrived at as equation 5.28, a version of the law of charge conservation, before. In this equation, the normals $\hat{\boldsymbol{n}}$ in the two integrals on the left are directed from the left to the right, in the direction of the current's apparent flow. You should feel free to go back and review the relevant part of chapter 5 and reread the discussion there if this is not clear.

The integral on the *right* looks strangely familiar! In fact, it is part of Gauss's Law for Electricity! Using Gauss's Law (multiplied on both sides by ϵ_0) we can substitute:

$$\int_{V/S} \rho dV = \epsilon_0 \oint_S \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}}' dA \tag{9.46}$$

(where ***note well*** that $\hat{\boldsymbol{n}}'$ on the right is the ***outward directed normal*** in GLE).

We substitute this back into the first equation to get:

$$\begin{aligned} \int_{S_1/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA - \int_{S_2/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA &= \frac{d}{dt} \int_{V/S} \rho dV = \frac{d}{dt} \epsilon_0 \oint_S \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}}' dA \\ &= \frac{d}{dt} \epsilon_0 \left\{ \int_{S_1} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}}' dA + \int_{S_2} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}}' dA \right\} \end{aligned} \tag{9.47}$$

where I have broken the flux integral on the right hand side into two pieces, one over S_1 and one over S_2. This is perfectly all right since the entire closed surface $S = S_1 + S_2$.

Next, since $\hat{\boldsymbol{n}}'$ in the two field flux integrals is the *outward directed normal* of GLE, it goes from left to right on S_2, but on S_1 *it goes from right to left!* I want to make $\hat{\boldsymbol{n}}$ exactly the same (left to right) in the field flux integrals as it is in the current density flux integrals on the *left hand side* of the equation, so I *change the sign of the S_1 integral* (and thereby can change $\hat{\boldsymbol{n}}'$ to $\hat{\boldsymbol{n}}$ in *both* integrals):

$$\int_{S_1/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA - \int_{S_2/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA = -\epsilon_0 \frac{d}{dt} \int_{S_1/C} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA + \epsilon_0 \frac{d}{dt} \int_{S_2/C} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA \tag{9.48}$$

Finally, we move all of the S_1 integrals to the left, and all of the S_2 integrals to the right:

$$\int_{S_1/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA + \epsilon_0 \frac{d}{dt} \int_{S_1/C} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA = \int_{S_2/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA + \epsilon_0 \frac{d}{dt} \int_{S_2/C} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA \tag{9.49}$$

The left side only depends on S_1/C. The right depends only on S_2/C. We used no special properties of these curves or surfaces beyond the fact that any two non-coincident open surfaces bounded by the same closed curve C enclose a volume. The two sides are thus *invariant*

under any possible change in the curves C or surfaces S. We thus define the *invariant current* to be:

$$I_{\text{invariant, "through C"}} = \int_{S/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA + \epsilon_0 \frac{d}{dt} \int_{S/C} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA \tag{9.50}$$

where the result now holds for *any* surface S bounded by *any* given closed curve C!

Let us now *guess* – not prove – that this invariant current is the *correct* one to use in Ampere's Law. If it isn't, Ampere's Law is in deep trouble, as any other form will be inconsistent with the Law of Charge Conservation and/or will produce a "current" we'll have to "cheat" to evaluate by knowing the answer and selecting particular choices for C and S so that it somehow works out.

Fortunately, we can immediately check to see if this invariant current works in at least one problem where we *do* know the correct answer – the specific capacitor example above where Ampere's Law as it was before got it wrong.

That is, we suppose Ampere's Law is really:

$$\oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} = \mu_0 I_{\text{invariant, "through C"}} = \mu_0 \left\{ \int_{S/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA + \epsilon_0 \frac{d}{dt} \int_{S/C} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA \right\} \tag{9.51}$$

which *no longer depends* on our choice of S/C. Note well the location of the brackets: the μ_0 is *outside* of them, and everything inside of them has the units of current.

If we apply this version of Ampere's Law to our pathological counterexample, the capacitor problem above, when we compute the *invariant* current through S_1 we still get the actual current I in the wire (because the $\vec{\boldsymbol{E}}$-field due to the capacitor is confined to live in between the plates of the capacitor and doesn't pass through S_1 at all in our usual idealization). This leads us to the expected answer for the field of a long straight wire, which we *also* evaluated directly from the Biot-Savart Law in week/chapter 7 and which agrees with Ampere's original current balance experiments – empirically, this is bound to be right.

If we apply it to the surface S_2, then as before no *physical* current gets through, but the magnitude of the *field in between the plates of the capacitor* is (recall):

$$E = \frac{\sigma}{\epsilon_0} = \frac{1}{\epsilon_0} \frac{Q}{A} \tag{9.52}$$

where A is the area of the capacitor, where the field goes from left to right across S_2. This field is nonzero only between the plates of the capacitor, so integrating over S_2 only gets a contribution from the area A where the field is non-zero, uniform in magnitude, and parallel to $\hat{\boldsymbol{n}}$ as we have drawn S_2. Therefore the flux integral is:

$$\int_{S_2/C} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA = EA = \frac{Q}{\epsilon_0} \tag{9.53}$$

exactly as one expects, and hence:

$$I_{\text{inv}} \text{ (through } S_2/C) = \epsilon_0 \frac{d}{dt} \int_{S_2/C} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA = \epsilon_0 \frac{d(EA)}{dt} = \frac{dQ}{dt} = I_{\text{inv}} \text{ (through } S_1/C) \tag{9.54}$$

because I in the wire is, in fact, the rate at which the capacitor is charging! ***We get the same I for both surface, and for both surfaces this leads to the correct field around C!***

Needless to say (I wouldn't have presented all of this hard work for nothing, after all) this version of Ampere's Law gives the right answers for *all* classical charge-current densities where "right" means only "in agreement with experiment within experimental error and the fact that we are ignoring quantum mechanics" as it should. From now on, if someone asks for "Ampere's Law", while you can still use the broken version in magnetostatic, steady state current density problems, you should remember *this* version:

$$\boxed{\oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} = \mu_0 \left\{ \int_{S/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA + \epsilon_0 \frac{d}{dt} \int_{S/C} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA \right\}} \tag{9.55}$$

with its invariant current as the *one and only correct* version of Ampere's Law and remind yourself that you are neglecting the term involving the derivative of field flux *because it is zero* in problems of this sort.

The extra term we have added to the physical current was originally added by James Clerk Maxwell, and the implications of this term ae so profound, so overwhelming, that the entire set of equations (and the term itself) were named in his honor. It is called the *Maxwell Displacement Current*:

$$I_{\mathrm{MDC}} = \epsilon_0 \frac{d}{dt} \int_{S_2/C} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA \tag{9.56}$$

As we've seen, for many "static" problems where there is no time-varying electric field we can use the old form without error, but it won't work when charge is building up and the electric field "through C" is varying! In fact, there is one *very important* place where the old form fails. It fails to describe the magnetic field *inside* the parallel plate capacitor. Let's work that out as an example.

Example 9.1.1: The Magnetic Field Inside a Parallel Plate Capacitor

In figure 9.4 we see a parallel plate capacitor with cylindrical symmetry being charged by a (momentarily) steady current I. As charge flows onto the capacitor, the field (assumed as usual to be strictly confined to be between the two plates, ignoring the fringe) increases uniformly. This increasing field creates an increasing flux through cylindrically symmetric Amperian loops of radius r in between the plates, generating a magnetic field there. Our job is to evaluate this field, both between the plates and in free space outside of the plates (but in the plane that separates them).

This description is a perfect recipe for our algebraic work, yet another example of how a *verbal* understanding of the physics plus knowledge of the laws and ability to do relatively simple math suffices to enable one to solve problems that at first glance are quite difficult. We imagine that at some time t the capacitor has a total charge $Q(t)$ on it such that $I = dQ/dt$.

Then (from Gauss's Law):

$$E = \frac{\sigma}{\epsilon_0} = \frac{Q}{\epsilon_0 A} \qquad r < R \tag{9.57}$$

(from left to right – remember that the field is a vector) and $E = 0$ for $r > R$. This just represents in an equation and a solution that by now should be *very* familiar to you the first step in the recipe above.

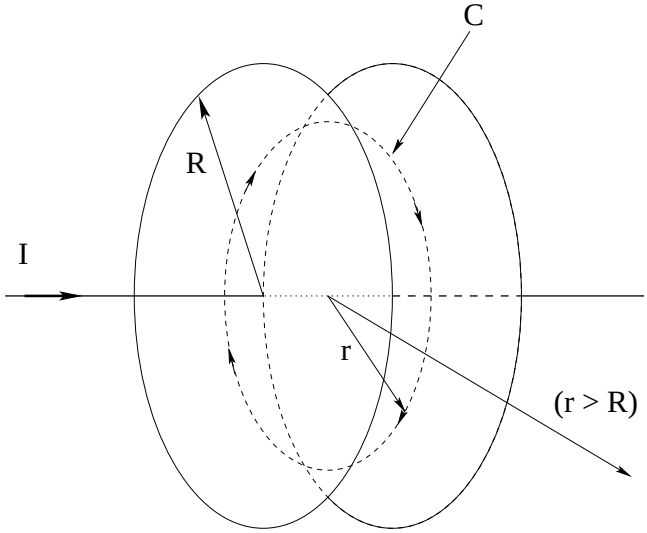


Figure 9.4: A capacitor made up of two circular disks is being charged by a current I. The increasing electric field between the two plates becomes a *Maxwell Displacement Current* that creates a magnetic field identical to the one that would exist inside a uniform conductor of the same radius (assuming the conductor had a magnetic permeability and electric permittivity identical to the vacuum value, not really a very good assumption).

Second, we have have to evaluate the flux through the Amperian path C (for $r < R$) in figure 9.4:

$$\phi_C = \int_{S/C} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA = EA = \frac{Q\pi r^2}{\epsilon_0 A} = \frac{Q\pi r^2}{\epsilon_0 \pi R^2} \tag{9.58}$$

(where we have used $A = \pi R^2$ at the end).

Third, we have to write Ampere's Law for this Amperian path:

$$\oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} = B_\phi 2\pi r = \mu_0 \left(\int_{S/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA + \epsilon_0 \frac{d}{dt} \int_{S/C} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA \right) \tag{9.59}$$

$\vec{\boldsymbol{J}} = 0$ (no actual current flows through the insulating vacuum between the plates) and the only thing that varies with time in the flux is the charge Q, so this becomes:

$$B_\phi 2\pi r = \mu_0 \frac{\frac{dQ}{dt} r^2}{R^2} = \frac{\mu_0 I r^2}{R^2} \tag{9.60}$$

We rearrange this to obtain half of our answer:

$$B_\phi = \frac{\mu_0 I r}{2\pi R^2} \qquad r < R \tag{9.61}$$

(in the right handed direction around the current onto the disk as shown).

If we choose the larger Amperian path C at $r > R$, the only thing that changes is that the flux is no longer a function of r, as the field is nonzero only in between the plates and equals $\phi_C = \frac{Q}{\epsilon_0}$ there. The field (after the same basic algebra) becomes:

$$B_\phi = \frac{\mu_0 I}{2\pi r} \qquad r > R \tag{9.62}$$

Note two things. First, the two algebraic forms for B_ϕ are equal at $r = R$, the boundary between the two regions. Second, on the *inside* the field is the same as the field one would expect in a wire of radius R carrying a uniform current I (and vanishes at $r = 0$ as might be expected), while on the *outside* the field is that of an infinitely long straight wire. These two observations are strong algebraic evidence that our displacement current has indeed "solved" the problem of finding an invariant current that gives us sensible answers regardless of the path C or surface S chosen that is bounded by it.

9.2: Advanced Topic: Origins of the Magnetic Field

In chapter 7, we noted that at least in the non-relativistic limit where $v \ll c$, the $\vec{\boldsymbol{B}}$-field of a unformly moving point charge could be sort-of derived from the Biot-Savart Law, and that the Biot-Savart Law could sort-of be turned into Ampere's Law (broken version). We *also* saw in chapter 8 that there were some very interesting consequences from looking at what happens to the magnetic field and force when we hop from a frame where a charge is moving in a uniform $\vec{\boldsymbol{B}}$-field and no $\vec{\boldsymbol{E}}$-field is present (so it *only* experiences a magnetic force) to a frame where the charge is *not* moving, experiences *no* magnetic force, but still experiences a force that must, somehow, be due to an electric field that appeared out of nowhere. That field, as it turned out, was directly related to the changing flux of the magnetic field!

This leads us to the question: Can we play the same sort of game now, only backwards? Can we (for point charges moving in the non-relativistic $v \ll c$ limit) look *only* at the changes in electric field flux – since there will be *no* physical charge current through *almost all* possible surfaces S/C in Ampere's Law, after all – the surfaces where the point charge is *on* S are a set of measure zero in the set of all surfaces in mathematese, and require careful treatments of various infinities surely unsuitable for an introductory course – and obtain e.g. the magnetic field of a point charge from Ampere's Law with the MDC *only*, and hence work backwards to the Biot-Savart Law etc?

The answer, amazingly enough, is *yes*. The following is due to Robert Buschauer[128]. We start with a point charge q sitting at rest at the origin. As we well know, its *static* fields are then:

$$\vec{\boldsymbol{E}} = \frac{k_e q}{r^2}\hat{\boldsymbol{r}} \qquad \vec{\boldsymbol{B}} = 0 \tag{9.63}$$

Now we imagine changing frames into a frame that is moving (say) in the $-\hat{\boldsymbol{z}}$ direction at speed $v \ll c$. In this frame, the charge is moving at velocity $\vec{\boldsymbol{v}} = v\hat{\boldsymbol{z}}$ in the positive z direction, and we'll consider it to be at the origin at the time $t = 0$ in both frames. Now consider the electric flux of the *static* electric field through the spherical "cap" of a sphere of radius r whose boundary is a circle parallel to the x-y plane at a height z above the origin. This is simple to evaluate:

$$\int_{\text{disk}} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA = \frac{q}{4\pi\epsilon_0\, r^2} \int_0^\theta \int_0^{2\pi} r^2 \sin\theta d\theta d\phi \tag{9.64}$$

where

$$\theta = \cos^{-1}\frac{z}{r} \tag{9.65}$$

[128]The Physics Teacher 51, 542 (2013)

is the angle that defines the curve C that bounds the cap on the sphere of radius r. This is easy to evaluate. The r^2 cancels, we get 2π from the $d\phi$ integral, and the remaining theta integral yields

$$\phi_e = (1 - \cos\theta)\frac{q}{2\epsilon_0} = (1 - \frac{z}{r})\frac{q}{2\epsilon_0} \tag{9.66}$$

A quick check: If $\theta_0 \to \pi$, we get the flux through the entire sphere, which is (correctly, according to GLE) q/ϵ_0.

Now let's write down Ampere's Law with the MDC, evaluated only for the electric flux through this cap:

$$\oint \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} = B_\phi 2\pi r \sin\theta = \mu_0\epsilon_0 \frac{d}{dt}\left\{(1 - \cos\theta)\frac{q}{2\epsilon_0}\right\} \tag{9.67}$$

We get no contribution from the 1, ϵ_0 cancels, and (to ring in a v) we can use the *chain rule*:

$$\frac{d\cos\theta}{dt} = \frac{d\cos\theta}{dz}\frac{dz}{dt} = \frac{d\cos\theta}{dz}v \tag{9.68}$$

All that remains before putting it back together is to evaluate $\frac{d\cos\theta}{dz}$. First, let's let $a = r\sin\theta$. Since we are not letting the actual curve C change in time, *this is a constant!* In terms of this:

$$\cos\theta = \frac{z}{r} = \frac{z}{(a^2 + z^2)^{\frac{1}{2}}} \tag{9.69}$$

The derivative with respect to z (holding a constant) is straightfoward to evaluate:

$$\frac{d}{dz}\frac{z}{(a^2 + z^2)^{\frac{1}{2}}} = \frac{1}{r} - \frac{z^2}{r^3} = \frac{(a^2 + z^2) - z^2}{r^3} = \frac{a^2}{r^3} \tag{9.70}$$

Now we can plug this in and reassemble the parts. Ampere's Law becomes:

$$\oint \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} = B_\phi 2\pi a = \frac{\mu_0}{2}\frac{qva^2}{r^3} \tag{9.71}$$

If we divide out the $2\pi a$, and do some renaming, we find that:

$$B_\phi = \frac{\mu_0}{4\pi}\frac{qvr\sin\theta}{r^3} = k_m\frac{|q\vec{\boldsymbol{v}} \times \vec{\boldsymbol{r}}|}{r^3} \tag{9.72}$$

which is *exactly our "Biot-Savart Law" for a point particle* with *exactly the usual right-hand rule determining direction!* We can then easily coarse-grain average this over (say) a chunk of conducting wire to find that the field in a short length $d\vec{\boldsymbol{\ell}}$ of wire (pointing in the direction of the positive charge carrier current):

$$d\vec{\boldsymbol{B}} = k_m\frac{(nqv_dA)d\vec{\boldsymbol{\ell}} \times \vec{\boldsymbol{r}}}{r^3} = k_m\frac{I(d\vec{\boldsymbol{\ell}} \times \hat{\boldsymbol{r}})}{r^2} \tag{9.73}$$

our usual Biot-Savart Law. We thus see that the Biot-Savart Law is really just ***Ampere's Law in Disguise***, valid in the *non-relativistic* (that is, $v_d \ll c$), *coarse-grained average limit!*

We had it exactly *backwards* in chapter 7! Rather than starting with Biot-Savart and then handwaving to get Ampere, we merely needed to write down Ampere correctly to be able to derive both the field of a point charge in the non-relativistic limit and the associated Biot-Savart Law. Note that the Biot-Savart Law is *not* a replacement for Ampere's Law and the

Maxwell Displacement current – it doesn't "know what do" with regions of space where charge is *accumulating* or where *there are no charges or currents* any more than the broken version of Ampere's Law did.

Still, it's pretty cool to see how amazingly well this idea of *invariance* works. Simply *changing frames* from a frame where there was only an electrostatic field causes a magnetic field to magically appear! Changing frames from a frame where there is only a magnetic field causes an electric field to magically appear! Clearly magnetic and electric fields must both be components of some higher order "superfield"[129] where frame changes alter the components, and in fact that is precisely the conclusion of the theory of special relativity, as you might learn in future Electrodynamics course or by following the provided link and doing a bit of link-hopping.

So what's (still) wrong with this? Why *must* we insist on $v \ll c$ (and, really, assume that r is "small enough")? The problem is that Faraday's Law and Ampere's Law, working together, permit a peculiar kind of *electromagnetic* field that *propagates like a wave* through regions of space that are otherwise free of charge. Indeed, *all changes* to the electromagnetic field observed at a point due to the motion of a charge only arrive at the point of observation after a time $t = r/c$, where c is the speed of light! The magnetic field we calculated above is correct enough as far as it goes, but only if one uses $\vec{r}$ *at the time in the past* when the electromagnetic field given off by the source charge would arrive at the point of observation in the present. This is called *retardation* – we are always "seeing" the electromagnetic field corresponding to sources that are time-of-travel in the past, just as we only hear explosions after the speed of sound carries the boom to our ears.

The Earth is around 15×10^7 kilometers away from the Sun. The speed of light is 3×10^5 kilometers per second. If the Sun were to go supernova and explode at the instant you read these words, it would be:

$$\Delta t = \frac{15 \times 10^7}{3 \times 10^5} = 500 \text{ seconds} \tag{9.74}$$

or a bit over eight *minutes* before you would – very briefly – *see* it before being vaporized (along with the rest of the Earth) by the burst of incredibly intense light propagating out from the blast. Our equations above left this retardation factor out, but a more precise treatment that leaves it in yields an entire system of electromagnetism that is completely invariant under *all* inertial reference frame changes. And that is *very* cool indeed!

First things first, of course, Our next step – which may be your first step, if you opted to omit this "advanced" sectioun – is to take Maxwell's Equations in all of their mostly complete (except for hypothetical magnetic monopoles) glory, and learn about this propagating electromagnetic wave that we know by a different name – *light!*

[129] Wikipedia: http://www.wikipedia.org/wiki/Electromagnetic_tensor#Relationship_with_the_classical_fields. Called the "second rank field strength tensor in four dimensions" in relativity theory, with symbol $F^{\mu\nu}$, if you care to look it up.

9.3: Maxwell's Equations for the Electromagnetic Field: The Wave Equation

OK, so let's rewrite the complete set of Maxwell's Equations, but this time *with* Maxwell's teensy weensy little contribution and see if we can figure out why it is so all-fired important that physicists speak in hushed tones when they mention Maxwell's name, much as they do for Newton and Einstein and a handful of others:

$$\oint_S \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA = \frac{1}{\epsilon_0} \int_{V/S} \rho_e dV \tag{9.75}$$

$$\oint_S \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA = \mu_0 \int_{V/S} \rho_m dV = 0 \tag{9.76}$$

$$\oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} = \mu_0 \left(\int_{S/C} \vec{\boldsymbol{J}} \cdot \hat{\boldsymbol{n}} dA + \epsilon_0 \frac{d}{dt} \int_{S/C} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA \right) \tag{9.77}$$

$$\oint_C \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{\ell}} = -\frac{d}{dt} \int_{S/C} \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA \tag{9.78}$$

The *symmetry* will now be a apparent if I put all of the terms involving *charges* as *sources* of the fields on the right and all of the terms involving the fields themselves on the left:

$$\oint_S \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA = \frac{1}{\epsilon_0} \int_{V/S} \rho_e dV \tag{9.79}$$

$$\oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} - \mu_0 \epsilon_0 \frac{d}{dt} \int_{S/C} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA = \mu_0 \int_{S/C} \vec{\boldsymbol{J}}_e \cdot \hat{\boldsymbol{n}} dA \tag{9.80}$$

$$\oint_S \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA = 0 \tag{9.81}$$

$$\oint_C \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{\ell}} + \frac{d}{dt} \int_{S/C} \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA = 0 \tag{9.82}$$

The *only* asymmetry now arises from the empirical non-observation of magnetic monopoles, and even you, humble beginning physics student that you are, can already see exactly what we would have to do to "fix" Maxwell's Equations if tomorrow somebody performed a reproducible experiment that discovered them.

But this symmetry isn't (yet) why Maxwell is cool. No, there is something much more profound buried in these equations now. Faraday's Law already showed us that changing magnetic fields make electric fields. Maxwell showed us that *at the same time*, changing electric fields make magnetic fields! Why is this significant? Because a changing electric field can make a changing magnetic field that makes a changing electric field that makes a changing magnetic field that makes – wait a minute! Is it possible that we could have an *electromagnetic wave?*

It is!

To see this is a bit tricky. It is tricky because we are taking an intro course where we have to avoid "real" differential multivariate calculus and the dread $\vec{\boldsymbol{\nabla}}$ differential operator. We have learned only the integral equation forms, which means basically that we have to convert them into derivatives in order to end up with a wave (differential) equation for the electric and magnetic field. Let's get to it.

9.3.1: The Wave Equation

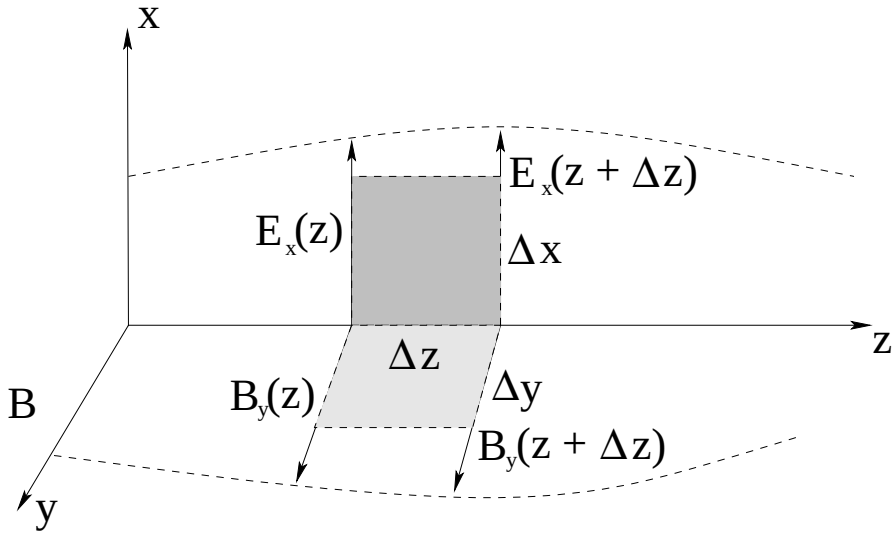


Figure 9.5: Two particular components of the electric and magnetic field, in a coordinate frame "far" from any sources and varying in space and time. The graph is a snapshot at a particular time t, but we can imagine that $E_x(z,t)$ and $B_y(z,t)$ generally and ignore any other variation with x or y for the moment.

Let us start, then, with *no* source terms in Maxwell's equations, or rather, in a region of space far from any sources. That doesn't mean that the fields there are zero, only that we don't have to worry about how the fields were originally produced – we know that they were somehow created by electric charges and currents but we don't care about the details. Maxwell's equations are then somewhat simpler:

$$\oint_S \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA = 0 \tag{9.83}$$

$$\oint_S \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA = 0 \tag{9.84}$$

$$\oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} = \mu_0 \epsilon_0 \frac{d}{dt} \int_{S/C} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA \tag{9.85}$$

$$\oint_C \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{\ell}} = -\frac{d}{dt} \int_{S/C} \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA \tag{9.86}$$

as now there are no magnetic *or* electric monopoles present, only the fields.

Let us graph the fields on an arbitrary coordinate system and apply Ampere's Law and Faraday's Law (only) to our graph. $\vec{\boldsymbol{E}}$ and $\vec{\boldsymbol{B}}$ have many components each, of course, and can be varying with respect to both position and time, so we need to simplify a bit to make sense of things. We will then imagine that either our distant source created only x-directed electric fields and y-directed magnetic fields or that, equivalently, we are only considering E_x and B_y components in particular of a more complicated field. Since the fields satisfy the superposition principle, any results we get for this pair of components can be generalized to any actual directions we like.

The graph is shown in figure 9.5, along with two dashed curves (bounding the shaded surfaces) to which we will apply Ampere's and Faraday's Laws. We will assume that $E_x(z,t)$ is

a function of z and t only – it may vary with respect to x or y as well, but for the moment we'll ignore any such variation[130]. Similarly we will assume $B_y(z,t)$ only. Our graph is a snapshot at some particular time t, so we don't bother writing t in on the figure (but it is really there). I'm sorry if it is a bit confusing to constantly ignore variation with respect to this or that variable – if/when you take multivariate calculus you'll learn once and for all how to deal with this sort of thing and encode it into the notion of the *partial derivative* but for the moment we're working our way towards a *result* that should be expressed in partial derivatives without actually using them or their (honestly, much simpler) notation.

Now let us apply Faraday's Law to the small differential loop in the x-z plane. This loop has an area $\Delta A = \Delta x \Delta z$, and we need to define a *right handed normal* to the loop in the y-direction (parallel to $\vec{\boldsymbol{B}}$). That means that we need to go around the loop *counterclockwise* as drawn in the page. Then:

$$\begin{aligned}
\oint \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{\ell}} &= -\frac{d}{dt}\int_{\Delta A} \vec{\boldsymbol{B}} \cdot \hat{\boldsymbol{n}} dA \\
0 \cdot \Delta z + E_x(z+\Delta z)\Delta x - 0 \cdot \Delta z - E_x(z)\Delta x &= -\frac{d}{dt}\left(B_y \Delta A\right) \\
\left(E_x(z+\Delta z) - E_x(z)\right)\Delta x &= -\frac{dB_y}{dt}\Delta x \Delta z \\
\frac{\left(E_x(z+\Delta z) - E_x(z)\right)}{\Delta z} &= -\frac{dB_y}{dt}
\end{aligned} \tag{9.87}$$

where we do the loop piecewise and get no contribution when we go in the z direction (because $\vec{\boldsymbol{E}}$ is in the x-direction perpendicular to z). If we take the limit $\Delta z \to 0$ of the left hand side this is just the *definition* of the derivative and we get[131]:

$$\frac{dE_x}{dz} = -\frac{dB_y}{dt}$$

Let's do exactly the same thing for Ampere's Law, this time using the more lightly shaded surface and curve in the y-z plane with area $\Delta A = \Delta y \Delta z$. Again we must go around it so that the right handed normal is parallel to $\vec{\boldsymbol{E}}$ in the x direction, or again counterclockwise as seen on the page from above. The *only* term on the right is the Maxwell Displacement Current – this is where Maxwell's contribution shines!

$$\begin{aligned}
\oint \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} &= \mu_0\epsilon_0 \frac{d}{dt}\int_{\Delta A} \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{n}} dA \\
B_y \Delta y + 0 \cdot \Delta z - B_y(z+\Delta z)\Delta y - 0 \cdot \Delta z &= \mu_0\epsilon_0 \frac{d}{dt}\left(E_x \Delta A\right) \\
-\left(B_y(z+\Delta z) - B_y(z)\right)\Delta y &= \mu_0\epsilon_0 \frac{dE_x}{dt}\Delta y \Delta z \\
\frac{\left(B_y(z+\Delta z) - B_y(z)\right)}{\Delta z} &= -\mu_0\epsilon_0 \frac{dE_x}{dt} \\
\frac{dB_y}{dz} &= -\mu_0\epsilon_0 \frac{dE_x}{dt}
\end{aligned}$$

[130] It isn't too difficult to imagine how such a field could be produced by (say) a distant oscillating electric dipole in the $-z$ direction, actually.

[131] Technically, this should be expressed as *partial* derivatives: $\frac{\partial E_x}{\partial z} = -\frac{\partial B_y}{\partial t}$, but since we cleverly arranged it so that E_x is a function of only one spatial coordinate and x and t are independent, it doesn't matter in this case.

where we have taken the limit $\Delta z \to 0$ as before in the last step[132].

Since we're going to use these two results a *lot*, let's write them down right next to each other:

$$\frac{dE_x}{dz} = -\frac{dB_y}{dt} \tag{9.88}$$

$$\frac{dB_y}{dz} = -\mu_0\epsilon_0\frac{dE_x}{dt} \tag{9.89}$$

Although they don't *look* much like it, these are both still Faraday's Law and Ampere's Law (with the MDC) respectively, although expressed only for two particular components of the electric and magnetic field.

Well, we could have had (say) a y-directed electric dipole instead, or (since our coordinate system was arbitrary) we could just rotate it by $\pi/2$ around the z axis to make E_x into E_y and B_y into $-B_x$ in the new coordinate system (imagine lifting the y-axis *up* and push x-back into the page as you mentally rotate figure 9.5). In that case one expects to get:

$$\frac{dE_y}{dz} = \frac{dB_x}{dt} \tag{9.90}$$

$$\frac{dB_x}{dz} = \mu_0\epsilon_0\frac{dE_y}{dt} \tag{9.91}$$

from an identical argument to the one above, something you can verify by completely recapitulating the derivation above as part of your homework[133].

This is all very well, but so far it is still not spectacular. To make it spectacular, we (say) differentiate the first of these equations with respect to z:

$$\frac{d}{dz}\frac{dE_x}{dz} = \frac{d^2E_x}{dz^2} = -\frac{d}{dz}\frac{dB_y}{dt} = -\frac{d}{dt}\frac{dB_y}{dz} \tag{9.92}$$

If we substitute the second equation in for the last term, we get:

$$\frac{d^2E_x}{dz^2} = -\frac{d}{dt}\mu_0\epsilon_0\frac{dE_x}{dt} = \mu_0\epsilon_0\frac{d^2E_x}{dt^2} \tag{9.93}$$

or

$$\frac{d^2E_x}{dz^2} - \mu_0\epsilon_0\frac{d^2E_x}{dt^2} = 0 \tag{9.94}$$

[132]Once again, this should be $\frac{\partial B_y}{\partial z} = -\mu_0\epsilon_0\frac{\partial E_x}{\partial t}$, but in this one dimensional, non-relativistic treatment it doesn't matter.

[133]Sure, sure, they should all be partials. In fact, you are basically deriving:

$$\vec{\nabla}\times\vec{E} = -\frac{\partial\vec{B}}{\partial t}$$

$$\vec{\nabla}\times\vec{B} = \mu_0\epsilon_0\frac{\partial\vec{E}}{\partial t},$$

the grown-up way of writing the source free Faraday's and Ampere's Laws in terms of the *curl*, a component pair at a time. You can actually get all six terms in these two equations from our one original result by mentally rotating the arbitrary right-handed coordinate system into all six indepedent orientations. Or you can use Stokes Theorem, which we basically just derived. Since advanced students derived the partial differential form for Gauss's Law in the second week, we have now derived the partial differential form for the whole set of Maxwell's Equations, at least once the source terms are put back in...

We stare at this for a moment, our brains dulled by too much algebra. Then, through the fog, a *light* begins to shine through, dim at first, then ever brighter until it rivals the sun! Holy Smoke, Batman, haven't we seen that equation, or one sort of like it, before?

We *have!* In the first part of the course we went to considerable (although much less) pains to derive the one-dimensional *wave equation for a string*:

$$\frac{d^2y(x,t)}{dx^2} - \frac{1}{v^2}\frac{d^2y(x,t)}{dt^2} = 0 \tag{9.95}$$

for a y-displaced string, where the wave propagated at speed v in the $\pm x$ direction! Well, it seems that Maxwell's Equations tell us that the x-component of the electric field in a region of space far from any sources satisfies a wave equation too! I wonder (you ask yourself) what the *speed* of this wave is?

Well, comparing the two equations, we see that:

$$v^2 = \frac{1}{\mu_0\epsilon_0} = \frac{4\pi}{\mu_0 4\pi\epsilon_0} = \frac{4\pi}{\mu_0}\frac{1}{4\pi\epsilon_0} = \frac{k_e}{k_m} \tag{9.96}$$

and if we do only a *tiny* bit of arithmetic with the only two constants I really required you to memorize/learn for this part of the class we get:

$$v^2 = \frac{9\times 10^9}{10^{-7}} = 9\times 10^{16} \quad \frac{\text{meters}^2}{\text{second}^2} \tag{9.97}$$

or:

$$v = c = 3\times 10^8 \quad \frac{\text{meters}}{\text{second}}. \tag{9.98}$$

This particular speed was first estimated during the very first days of systematic scientific exploration based on observations of variations in the period of one of Jupiter's moons. It was known within a few percent by the mid-1800s, and experiments were being done that were rapidly adding significant digits to the quantity (it is currently one of the most accurately known physical constants). This quantity is the *speed of light*.

The electric field wave propagates at the ***speed of light!***

And *that*, boys and girls, is why Maxwell got his name on the *whole set* of Maxwell's Equations for his one measely term. He proposed (correctly) that ***light is an electromagnetic wave*** and in so doing, transformed the still partially disparate electric and magnetic fields into a single unified field theory and revolutionized our understanding of, well, *everything*. You. Me. Stuff. What *isn't* made up of electric charges and *doesn't* interact via the electromagnetic interaction[134]?

Well, we haven't quite shown *all* of that yet. But now you can see how it goes well enough to complete most of what we still need to do even without my help. If we take the *second* of the two equations (Ampere's Law) and differentiate both sides with respect to z and substitute in the first (Faraday's Law) for the right hand side we get:

$$\frac{d^2B_y}{dz^2} - \mu_0\epsilon_0\frac{d^2B_y}{dt^2} = \frac{d^2B_y}{dz^2} - \frac{1}{c^2}\frac{d^2B_y}{dt^2} = 0 \tag{9.99}$$

[134]The correct answer: ***Apparently*** some 85% of the massive part of the Universe, maybe, possibly – so-called Wikipedia: http://www.wikipedia.org/wiki/Dark Matter. But we can't see it because, well, it's *dark!*

for example (you should verify this, obviously, by *doing* it). So yes, $B_y(z,t)$ is also a wave that propagates at the speed of light c. The two components were presented together because they are *coupled* by Ampere's and Faraday's Laws. The variation of E_x in space and time produces the variation of B_y in space and time, so that either one propagates like a wave, but the waves are not independent. Similarly, E_y and B_x are coupled as they vary along the z axis in time, and obviously they satisfy the same wave equation and propagate at the same speed as well.

The rest of the course is basically devoted to understanding light as an electromagnetic wave. Although we will restrict ourselves to "one dimensional" wave forms, we will talk a bit about how light varies with distance as it spreads out in three dimensions from a central source. We will think at least a bit about sources, relying heavily on the oscillating electric dipole as a model source. As a source, the dipole has one ideal feature: It is a *harmonic* source. Consequently, although light in general does not *have* to be harmonic, we will find it very convenient to focus on understanding it as a harmonic wave[135].

9.4: Light as a Harmonic Wave

Before we study light as a harmonic wave, let's very quickly recapitulate things we know – or *should* know – about waves based on our study of waves on a string and sound waves in the first (Mechanics) part of the course, assuming you used my/this textbook for it. Recall that we showed that a very general solution to the wave equation for waves on a string was:

$$y(x,t) = f(x \pm vt) \tag{9.100}$$

where $f(u)$ is an arbitrary one-dimensional function. Basically *any* functional form that propagates to the right or left along the x-axis (the string axis) was a solution to the wave equation.

Since the electric and magnetic fields both satisfy one-dimensional wave equations for propagation along the z-axis, we can expect this to be true for them as well. Any electric field that we can create that has some shape at time $t = 0$ can be made to propagate in the $\pm z$ direction by pairing it with the appropriate magnetic field. However, *most* of those arbitrary shapes are going to be very difficult to arrange, and arranging them to occur with their correctly paired partner field even more difficult, and then – we really should be working with a *3 dimensional* wave equation for the *coupled vector fields*. We will thus *ignore* this *general* "one dimensional" solution and concentrate on a much more specific one, one tied to a particular easy-to-imagine source.

Suppose the source of the wave we observe is indeed an oscillating electric dipole located at the (distant) origin and aligned with the x-axis. Then we know that at any given instant in time, if the dipole points up in the $+x$ direction, its field curls around and points down in the $-x$ direction as it passes through the z-axis. At least, this was our *static* result. Now, however,

[135] Even when we treat light as a non-harmonic wave, we usually begin by transforming e.g. the initial conditions or boundary conditions into the harmonic/frequency/wavenumber domain, solve the problem for harmonic waves, and then use the *Fourier transform* to transform back and obtain the general non-harmonic result. Of course this once again requires more math to pursue. Physics majors, do you get the idea that you will need more math, sooner or later? Math majors, do you see why you need to take more physics? Everybody else, aren't you glad you *don't* need to in order to pretty much understand light waves perfectly well?

we see that this result can't quite be correct. If the electric field propagates at speed c and the dipole is *oscillating*, the field itself has to oscillate too, and furthermore the "up" regions have to move away from the source at c, as do the "down" regions. In other words, at any specific distance z (much greater than the size of the physical dipole) we'd expect the field to have the form of a *harmonic wave*:

$$E_x(z,t) = E_{0x}\sin(kz \pm \omega t) \tag{9.101}$$

where ω is the frequency of the oscillating dipole source that is producing the wave and E_{0x} *might* have a bit of leftover z-dependence as we know the dipole fields themselves must get weaker farther from the source[136].

We are fortunate in that this actually *is* a function of the form $f(z \pm vt)$! To see this, let's factor the argument:

$$E_x(z,t) = E_{0x}\sin(k(z \pm \frac{\omega}{k}t) = E_{0x}\sin\left(k(z \pm ct)\right) \tag{9.102}$$

which has the desired form if $c = \omega/k$. Indeed, if you substitute this harmonic wave into the wave equation, you get:

$$\begin{aligned} \frac{d^2}{dz^2}E_{0x}\sin(kz \pm \omega t) = -k^2 E_{0x}\sin(kz \pm \omega t) &= \frac{1}{c^2}\frac{d^2}{dt^2}E_{0x}\sin(kz \pm \omega t) \\ &= -\frac{1}{c^2}\omega^2 E_{0x}\sin(kz \pm \omega t) \end{aligned} \tag{9.103}$$

or (dividing out)

$$c^2 = \frac{\omega^2}{k^2} \tag{9.104}$$

and $c = \omega/k$ as promised.

Again recalling our work with harmonic waves, we expect that in these equations:

$$k = \frac{2\pi}{\lambda} \tag{9.105}$$

is the *wave number* of the wave, the "spatial angular frequency" in terms of the *wavelength* of the wave λ, just as:

$$\omega = \frac{2\pi}{T} \tag{9.106}$$

is the *temporal* angular frequency of the wave in terms of its period T. Thus:

$$c = \frac{\omega}{k} = \frac{2\pi}{T}\frac{\lambda}{2\pi} = \frac{\lambda}{T} = f\lambda \tag{9.107}$$

are *all useful ways of relating the frequency, wavelength, angular frequency, wave number, period, and speed* of the wave. Yes, you can remember just one of these and figure out the rest, but on an exam *speed* counts and I recommend learning *all* of these forms so that they are second nature and you don't have to think about them.

[136]Note well that we could have equally well used $E_{0x}\cos(kz \pm \omega t + \phi)$ for some arbitrary phase angle ϕ, or better yet $E_{0x}e^{ikz}e^{\pm i\omega t}$ where $E_{0x} = |E_{0x}|e^{i\phi}$ is an arbitrary complex amplitude. We choose to use $\sin(kz \pm \omega t)$ for no other reason than to have something specific to work with, but these all satisfy the wave equation and are equally valid possibilities. The phase angle ϕ in particular corresponds to determining simply the shape of the wave when we start the "clock" of our harmonic wave in our particular reference frame.

We expect that:

$$B_y(z,t) = B_{0y}\sin(kz \pm \omega t + \phi) \tag{9.108}$$

where we cannot yet assume that E_x and B_y have the same phase, although we do insist (since they are parts of the same wave) that they have the same frequency. Now let's work some magic. We'll restrict our interest for the moment to a wave propagating to the *right* ($+z$):

$$E_x(z,t) = E_{0x}\sin(kz - \omega t) \tag{9.109}$$

$$B_y(z,t) = B_{0y}\sin(kz - \omega t + \phi) \tag{9.110}$$

We substitute these two forms into (your choice of) Ampere's or Faraday's Law in differential form. Let's choose Faraday as being marginally simpler:

$$\frac{d}{dz}E_{0x}\sin(kz - \omega t) = -\frac{d}{dt}B_{0y}\sin(kz - \omega t + \phi) \tag{9.111}$$

$$kE_{0x}\cos(kz - \omega t) = \omega B_{0y}\cos(kz - \omega t + \phi) \tag{9.112}$$

$$E_{0x}\cos(kz - \omega t) = \frac{\omega}{k}B_{0y}\cos(kz - \omega t + \phi) \tag{9.113}$$

$$E_{0x}\cos(kz - \omega t) = cB_{0y}\cos(kz - \omega t + \phi) \tag{9.114}$$

In order for this to be true, $\phi = 0$ – the electric and magnetic fields *do* have to have the same phase (and frequency and wavelength) and we have now *proven* this, and:

$$E_{0x} = cB_{0y} \tag{9.115}$$

The electric and magnetic fields are not independent! The magnitude, phase, and frequency of one is determined completely by the other.

This is a wave propagating to the *right*, as noted. Let's try the exact same solution for the independent solution:

$$E_y(z,t) = E_{0y}\sin(kz - \omega t) \tag{9.116}$$

$$B_x(z,t) = B_{0x}\sin(kz - \omega t + \phi) \tag{9.117}$$

Note that we have assumed nothing other than E_y is coupled to B_x (because that's what Ampere/Faraday tell us). Again we substitute – using the form of Faraday's Law we derived for E_y – and get:

$$\frac{d}{dz}E_{0y}\sin(kz - \omega t) = \frac{d}{dt}B_{0x}\sin(kz - \omega t + \phi) \tag{9.118}$$

$$kE_{0y}\cos(kz - \omega t) = -\omega B_{0x}\cos(kz - \omega t + \phi) \tag{9.119}$$

$$E_{0y}\cos(kz - \omega t) = -\frac{\omega}{k}B_{0x}\cos(kz - \omega t + \phi) \tag{9.120}$$

$$E_{0y}\cos(kz - \omega t) = -cB_{0x}\cos(kz - \omega t + \phi) \tag{9.121}$$

This time we see that the two fields must be in phase and that:

$$E_{0y} = -cB_{0x} \tag{9.122}$$

For a wave propagating to the right, *both* of the independent components of $\vec{\boldsymbol{E}}$ are related to the coupled components of $\vec{\boldsymbol{B}}$ such that:

$$|\vec{\boldsymbol{E}}| = c|\vec{\boldsymbol{B}}| \tag{9.123}$$

and so that the E-field *crossed into* the B-field – $\vec{\boldsymbol{E}} \times \vec{\boldsymbol{B}}$ – points in the direction of the wave's propagation. That is, if we let the fingers of our right hand line up with $\vec{\boldsymbol{E}}$ and curl them into $\vec{\boldsymbol{B}}$, our thumb points in the direction of propagation. The right hand rule $\vec{\boldsymbol{E}} \times \vec{\boldsymbol{B}}$ also works for waves propagating in the $-z$/left direction, e.g. $E_{0x}\sin(kz + \omega t)$ (try it!).

9.5: The Poynting Vector

OK, so now we have the harmonic electric and magnetic field, and both are in phase and have amplitudes related by c. We know that there is some *energy* in these fields described by the *energy density* of the electric and magnetic fields respectively:

$$\eta_e = \frac{1}{2}\epsilon_0 E^2 \tag{9.124}$$

$$\eta_m = \frac{1}{2\mu_0}B^2 \tag{9.125}$$

For electromagnetic waves, however, that energy isn't sitting still. It is *moving*, being carried by the wave from one point to another!

We can easily see that energy must be carried by the wave by imagining a source that is turned on (the dipole moment is pulled out and released to oscillate, if you like) at time $t = 0$. Some distance away from the source at first there is no field – our "Lorentz model" atom was originally spherically symmetric and produced no field until polarized – and then the field *reaches* it some time after the dipole is excited and starts to oscillate. No electromagnetic field energy in that region of space before the field reaches it, *yes* electromagnetic field energy after the field reaches it, therefore ***energy is carried by the field from the source to the region of space***. Simple!

Naturally, we'd like to be able to compute how *much* energy is being carried along by the field. To find out, we resort to what should now be a very familiar argument. In a time Δt, all of the energy in a box of length $c\Delta t$ will be carried through the cross-sectional area A of it's end. The amount of energy is:

$$\Delta U = \frac{1}{2}\left(\epsilon_0 E^2 + \frac{1}{\mu_0}B^2\right)c\Delta t A \tag{9.126}$$

The power per unit area per unit time that is carried through A is a quantity we define to be the *intensity* of the light wave:

$$I = \frac{P}{A} = \frac{\Delta U}{A\Delta t} = \frac{1}{2}\left(\epsilon_0 E^2 + \frac{1}{\mu_0}B^2\right)c \tag{9.127}$$

Let's do a bit of algebra. For the moment, let's once again concentrate on our familiar harmonic pair $E_x(z,t)$ and $B_y(z,t)$ (omitting the arguments (z,t) for simplicity). Note that We can now freely substitute

$$E_x = cB_y \quad \Leftrightarrow \quad B_y = \frac{E_x}{c}$$

obtained above and use our new expression for the speed of light to get the following (extremely useful) result:

$$c^2 = \frac{1}{\epsilon_0\mu_0} \quad \Rightarrow \quad \frac{1}{\mu_0}B_y^2 = \frac{1}{\mu_0}\frac{E_x^2}{c^2} = \frac{\epsilon_0\cancel{\mu_0}}{\cancel{\mu_0}}E_x^2 = \epsilon_0 E_x^2 \tag{9.128}$$

In words, ***the energy density of the electric field equals the energy density of the magnetic field in a propagating electromagnetic wave!***

This lets us "instantly" simplify our expression for I above:

$$I = \frac{1}{2}\left(\epsilon_0 E^2 + \frac{1}{\mu_0}B^2\right)c = \epsilon_0 c E_x^2 = \epsilon_0 c^2 E_x B_y = \frac{1}{\mu_0}E_x B_y \tag{9.129}$$

An identical expression holds, of course, for the other two components of the electromagnetic field propagating in the z direction (or *any* direction) and we can identify this direction as the direction of $\vec{\boldsymbol{E}} \times \vec{\boldsymbol{B}}$. This motivates us to turn the intensity into a *vector* pointing in this direction. Indeed, we have just derived:

$$\vec{\boldsymbol{S}} = \frac{1}{\mu_0}\vec{\boldsymbol{E}} \times \vec{\boldsymbol{B}} \tag{9.130}$$

– a quantity eponymously named the *Poynting vector* (yes, it *poynts* in the direction that the wave propagates, har har, although it is actually named after John Henry Poynting, its original discoverer).

We derived this expression for the "special" case of a surface ΔA that is perpendicular to the direction of propagation. By now we should easily be able to see that if we tip this surface into $\Delta A'$ at some angle θ, we will increase its area by $1/\cos(\theta)$ and will need to compensate by multiplying it by $\cos(\theta)$. This makes the power through the surface not just $P = I\Delta A$ but $P = I\Delta A' \cos(\theta)$. This, of course, makes the power delivered through *any* surface S the *flux of the Poynting vector* through that surface:

$$P = \int_S \vec{\boldsymbol{S}} \cdot \hat{\boldsymbol{n}}\, dA \tag{9.131}$$

This is *general* – the electromagnetic wave doesn't have to propagate in the z direction. Indeed, as we shall see, it doesn't even have to be an electromagnetic *wave* – crossed $\vec{\boldsymbol{E}}$ and $\vec{\boldsymbol{B}}$ fields always carry power even if they are static! Our discovery above is just the field part of a general result in electrodynamics called **Poynting's Theorem**[137] that is effectively the field energy transmission component of the general work-energy theorem for a system of electrical charges and currents! The field intensity defined above is the *magnitude of the Poynting vector*:

$$I = |\vec{\boldsymbol{S}}(z,t)| = \frac{1}{\mu_0}\left|\vec{\boldsymbol{E}} \times \vec{\boldsymbol{B}}\right| \tag{9.132}$$

In your homework, you will show that the flux of the Poynting vector precisely describes the way power is delivered to resistors, capacitors, and inductors, just as we have already seen that the potential energy stored in (say) a capacitor is ultimately the integral of the electric energy density. For now, though, we will concentrate on just two coupled field components, e.g. E_{0x} and B_{0y} of a *harmonic traveling wave* propagating in the z direction. In this case the instantaneous intensity is:

$$I(z,t) = \frac{1}{\mu_0}E_{0x}B_{0y}\sin^2(kz - \omega t) = \frac{1}{\mu_0 c}E_{0x}^2 \sin^2(kz - \omega t) \tag{9.133}$$

As we will see in the next chapter, when dealing with *visible light* especially the frequencies are likely to be *enormous* – order of 10^{15} hertz! In this case it makes the most sense to only

[137] Wikipedia: http://www.wikipedia.org/wiki/Poynting's_theorem.

consider the *time average intensity*. Remember, the average of any harmonic function squared is $1/2$! For monochromatic harmonic traveling waves, then, we will usually write:

$$\boxed{I = \frac{1}{2}\epsilon_0 c E^2} \tag{9.134}$$

where E is the *magnitude* of the harmonic electric field vector perpendicular to the direction of propagation. Note that for this kind of use, we don't actually indicate explicitly that the intensity is time-averaged; we just don't include any explicit time dependence as the leading factor of $1/2$ says it all. To turn this into a time-averaged Poynting vector, we just multiply it by a unit vector pointing in the direction of the wave's propagation.

As you can hopefully see, the Poynting vector is pretty much magical[138]. It works equally well for steady state or slowly varying electromagnetic fields as it does for harmonic electromagnetic standing or traveling waves.

Since the electromagnetic field can carry energy – essentially transporting the work required to create the propagating field in one place to a *different*, quite possibly *very distant* place[139] where it turns back into work once again – it *must* carry some *momentum*[140] as well, as force is the time rate of change of the momentum and one cannot do work without a force!

If you recall our arguments back when we discussed the failure of Newton's Third Law in the context of magnetic fields and forces, we knew even then that it must be so – the missing momentum and energy described then has to go *someplace* or momentum violation would be ubiquitous in electromagnetism! It is time to run this down.

This is actually rather tricky to do. It isn't easy to derive the momentum carried by the electromagnetic field, because ***the electromagnetic field has no mass***, and our original definition of momentum of at least a massive particle was $\vec{p} = m\vec{v}$. Clearly we can't just let m go to zero in this expression, so we have to find a new relationship between e.g. field energy or the Poynting vector and momentum, one that doesn't rely on mass. The easiest way to see what it must be is to examine the net force exerted on charges in the electromagnetic field of a harmonic traveling wave. We'll do this (and define the associated idea of *radiation pressure*) in the next section.

[138]Or, perhaps, the *opposite* of magical – a fundamental and critical component of a consistent theory of electrodynamics. Maybe the word I'm reaching for is astounding, or awesome, or amazing...

[139]Star light, star bright, first star I see tonight... work done by charges many *trillions* of kilometers away is routinely delivered to charges in the rhodopsins in the rods and cones in your retina!

[140]And often angular momentum as well, but that is beyond the scope of this course.

9.6: Radiation Pressure and Momentum

Although it is beyond the scope of this course to *properly* derive the electromagnetic stress-energy tensor that consistently incorporates *both* Poynting's theorem *and* momentum transfer by the field[141], there are at least a couple of elementary arguments that we can use to motivate the results that matter at the introductory level.

The first argument is that of *dimensional scaling*. The fundamental SI units of momentum and energy are clearly:

$$\text{Energy} = \frac{\text{kilogram-meter}^2}{\text{second}^2} \qquad \text{Momentum} = \frac{\text{kilogram-meter}}{\text{second}}$$

They differ by a velocity. In fact:

$$\text{Velocity} = \frac{\text{meter}}{\text{second}} \quad \Rightarrow \quad \text{Momentum} = \frac{\text{Energy}}{\text{Velocity}}$$

We've already seen that – massless or not – electromagnetic fields contain energy. We also have seen that electromagnetic fields – harmonic or not – can *propagate* at only one speed in free space – that of light! We can then at least guess on dimensional grounds that the magnitude of the momentum in an electromagnetic wave might be given by:

$$p = \frac{\mathrm{E}}{c}$$

(where E is energy, sorry, not electric field strength for the moment). This leaves us a couple of things to resolve, though. For one thing, we don't have an expression for "the energy in the electromagnetic field", we have an expression for the energy *density* of the electromagnetic field. For a harmonic traveling wave in the z direction represented by E_x and B_y , then, we might expect something like:

$$\frac{\Delta p}{\Delta \mathcal{V}} = \frac{1}{c}\frac{1}{\mu_0}B_y^2 = \frac{1}{c^2}\frac{1}{\mu_0}E_x B_y = \frac{1}{c^2}|\vec{\boldsymbol{S}}| \tag{9.135}$$

(where Δp is the magnitude of the momentum in a small/differential volume $\Delta\mathcal{V}$) to represent the momentum *density* in terms of the energy *density*, expressed in terms of the Poynting vector's magnitude for these two coupled field components. This *does* leave us with the problem of assigning the momentum density a *direction* (as a vector quantity) but there is at really only one choice that seems reasonable – in the direction of propagation of the wave, that is, in the same direction of the Poynting vector!

Let us, then, define the *momentum density* of an electromagnetic waves as:

$$\vec{\boldsymbol{g}} = \frac{d\vec{\boldsymbol{p}}}{d\mathcal{V}} = \frac{1}{c^2}\vec{\boldsymbol{S}} \tag{9.136}$$

This makes dimensional sense. If light carries energy and momentum "like a particle" but distributed in a volume of space, it is difficult to imagine some *other* velocity to use to convert one into the other or some *other* direction for the momentum.

[141] Way, way beyond. It's an experience that must wait for a future, more serious course on electrodynamics that includes special relativity and much more...

In fact, this expression is exactly correct. As we'll see below, as charged matter absorbs or reflects an electromagnetic wave, it will experience a vector force representing the rate of momentum transfer to the matter much the same way collisions between massive objects do, but without the incident field having any mass. But where does this force come from?

To offer at least *some* sort of plausible physical model for this force, consider a plane wave of light incident on an ideal/perfect conductor as illustrated in figure 9.6. In this figure we see a harmonic electromagnetic wave incident at right angles onto a conducting surface. The $\vec{E}$-field is oscillating in the plane of the page, and the coupled $\vec{B}$-field is oscillating in and out of the page as the wave propagates down to hit the surface.

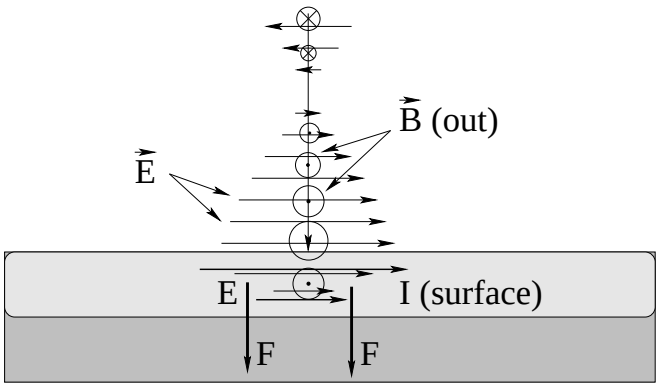


Figure 9.6: An electromagnetic wave incident on a conducting surface penetrates a short distance into the conductor, inducing a ***surface current*** in the direction of the electric field at the surface.

Now, it is a True Fact$_{\rm tm}$ that while conductors screen their *bulk* interior from electromagnetic fields (including electromagnetic radiation) they do not do this *instantly* at the surface. Just as static fields build up a static surface charge density a *few atoms thick* that cancels the field on the interior of the conductor, time varying fields penetrate a small distance into a conductor (a distance called the ***skin depth***) before being cancelled by an induced time-varying charge-current distribution confined to the surface. The skin depth depends on the frequency of the wave and the conductivity of the material (getting smaller as either one gets larger) but is usually at least a few atoms thick (and can be centimeters thick at very low frequencies such as that of household current).

The electrical field of this wave penetrates a short (grey-shaded) distance into the conductor before being attenuated, and within this distance the electric field pushes a ***surface current in the direction of the field*** as one expects from the relation $\vec{J} = \sigma\vec{E}$ (a form of Ohm's Law, recall, from our discussion of conduction and resistance). As this happens, the magnetic field *also* penetrates a short distance into the surface and exerts a force on this surface current. From our knowledge of the magnetic force exerted on an elecgtrical current and the right-hand rule, this force is expected to be ***in the direction of the wave's propagation*** and will be spread out on the entire conducting surface.

This simple picture demonstrates that just as the electromagnetic wave carries *energy* (per unit time) into the metal to be reradiated in a reflected wave and.or converted into heat by the resistance of the surface, it also carries *linear momentum* into the surface (per unit time) and exerts a force on that surface as it is absorbed or reflected. From our previous discussion

of dielectrics, which *also* develop a (bound) surface charge density that reduces the electric field, we expect a dielectric surface to also have a (much weaker) surface current parallel to the electric field and to still experience a force when impacted by an electromagnetic wave in direct proportion to the energy absorbed by the surface per unit time.

This transfer of momentum to the surface follows the same general rules we learned in the first half of this course when discussing momentum transfer by things like basketballs hitting a floor and bouncing off versus baseballs being caught. If any surface absorbs the energy transmitted by radiation, it also absorbs the momentum transmitted by the radiation (like a baseball being *caught* by an ice skater). If the surface reflects the energy of the radiation, it picks up ***twice*** the momentum transmitted by the radiation (less a small amount needed to balance energy and momentum simultaneously), like a baseball caught by an ice skater who then *throws it back* (almost) as fast as it was moving when it was caught. If an incident wave is reflected off at an angle, the force exerted on the surface comes only from the part of the wave's momentum that is *changed* by the implicit reflection, not the unchanged component parallel to the surface.

We will idealize these two rules and assume that absorption transfers exactly the momentum in the wave (per unit time) in the direction of the Poynting vector, and that reflection of a wave transfers twice the *component* of the momentum (per unit time) of the wave perpendicular to the surface.

The remaining question is how can we compute the force exerted by the wave on any given surface? We start with:

$$\vec{\boldsymbol{g}} = \frac{1}{c^2}\vec{\boldsymbol{S}} \tag{9.137}$$

obtained by vigorous hand-waving above (but correct for all of that) and reason with an argument that by now must be quite familiar (especially since we just used it in the previous section).

Again let us consider a simple plane (harmonic) wave incident at right angles onto a perfectly absorbing surface with area A. The magnitude of the momentum Δp transferred to the surface in a time Δt is *all of the momentum* in the box of volume

$$\Delta\mathcal{V} = Ac\Delta t$$

as usual. That is:

$$\Delta p = |\vec{\boldsymbol{g}}|\Delta\mathcal{V} = \frac{1}{c^2}|\vec{\boldsymbol{S}}|Ac\Delta t \tag{9.138}$$

If we divide both A and Δt to the left, we get the *force per unit area* exerted on the surface:

$$P_r = \frac{1}{A}\frac{\Delta p}{\Delta t} = \frac{|\vec{\boldsymbol{S}}|}{c} \tag{9.139}$$

This is called the ***radiation pressure***. Its SI units are exactly what one expects – pascals – although in this context it is probably best to think of as pascals as the natural units of *energy density*, joules per unit volume, as it describes the pressure in the *propagating electromagnetic field*.

Alternatively, one can leave the A behind to get just the vector *force*:

$$\vec{\boldsymbol{F}} = \frac{\Delta\vec{\boldsymbol{p}}}{\Delta t} = A\frac{\vec{\boldsymbol{S}}}{c} \tag{9.140}$$

which includes the *direction* of the momentum carried by the field in the wave, the Poynting vector direction itself.

It's a bit trickier than usual to considers a *tipped* surface (that still completely absorbs the wave). One has to compute the flux of the Poynting vector into the surface and reduce the effective force by the cosine of the angle of incidence. This much is straightforward – the wave has to pass through the "tipped window" presented by the surface to the incident wave. However, the result has to remain a *vector* in the direction of $\vec{\boldsymbol{S}}$! We can write this for a simple flat surface A tipped at the angle θ relative to the direction of incidence of a uniform plane wave as:

$$\vec{\boldsymbol{F}} = \frac{1}{c} A \cos\theta \vec{\boldsymbol{S}} \tag{9.141}$$

but we'd have to resort to tensor forms to compute the force exerted on a curved surface by an electromagnetic wave that varied significantly across that surface, especially one that is partially absorbed and partially reflected, *well* above our pay grade.

There is, however, one more case we can describe relatively cleanly along this same line – if a uniform plane wave is incident on a tipped *perfectly reflecting* surface, it exerts *twice* the force from the radiation pressure alone, but only along a line *perpendicular* to the surface, much like the homework problem involving beads bouncing on the pan of a balance in the Mechanics text. In this case we expect:

$$\vec{\boldsymbol{F}} = 2 \times \frac{1}{c} A \cos\theta \vec{\boldsymbol{S}} \cdot \hat{\boldsymbol{n}} \times \hat{\boldsymbol{n}} = \frac{2}{c} A \cos^2\theta |\vec{\boldsymbol{S}}| \hat{\boldsymbol{n}} \tag{9.142}$$

where $\hat{\boldsymbol{n}}$ is a normal unit vector pointing *in* to the surface in question.

In words, we compute the flux of the Poynting vector that actually strikes the surface A, introducing a factor of $\cos\theta$ to describe the tipped window of incidence. However, only the component of $\vec{\boldsymbol{S}}$ perpendicular to the surface itself is reflected/reversed (giving us the factor of 2) – the component of the momentum density of the incident wave parallel to the surface A is unchanged in the reflected wave. This gives us a *second* factor of $\cos\theta$. Finally, the direction of the transferred momentum is $\hat{\boldsymbol{n}}$, directly into the surface.

This result is an idealization as the reflected wave will always have *slightly* less energy density than the incident one if the surface itself is moving in the general direction of the incident wave and thereby gains energy from the wave, and will have slightly *greater* energy density if the surface is moving *towards* the source and hence *loses* energy to the wave[142].

These two cases, however simple and specific they might be, are well worth learning. Together, they help us understand how sunlight can drive the solar wind and how ***solar sails*** can be used to move – slowly – around in the solar system without the expenditure of fuel. Solar sails are actually being deployed as a way of prolonging the lifetime of satellites, whose orbits generally decay (due to e.g. drag forces from the very thin atmosphere through which they move).

Note well that the radiation pressure of visible light at intensities that wouldn't vaporize us instantly is *not large!* A one watt per square meter light source (think – a flashlight illuminating

[142]Or, you can think of the reflected wave being Doppler shifted by the motion of the reflector – in one case the reflected wavelength would be a bit longer in the reflected wave than in the incident one, so the energy is going to be more "diluted" in space; in the other the reflected wavelength is *shorter*, so there are more wavefronts per unit volume.

a square meter of wall) exerts a force of $\sim 3 \times 10^{-7}$ newtons on that same square meter if it is painted dead black and functions as a perfect absorber. Even direct sunlight at noon – call it 700 watts per square meter – is still only $\sim 2 \times 10^{-4}$ newtons – 200 *micro*newtons. This is why we simply never notice it – it is invisibly small compared to near-Earth gravitation, atmospheric pressure, drag forces, etc that act on our large masses. It's not like sunlight pushes you into the sand if you lie out in it to get a tan.

For solar sails to be able to move things around, they have to be *large* and *very, very light and shiny*. Fortunately, we can engineer e.g. mylar sheets that are indeed both! You'll get to explore the numbers in homework problems.

9.7: Sources of the Electromagnetic Field (Advanced/Optional)

One question that we failed to answer in our derivation of the wave equation for electromagnetic radiation above is this:

> Where the heck does the radiation *come from*?

At this point in the course, you should be able to see that it has to come from electric charges, because so far the *only* source for *either* electric *or* magnetic fields are electric charges, either stationary or moving.

"Radiation", in the form of a solution to the coupled wave equation for electric and magnetic fields we derived above, requires two more things: The electric and magnetic fields have to be ***changing in space and time*** and they have to be ***perpendicular to the direction the radiation propagates***. This strongly suggests that the charges that produce electromagnetic radiation have to be *moving*.

However, our derivation of the magnetic field produced by a uniformly moving point charge, plus a bit of reasoning, clearly shows that while a charge can produce both an electric and a magnetic field in any inertial reference frame in which it is moving, if we choose the *specific* inertial reference frame in which it is *not* moving, it produces only an electrostatic field. Without a magnetic field at all (in this frame) there can be no directed radiation of energy (that is, the Poynting vector associated with the fields produced by the charge is zero if the magnetic field is zero). If there is no energy radiated in one inertial frame, there can be no energy radiated in any other inertial frame, so we conclude that ***uniformly moving charges do not radiate***. This is essentially the statement that both the kinetic energy and momentum of an isolated, uniformly moving charge must be conserved.

Well, what about *accelerating* charges, charges whose magnetic field *cannot* be made to generally vanish in *any* inertial reference frame? As we will see below, the source of electromagnetic radiation is ***accelerating charge!*** Before we attempt to derive an explicit form for the radiation emitted by a uniformly accelerated charge, however, we have to add one key concept to our earlier treatment of the electrostatic field. In the first three chapters, we only considered the fields produced by more or less stationary charge (density) distributions. In our treatment of magnetostatics, we similarly only considered the fields produced by uniformly moving charge/current densities. In both of these treatments, we could determine the electric and magnetic fields "over all space".

When we added Faraday's Law and Ampere's Law with the Maxwell Displacement Current, however, things *changed*. In particular, we proved above that electromagnetic radiation must propagate in a vacuum at the speed of light, c, determined from the electric and magnetic constants.

We will now extend this just a tiny bit. It turns out that:

> ***Changes to the electric and magnetic fields and potentials produced by the motion of charges propagate away from the sources at the speed of light!***

Proving this is beyond the scope of this course – it really requires the vector differential formulation of Maxwell's equations, the so-called "vector potential", and gauge field theory, but for the time being we can simply accept it as an empirically true natural law in its own right. In other words, we can only tell that the electrostatic and magnetostatic fields and potentials propagate away from sources at speed c because if we *move* them in certain ways, we can observe the *changes produced by the motion* only after a lag time:

$$\Delta t = \frac{|\Delta \vec{x}|}{c} \tag{9.143}$$

where $|\Delta \vec{x}|$ is the distance between the source point (at the time t_e of emission) and the observation point (at the time $t_o = t_e + \Delta t$). Note that this doesn't affect any of our static field results above for (not really) "stationary" charges or continuous static currents as long as the charges or currents have been there long enough that their static fields have had time to fill space within at least the distance between the sources and our measuring apparatus.

This general idea should be fairly familiar to us already from our studies of sound waves. Because sound travels so much more slowly than light, it is common enough to e.g. see a flash of lightning and have to wait before the clap of thunder it produces arrives at our ears. Light is a bit trickier, because it turns out *nothing* carries information from one point of space to another faster than the speed of light, so we *can't* be located at one point in space and somehow "see" that a charged particle moves far away any faster than the field at our point of observation changes as a result of the motion!

Armed with this one small addition to our existing knowledge of electrostatics, we can fairly simply establish the connection between an accelerated charge and the appearance of electromagnetic radiation.

9.7.1: Larmor Radiation from Accelerating Charges

The following derivation of so-called "Larmor" radiation from an accelerated point charge is due to J. J. Thomson, dating back to the first decade of the 20th century just about the time Einstein was deriving special relativity. It is non-relativistic, and requires that our charges never move at a speed comparable to the speed of light.

Consider a point charge q that has been located at rest at the origin of a frame S "for a long time" (long enough that the *only* field visible in the nearby surrounding space is the usual electrostatic field of a point charge centered on the origin). At the time $t = 0$, we will give this charge a uniform "push" for a very short time $\Delta\tau$ so that it experiences a constant acceleration

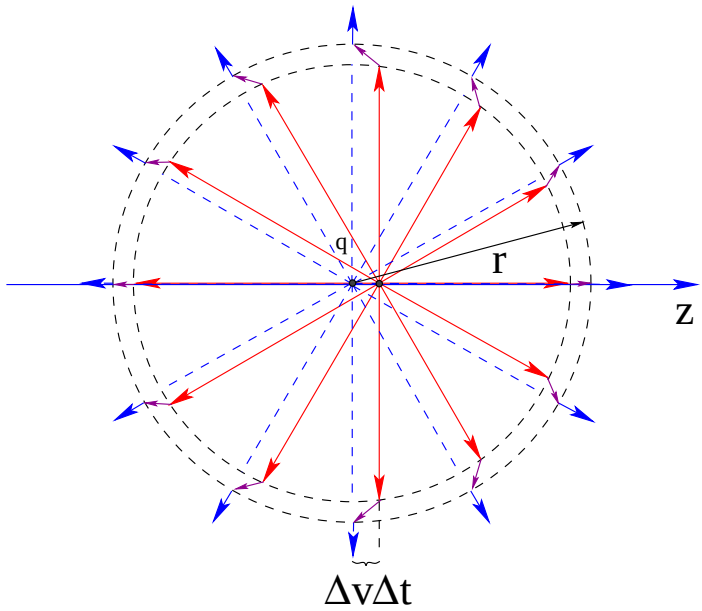


Figure 9.7: The field lines produced by a charge initially at rest at the origin that undergoes a brief acceleration to a final speed $\Delta v = a\Delta\tau \ll c$, at the time $\Delta t \gg \Delta\tau$ that the *change* in the electrostatic field reaches the sphere of radius $r = c\Delta t$. The red field lines are centered on the moving charge, the blue field lines are centered on the charge before it began to move, and the purple field lines represent the disjunction of the field across the spherical shell of thickness $c\Delta\tau$.

in the z direction. At the end of this time, the charge has the uniform velocity:

$$\Delta v = a\Delta\tau \tag{9.144}$$

in the z direction and has moved a very short distance:

$$\Delta z = \frac{1}{2}a\Delta\tau^2 = \frac{1}{2}\Delta v\Delta\tau \tag{9.145}$$

Note well that we require $\Delta v \ll c$ both to make the following geometric analysis work out and to avoid the need for special relativity when changing inertial reference frames.

This motion – plus our new rule above that *changes* to the electrostatic field can propagate no faster than the speed of light – creates a kind of a *disjunction* in the electrostatic field. If we are sitting a a point of observation $P = (r, \theta, \phi)$ in spherical polar coordinates (where ϕ is irrelevant because our z directed motion is azimuthally symmetric under rotation around the z-axis) then the field there will remain unchanged for a time

$$\Delta t = \frac{r}{c} \tag{9.146}$$

We will express this the other way around. For $r \geq c\Delta t$, the purely radial electrostatic field in the frame S will just be:

$$E_r = \frac{k_e q}{r^2} \tag{9.147}$$

After the time $\Delta\tau$ the charge is moving uniformly with speed Δv along the z-axis. As we know, there exists a frame S' co-moving with the charge (so it is at rest and at the origin of S')

and within an expanding sphere of radius $r - c\Delta\tau = c(\Delta t - \Delta\tau)$, there is an electrostatic field centered on the point charge. If we transform this field back into S, it will look just like a radial electrostatic field whose center points back to the position of the charge at the time Δt.

This situation is illustrated in figure 9.7, where the the (blue) field lines represent the field at the boundary of the sphere of radius $r = c\Delta t$ and beyond, and the (red) field lines all point back to the position of the charge at time Δt as they hit the sphere of radius $r - c\Delta\tau$.

We are implicitly assuming that $\Delta\tau \ll \Delta t$, and are hence simply ignoring the tiny displacement of the charge $\frac{1}{2}\Delta v\Delta\tau$ during the acceleration itself relative to $\Delta v\Delta t$. Also, because $\Delta v \ll c$, $\Delta v\Delta t \ll c\Delta t$. If we were to draw figure 9.7 *properly* to scale in this limit, $\Delta v\Delta t$ would be indistinguishable from the origin, and the two dashed spheres in the figure would form a spherical shell of thickness $c\Delta\tau$ *centered on the origin of* S*!*

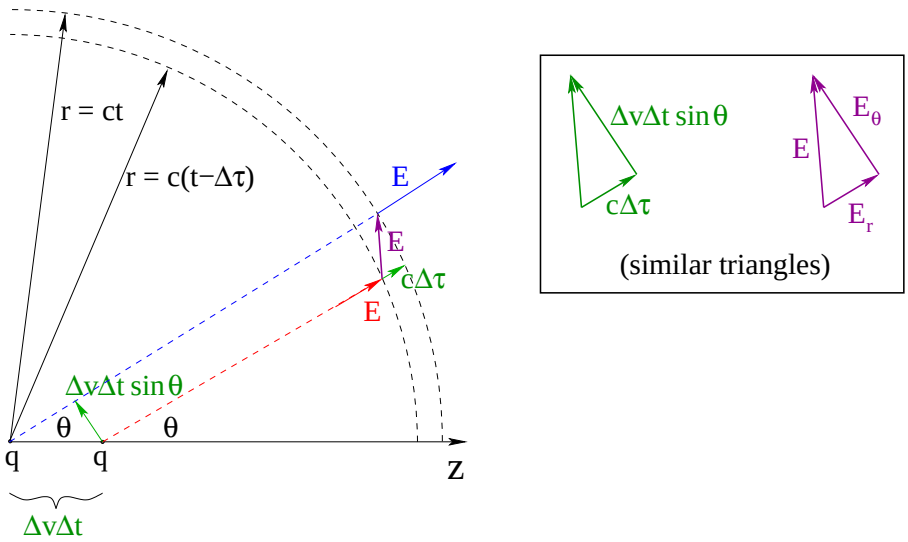


Figure 9.8: Close-up of the geometry of the $\vec{E}$-field across the spherical shell from $r = ct$ to $c(t + \tau)$. Note well that $t \gg \tau$ and $\Delta v = a\tau \ll c$ – the thickness of the shell and displacement of the charge are greatly magnified to make it easy to see the relevant triangles.

Figure 9.8 represents these two origin-centered spheres on a scale that greatly exaggerates the size of $\Delta v\Delta t$ relative to $r = c\Delta t$, but that *also* preserves the geometry essential to determining the components of $\vec{E}$ across the transitional spherical shell. Recalling the rule that electrostatic field lines can begin or end only upon a charge, the (purple) $\vec{E}$ field lines connect the lines emitted from the charge at the *common angle* θ (relative to the z-axis) on either side of the spherical shell of thickness $c\tau$. It is this purple connecting field that is responsible for radiation, and we need to estimate its spherical polar coordinates. As we proceed, we will effectively eliminate both $\Delta\tau$ and Δt from the result in favor of the acceleration a and r.

First, finding the radial field strength inside the shell is simple enough. We use the binomial expansion in the limit (established above) that $\Delta t \gg \Delta\tau$ to approximate the radial field strength

E_r at the *inner* sphere of radius $c(\Delta t - \Delta\tau)$:

$$\begin{aligned} E_r &= \frac{k_e q}{(c(\Delta t - \Delta\tau))^2} \\ &= \frac{k_e q}{(ct)^2\left(1+\frac{\Delta\tau}{\Delta t}\right)^2} \\ &\approx \frac{k_e q}{(c\Delta t)^2}\left(1 - 2\frac{\Delta\tau}{\Delta t} + ...\right) \\ &= \frac{k_e q}{(c\Delta t)^2} = \frac{k_e q}{r^2} \end{aligned} \tag{9.148}$$

which is the radial field strength at the *outer* sphere of radius $c\Delta t$. This establishes that E_r is effectively *continuous* across the spherical shell where the field direction "suddenly" shifts due to the acceleration of the charge and in any event produces no radiation of energy.

On the other hand, E_θ is the component *perpendicular* to $\vec{r}$ that cannot be shifted away by any inertial reference frame change and it is *this* component that we expect to result in electromagnetic radiation in the form of a nonzero radial Poynting vector. We can find this from the fact that in figure 9.8 the (purple) $\vec{E}$-field component triangle and the (green) triangle describing the relative displacement of charge location perpendicular to r and the distance light travels during τ are *similar*:

$$\frac{\Delta v \Delta t\ \sin\theta}{c\Delta\tau} = \frac{E_\theta}{E_r} \qquad \Rightarrow \qquad E_\theta = \frac{k_e q}{r^2}\frac{\Delta v \Delta t\ \sin\theta}{c\Delta\tau} \tag{9.149}$$

If we multiply by $c/c = 1$, turn $c\Delta t$ on top into r, and identify $a = \dfrac{\Delta v}{\Delta\tau}$ this becomes:

$$E_\theta = \frac{k_e q a\ \sin\theta}{c^2 r} = \frac{qa\ \sin\theta}{4\pi\epsilon_0 c^2 r} \tag{9.150}$$

where I've converted to a form in terms of ϵ_0 for later convenience. ***Note well!*** This *electrodynamic* component of the $\vec{E}$-field is *inversely proportional to r!* It diminishes *more slowly than the radial electrostatic field itself* with r.

Next we form the magnitude of the instantaneous (not time-averaged!) Poynting vector associated with this component of $\vec{E}$ perpendicular to $\vec{r}$:

$$|\vec{S}| = \epsilon_0 c E_\theta^2 = \frac{q^2 a^2\ \sin^2\theta}{16\pi^2\epsilon_0 c^3 r^2} = \frac{P_0}{r^2}\sin^2\theta \tag{9.151}$$

where:

$$P_0 = \frac{q^2 a^2}{16\pi^2\epsilon_0 c^3} \tag{9.152}$$

has units of ***power*** (watts). The magnitude of the Poynting vector is (recall) the *intensity* – power per unit area – of the electromagnetic radiation propagating away from the charge as it arrives a distance $r = ct$ away from its position where the acceleration occurred.

We can compute the total rate at which the charge radiated energy while it was accelerating by integrating the flux of the Poynting vector through the sphere of radius r in spherical polar coordinates:

$$P_{\rm tot} = \int_0^\pi \int_0^{2\pi} \left(\frac{P_0}{r^2}\sin^2\theta\right) r^2 \sin\theta\, d\theta\, d\phi = 2\pi P_0 \int_{-1}^{1}(1-x^2)dx = \frac{8\pi P_0}{3} \tag{9.153}$$

(making our usual $\sin\theta\, d\theta \Rightarrow -d\cos\theta$ change of variables) or:

$$\boxed{P_{\rm tot} = -\frac{dW}{dt} = \frac{q^2 a^2}{6\pi\epsilon_0 c^3}} \tag{9.154}$$

(where here and immediately below, W stands for the energy in the radiated field, not necessarily "work" per se).

This result is called ***Larmor's Formula*** and represents the rate at which one has to do *extra* work while accelerating the charge q. Some of the work goes into increasing the kinetic energy of the (massive) charged particle, but some of it is lost in the form of radiation from the particle as it radiates!

From Newton's third law, then, the emitted radiation field at the location of the charge must therefore *itself* act back on the charge as something called ***radiation resistance***. Radiation resistance looks suspiciously like a "self-force" and is worthy of far more discussion than we can reasonably include here[143] .

Apparently, if one takes an electric charge and accelerates it, it radiates away electromagnetic energy. Charge moving at a constant velocity (which is a frame transformation away from being charge at rest) does not radiate energy. It may produce an electric and magnetic field, but that field is guaranteed *not* to carry any energy away. Only when it accelerates does the charge radiate (and of course, there is no inertial frame that can get rid of that acceleration, so the radiation occurs in all frames).

Well, when *do* charges accelerate? One place is inside a linear accelerator used to study particle physics. Another is when a charged particle strikes a medium and slows to rest as it interacts with it (producing so-called "braking radiation"[144]). They accelerate if they are oscillating harmonically (see the next section). They accelerate if they move around in circles or follow any sort of curved path even at constant speed – circular motion is just harmonic oscillation in two dimensions simultaneously, and that pesky centripetal acceleration qualifies as one that would radiate energy *continuously* and not just in part of its cycle or when speeding up or braking linearly.

The Larmor formula, historically, has been one of the most profound results in all of physics, as it established a fundamental inconsistency between Newtonian classical mechanics and Maxwell's Equations. What it adds up to is this: ***There is no obvious way to make a model for an atom made up of an electrostatically bound state of electron(s) and proton(s) that does not involve orbiting or oscillating charge!*** No non-obvious way either, at least not classically, especially not one that agrees with the observation that atoms *do* radiate electromagnetic energy, but only at certain fairly sharp energies and frequencies!

In fact, if you build a simple model for a hydrogen atom consisting of a proton being orbited by a light electron in an orbit that is initially roughly the right size, you find that it collapses, with the electron spiralling into the proton while it radiates away energy, in less than a nanosecond.

[143]Wikipedia: http://www.wikipedia.org/wiki/Abraham-Lorentz_force. The actual force is called the Abraham-Lorentz force, and it appears as a *third* order differential equation, $\vec{F}_{\rm rad} = -\frac{q^2}{6\pi\epsilon_0 c^3}\frac{d\vec{a}}{dt}$. Equations of this sort have various problems with causality and have runaway solutions that must be excluded as non-physical, but in any event (as it turns out) the Universe *doesn't* explode with free energy due to radiation reaction...

[144]Wikipedia: http://www.wikipedia.org/wiki/Bremsstrahlung. Which sounds much cooler in German.

A strictly Newtonian classical Universe based on Maxwell's equations would last just about that long!

Of course, the visible Universe has been around for approximately *fourteen billion years* according to the most recent theories and observations[145] . We are forced to conclude:

> ***Either Maxwell's Equations are wrong or classical Newtonian mechanics itself is wrong (or both)!***

In either case nearly everything we've learned over the last two semesters is wrong, or, to be more charitable, approximate and/or incomplete.

Too bad, ladies and gentlemen. Maxwell's Equations appear to be more or less *correct*, although they do need to be slightly reexpressed in non-classical mechanics and ultimately augmented with other fields. Classical Mechanics is *not*. At this point you should visualize Isaac Newton (who was reportedly a dangerous man to cross in an argument and who had been elevated to a state of near worship for his manifold contributions to mathematics and science in general) spiralling down to the earth like Icarus, wings melted, never to rise again as high as he was before Maxwell and Faraday made their momentuous discoveries and their successors used them and additional evidence (such as the discovery of the electron and atomic nucleus) to arrive at this stark conclusion.

Over the last 140 years or so Newtonian mechanics (or its classical near-equivalents, Lagrangian and/or Hamiltonian mechanics) has been replaced by ***quantum*** mechanics, a *wave-like* theory of matter that is, to say the least, a lot more complicated than the relatively *simple* $\vec{\boldsymbol{F}} = m\vec{\boldsymbol{a}}$ that has governed nearly everything we have learned so far. Indeed, this is *almost* a good point to bridge over to a study of so-called "modern physics" – the non-Newtonian mechanics and non-Galilean relativity of the late 19th and 20th century – but there are a few more things we need to study first concerning ***light as a classical electromagnetic wave***.

Before we embark on a study of light per se, it is worth noting that *nearly all* of the electromagnetic radiation we directly observe comes from something that can be remarkably successfully modeled by *oscillating dipoles* acting as sources of electromagnetic radiation, predominantly *electric* dipoles at that. Our Lorentz model "linear response" polarization of an atom (or molecule) turns out to be very useful (all the way into graduate classical mechanics) for understanding many of the "generic" properties of radiating atoms and matter interacting with a time dependent electromagnetic field.

9.7.2: Dipole Radiation

Let's return to our undamped Lorentz model atom from our discussion of dielectrics. If we polarize the atom in the z-direction so that its dipole moment is initially $p_0 = qz_0$ for some (relatively mobile/polarizable) charge q, and release it, we'd expect that charge to oscillate har-

[145] Wikipedia: http://www.wikipedia.org/wiki/Age_of_the_universe. Note that there is some uncertainty in the most current estimate of 13.772 billion years old, and that even the error estimates on this number themselves are based on assumptions that could be overturned by, for example, a better understanding of so-called "dark matter" and "dark energy" or physics beyond the Standard Model.

monically, $z(t) = z_0 \cos \omega t$. In turn this means that the charge would *accelerate* harmonically:

$$a(t) = -\omega^2 z(t) = -z_0 \omega^2 \cos \omega t \tag{9.155}$$

so that:

$$qa = -q z_0 \omega^2 \cos \omega t = -p_0 \omega^2 \cos \omega t \tag{9.156}$$

However, we just observed that accelerated charges radiate! This means that instead of oscillating harmonically *without* damping, the atom would *radiate the total energy of the oscillator away* in the form of ***dipole radiation***

$$\vec{\boldsymbol{S}}(r, \theta, t) = \frac{(qa)^2}{16\pi^2 \epsilon_0 c^3 r^2} \sin^2 \theta \, \hat{\boldsymbol{r}} = \frac{p_0^2 \omega^4 \cos^2 \omega t}{16\pi^2 \epsilon_0 c^3 r^2} \sin^2 \theta \, \hat{\boldsymbol{r}} \tag{9.157}$$

or (averaging over time to get an extra factor of $\frac{1}{2}$) and noting that the intensity is the magnitude of the Poynting vector:

$$\boxed{I_{\text{avg}}(r, \theta) = \frac{p_0^2 \omega^4}{32\pi^2 \epsilon_0 c^3 r^2} \sin^2 \theta} \tag{9.158}$$

As before, we can integrate over the entire solid angle to get the Larmor formula for the *power* emitted from the dipole:

$$\frac{dW_{\text{tot}}}{dt}(t) = -\frac{q^2 a^2}{6\pi \epsilon_0 c^3} = -\frac{q^2 z_0^2 \omega^4}{6\pi \epsilon_0 c^3} \cos^2 \omega t \tag{9.159}$$

or (averaging over time to get an extra factor of $\frac{1}{2}$):

$$\boxed{P_{\text{avg}} = \frac{dW_{\text{avg}}}{dt} = -\frac{p_0^2 \omega^4}{12\pi \epsilon_0 c^3}} \tag{9.160}$$

This radiation acts much like a *damping* force, removing the energy of the oscillator not through "frictional" losses into some medium but through direct radiation of energy into the electromagnetic field. If one models this loss as a linear damping term, one obtains formulas for the *dispersion and scattering* of electromagnetic radiation that are the dynamical equivalent of the dielectric polarization response we studied before.

9.7.3: Why The Sky is Blue

You can see from the above that the total power radiated from an oscillating dipole ***scales with the fourth power of the angular frequency of the oscillator***. This is an enormously interesting result! As we will see in the *next* chapter on light, electromagnetic radiation ***polarizes*** atoms or molecules in the direction of the $\vec{\boldsymbol{E}}$-field of the incident radiation. The atom or molecule acts like a driven oscillator, with a dipole moment oscillating at the same frequency as the incident radiation, and *reradiates* energy it absorbs from the radiation field, *scattering* the radiation in all directions distributed like $\sin^2 \theta$ relative to its axis of polarization.

All things being equal, then, an atom or molecule oscillating in the violet end of the spectrum, with a frequency close to twice the frequency of red light, will reradiate almost ***16 times as much power*** as one oscillating in the red part of the spectrum.

White sunlight is made up of all the frequencies of visible light, but red-orange light tends to go *through* the atmosphere unscattered, where blue-violet light tends to be scattered sideways. This both makes direct, originally white, sunlight appear "oranger" by the time it has penetrated the atmosphere, while the blue sky we see when we look in any direction *but* at the sun is blue because the scattered light contains more blue-violet wavelengths than it does red-orange ones.

This is also, of course, why sunsets and sunrises appear red. The sunlight comes in at an oblique angle and traverses a much longer distance in the atmosphere than sunlight from straight overhead does. The blue-violet component is scattered out to the sides, leaving behind the far less attenuated red-orange light. On a particularly clear day, you can almost see a pastel "rainbow" of colors at sunset, where the colors vary from dark red up through violet as the light you see from the sky increases the path length it must follow from the sun, through the atmosphere, to your eye.

There are several things you can observe in the real world that reflect this. Emergency lighting, stop lights, brake lights (and more) tend to be red because red light can penetrate anything from a clear atmosphere to a foggy, dusty atmosphere much farther than any other visible color. On the other hand, the little lights they use on airport runways to direct planes on the ground tend to be blue, because blue lights are more or less *invisible* unless you are right on top of them! Landing planes cannot see them or be distracted by them – only planes almost on top of them can see them at all.

The next time you are driving at night on a foggy/rainy/dusty night, look at the stop lights along your route. When they are red, you will be able to fairly clearly see them, with relatively little of a scattered light "halo" surrounding them. When they turn green, on the other hand, the green light will be surrounded by a substantial blue-green halo of scattered light! Similarly, if you look at red LEDs on (say) electronics in your home, you will see the LEDs as point-like sources in the dark, with little halo. Blue LEDs on the same appliance will be surrounded by a substantial blue halo from the ω^4 increase in the intensity of scattered radiation at all angles BUT the primary line between the source and your eye!

We will not pursue the theory of dipole radiation any further at this time, but we *do* want to apply our result to the radiation pattern we expect in a physically important case:

Example 9.7.1: The Dipole Antenna

We can use the results above for a single charge oscillating as part of a stationary dipole, together with the superposition principle, to get the Poynting vector and Larmor formula for an *arbitrary* z-directed oscillating dipole moment, one made up of the collective oscillation of *many* charges in the usual coarse-grained limit

With this concept in hand, consider figure 9.9, a *dipole antenna*[146] . An alternating voltage at angular frequency ω is applied to two conducting rods separated in the middle by a small gap. The top rod is first charged up to be positive, then negative, then positive, as the voltage

[146]Wikipedia: http://www.wikipedia.org/wiki/Dipole_antenna. If you follow this link, learn at least one thing from the article: Classical dipole antennae are *a lot more complicated* than what I present here, which is something that is at best valid – as noted – only in the limit that the dipole is much smaller than a wavelength and oscillates more or less like a "perfect dipole".

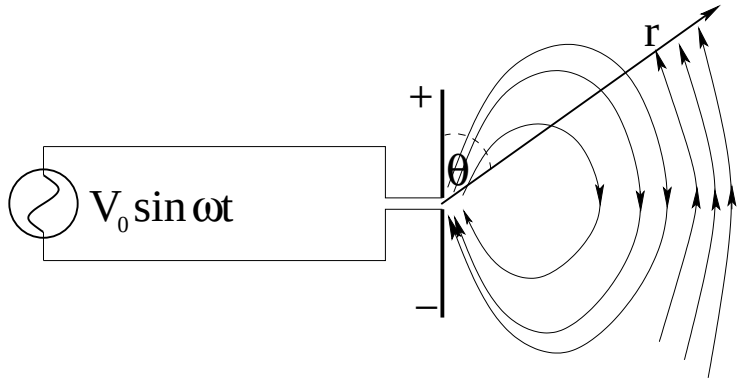


Figure 9.9: A simple "dipole antenna". An alternating voltage is applied to two simple wires, creating an *electric dipole* whose moment varies harmonically in time.

applied to the two rods oscillates, with the bottom rod charging up the exact opposite way. This makes the two rods at least approximately into a *harmonically oscillating dipole moment* centered between the two rods:

$$\vec{\boldsymbol{p}}(t) = p_0 \hat{\boldsymbol{z}} \cos \omega t \tag{9.161}$$

In order for the previous section (that really only worked for "pointlike" dipoles, ones where $p_z = qz_0$ with $z_0 \ll \lambda = c/f$ for the frequency of electromagnetic radation being produced) to apply, we will assume that the length of the dipole antenna is much smaller than a wavelength; otherwise we would have to work much harder (as you may well do in some future course). Witht his done, we can simply apply the results obtained above, substituting the *collective* dipole moment of the antenna for the dipole moment of the oscillating point charge above:

$$\vec{\boldsymbol{S}}(r,\theta) = \frac{p_0^2 \omega^4 \cos^2 \omega t}{16\pi^2 \epsilon_0 c^3 r^2} \sin^2\theta \, \hat{\boldsymbol{r}} \tag{9.162}$$

and average over time as before to get the average intensity radiated in the $\hat{\boldsymbol{r}}$ direction at the angle θ relative to the z-axis:

$$\boxed{I_{\text{dipole}}(r,\theta) = \frac{p_0^2 \omega^4}{32\pi^2 \epsilon_0 c^3 r^2} \sin^2\theta = \left(\frac{3P_{\text{tot}}}{8\pi}\right) \frac{\sin^2\theta}{r^2}} \tag{9.163}$$

where

$$\boxed{P_{\text{tot}} = -\frac{dW}{dt} = \frac{p_0^2 \omega^4}{12\pi \epsilon_0 c^3}} \tag{9.164}$$

is the total average power radiated from the antenna.

Note well that in order for the radiation field to take on this nice form, we *also* need to be *far enough away* that the assumption for r at the point of observation that we built into our derivation of the Larmor formula continues to hold, that is, that we are in the "far zone" where $r \gg z_0$ with z_0 the *maximum* displacement of charge from the origin in our antenna. If this condition is satisfied, we expect the radiation to at least have an angular distribution of $\sin^2\theta$ relative to the axis of the dipole itself and drop off like $1/r^2$ with distance. Otherwise, if we are too near the dipole, or if the dipole is too large relative to the wavelength, then – *you guessed it!* – we have to work a lot harder and our simple geometric form and estimate above (while usually not completely crazy) won't necessarily be particularly *quantitatively* accurate.

This is a great place to stop talking about sources (pending some possible future course in Electrodynamics where physics majors will fill in many of the missing details). Physics major or not, understanding that:

- accelerated charges radiate;
- electromagnetic radiation in the form that is most important/useful to our biology and technology comes from *oscillating electric dipoles*;

are the two critical facts you need to kickstart our explanation of that particular *band* of electromagnetic radiation we refer to as ***light*** as well as the rest of the electromagnetic spectrum.

Homework for Week 9

Problem 1.

Physics Concepts

Make this week's physics concepts summary as you work all of the problems in this week's assignment. Be sure to cross-reference each concept in the summary to the problem(s) they were key to. Do the work carefully enough that you can (after it has been handed in and graded) punch it and add it to a three ring binder for review and study come finals!

Problem 2.

Use Faraday's Law and Ampere's Law (in the absence of local currents) ***or*** a mix of frame rotation and relabeling to show that for a z-directed plane wave (where $\vec{E}$ and $\vec{B}$ are independent of x and y):

$$\frac{\partial E_y}{\partial z} = \frac{\partial B_x}{\partial t}$$
$$\frac{\partial B_x}{\partial z} = \mu_0 \epsilon_0 \frac{\partial E_y}{\partial t}$$

In one case you will need to start with the integral forms applied to small (differential scale) loops to recapitulate the method shown in lecture and the textbook; in the other you can start from the result itself derived in the textbook, but you will have to be clever! Extra kudos for doing it both ways!

Problem 3.

Use the differential form of Ampere's Law (in the absence of local currents) and Faraday's Law:

$$\begin{aligned}
\frac{\partial E_x}{\partial z} &= -\frac{\partial B_y}{\partial t} \\
\frac{\partial B_y}{\partial z} &= -\mu_0\epsilon_0\frac{\partial E_x}{\partial t} \\
\frac{\partial E_y}{\partial z} &= \frac{\partial B_x}{\partial t} \\
\frac{\partial B_x}{\partial z} &= \mu_0\epsilon_0\frac{\partial E_y}{\partial t}
\end{aligned}$$

to show that time-varying coupled fields (E_x, B_y) and (E_y, B_x) can exist that both satisfy the 1D wave equation for a $+z$-directed wave as long as $\vec{\boldsymbol{E}} \times \vec{\boldsymbol{B}}$ points in the $\hat{\boldsymbol{z}}$ direction, and that the speed of the electromagnetic wave is

$$c = \sqrt{\frac{k_e}{k_m}} = \frac{1}{\sqrt{\mu_0\epsilon_0}} \approx 3.00 \times 10^8 \ \mathrm{m/sec}$$

(accurate to three significant digits).

Problem 4.

Suppose you are given two regions of space where $\mp\hat{\boldsymbol{z}}$-directed (only) electric waves are propagating of the form(s):

L/+ $\vec{\boldsymbol{E}}_L(z,t) = E_0 \sin(kz + \omega t)\hat{\boldsymbol{x}}$ **R/-** $\vec{\boldsymbol{E}}_R(z,t) = E_0 \sin(kz - \omega t)\hat{\boldsymbol{x}}$

respectively. Use the differential forms of Faraday's Law and Ampere's Law (in the absence of local currents):

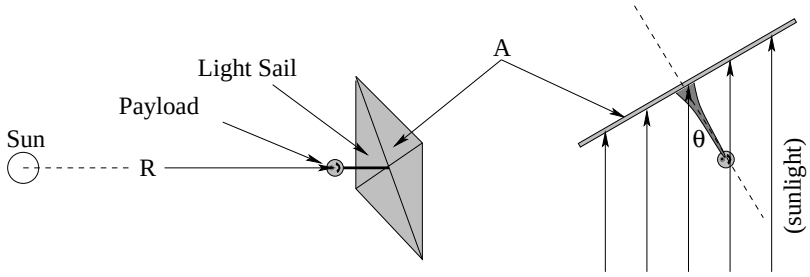
$$\begin{aligned}
\frac{\partial E_x}{\partial z} &= -\frac{\partial B_y}{\partial t} && \textbf{Faraday} \\
\frac{\partial B_y}{\partial z} &= -\mu_0\epsilon_0\frac{\partial E_x}{\partial t} && \textbf{Ampere + MDC}
\end{aligned}$$

to determine the form of the coupled $\vec{\boldsymbol{B}}(z,t)$ waves in the two regions. Use the result to prove the important wave properties presented in class and the textbook: That the amplitude of the B-field is given by $B_0 = \frac{1}{c}E_0$, that the $\vec{\boldsymbol{E}}$ and $\vec{\boldsymbol{B}}$ waves have the same frequency, wavelength and are ***in phase***, and that the **electromagnetic** wave they describe propagates in the

$$\vec{\boldsymbol{E}} \times \vec{\boldsymbol{B}}$$

direction in both cases

Problem 5.



perspective view (not to scale) side view

This problem analyzes the plausibility of using a perfectly reflective ***Solar Sail***[147] as a means of e.g. boosting communications satellites back up into stable orbits as they decay or cheaply sending fuel and life support materials to Mars ahead of any human voyage. A table of useful data is included below to help support your analysis.

a) Prove that the vector force exerted on a tipped, flat, solar sail a distance $R = rR_0$ from the sun by perfectly reflected sunlight can be expressed as:

$$\vec{\boldsymbol{F}}(r) = P_0 \frac{1}{r^2} A \cos^2(\theta) \hat{\boldsymbol{\theta}}$$

and determine the numerical value of P_0 in newtons/meter^2. Note that r is the distance from the sun in "Astronomical Units" (AU) see data below.

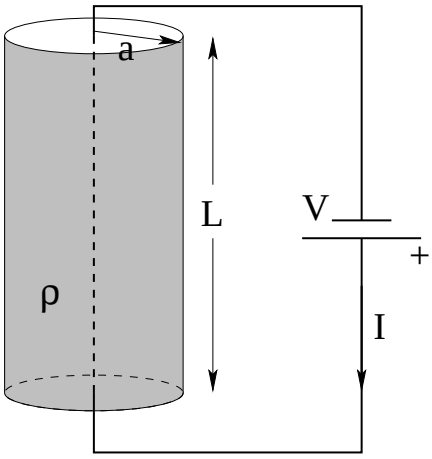
b) A solar sail material currently being tested is ***reflective mylar*** with a thickness of roughly 5 μm (microns). Assuming that the density of mylar is 1.5 times that of water, determine the area A of a solar sail with a mass of 500 kg.

c) The sail is only useful if it can lift a payload. Assume that the payload and all support and control structures have a mass of 500 kg as well, and determine the maximum ***acceleration*** that can be produced by the sail at the radius of the Earth's orbit (where $r = 1$).

Useful Data: The power output of the sun is 3.8×10^{26} watts. Its mass is 2×10^{30} kg. The mean radius of the Earth's orbit is $R_0 = 1.5 \times 10^{11}$ meters or $r = 1$ AU. The density of mylar is $\rho_m = 1.5 \times 10^3$ kg/meter^3. The mean radius of Mars' orbit is $r = 1.5$ AU.

[147] Wikipedia: http://www.wikipedia.org/wiki/Solar Sail.

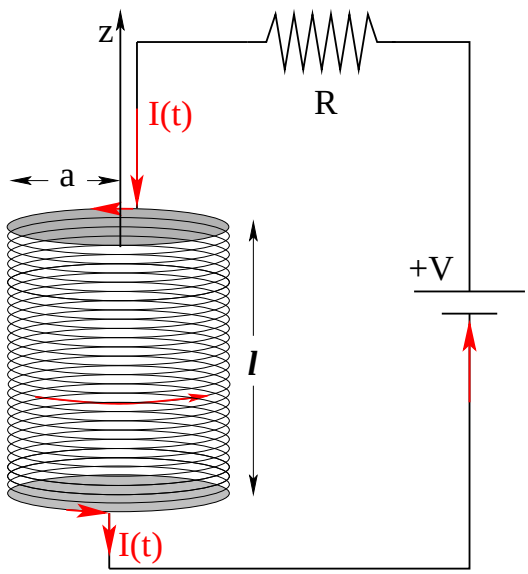
Problem 6.



Consider a resistor capped with perfectly conducting ends. The resistor is a cylinder of radius a and length L and is filled with a material of resistivity ρ. A voltage V is hooked up across the resistor so that uniform current (density) flows through it. Following the methods demonstrated in lecture and the textbook, find the $\vec{E}$ and $\vec{B}$ fields inside and at the surface of the resistor and prove that the power from Joule heating dissipated by this resistor equals the *flux of the Poynting vector* ***in*** *through the cylindrical side of the resistor*:

$$P_R = I^2 R = \frac{V^2}{R} = \int_{\text{side}} \vec{S} \cdot \hat{n} dA$$

Problem 7.

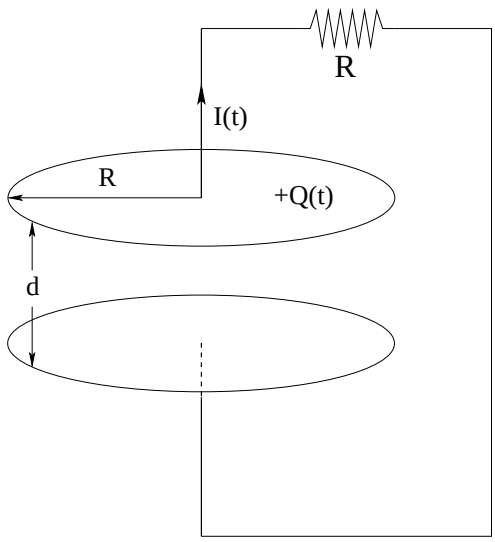


A "long" cylindrical solenoid of length ℓ, N turns, and radius a is in a circuit with a voltage V and resistor R as shown. A switch has just been closed so the current through the solenoid is $I(t) = I_0(1 - e^{-t/\tau})$ (current ***increasing*** in the red direction shown).

Find the magnitude and direction of the Poynting vector on an imagined surface of constant radius just inside the windings at radius a. Show that the flux of the Poynting vector through the cylindrical side of solenoid through the coils equals the rate at which the energy stored in the inductor changes:

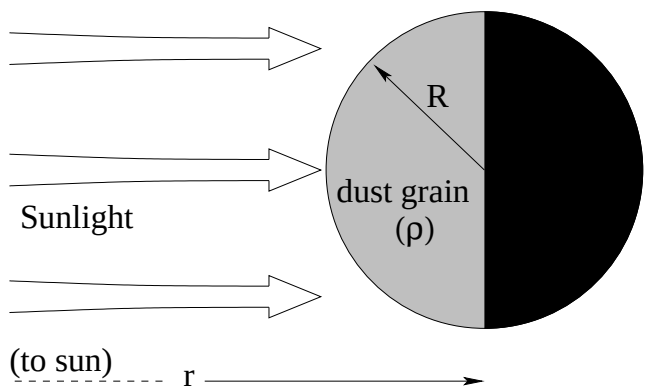
$$P_L = LI\frac{dI}{dt}$$

Problem 8.



A capacitor consisting of two circular conducting disks of radius R separated by a distance d was initially charged and is now discharging through a resistor as shown. Form the Poynting vector at a point on the "boundary" of the E field, assuming no fringing fields, and integrate the flux of the Poynting vector through this cylindrical surface. Show that the result equals $P_C = \frac{Q}{C}I$, the rate of change of the energy stored in the capacitor.

Problem 9.



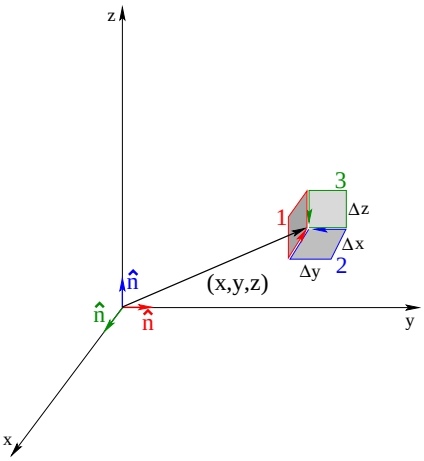
Consider a spherical grain of dust of radius R with a mass density ρ. Given the mass of the sun M_s, the total power emitted by the sun as electromagnetic radiation P_r, and your knowledge of G (the gravitational constant) and the insight that the radiation force exerted on this particle by sunlight is (within a factor of 2) produced by the total absorption of radiation flux incident on the ***transverse*** cross-sectional area of the sphere πR^2, ***estimate*** the "critical" radius R_c of a dust grain for which the force exerted by light pressure *away* from the sun *exactly balances* the gravitational force *towards* the sun independent of its distance from the sun. As always, solve this problem algebraically first, then substitute in $P_r = 3.8 \times 10^{26}$ watts, $M_s = 2 \times 10^{30}$ kg, $\rho = 10^3$ kg/m^3 (and other constants you should already know) and obtain a numerical answer as well.

This explains why small particles with $R \lesssim R_c$ are accelerated *away* from stars, forming a constant "wind" of microparticle radiation.

Problem 10.

A vertical cell phone radio tower acts as a dipole antenna. Suppose such a tower is located 1 km away from your cell phone. It radiates a power of 1 kilowatt. What is the approximate intensity of this radiation when it reaches your phone? Now consider your phone. It's dipole antenna radiates roughly one watt when it operates. What is the radiation intensity of your cell phone back at the tower?

Advanced Problem 11.



In the textbook (and lecture, and problem 2 of the homework) you learned to (for example) take a single loop such as (green) loop 3 to the left to find (partial) differential relations such as:

$$\mu_0\epsilon_0\frac{\partial E_x}{\partial t} = -\frac{\partial B_y}{\partial z}$$

In doing this, however, we only considered the y component of $\vec{\boldsymbol{B}}$ when doing the loop integral (working towards getting a 1-D wave propagating in the z direction). More generally, however, *both* B_y *and* B_z will contribute to the loop integral in Ampere's law leading to the partial of E_x with respect to time!

Either redo the derivation in the text for all *three* loops illustrated above, being sure to include the contributions to the loop integrals from the appropriate field components parallel to *all four sides* of each loop or (in my opinion, easier) use the *symmetry* and the right-hand rule to guess the form of the missing term in the equation above (and its equivalent from Faraday's law) and then cyclically permute the coordinate labels to get the other two components. Either way, you should show that:

$$\mu_0\epsilon_0\frac{\partial \vec{\boldsymbol{E}}}{\partial t} = \vec{\boldsymbol{\nabla}} \times \vec{\boldsymbol{B}}$$
$$\frac{\partial \vec{\boldsymbol{B}}}{\partial t} = -\vec{\boldsymbol{\nabla}} \times \vec{\boldsymbol{E}}$$

are the vector differential forms of Ampere's and Faraday's law respectively in free space (where ρ and $\vec{\boldsymbol{J}}$ are both zero). **Hint:**

$$(\vec{\boldsymbol{\nabla}} \times \vec{\boldsymbol{F}})_x = \frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z}$$

and cyclic permutation will give the other two components of the curl as well. **Note Well!** The equivalence of:

$$\oint_C \vec{\boldsymbol{F}} \cdot d\vec{\boldsymbol{\ell}} = \int_{S/C} (\vec{\boldsymbol{\nabla}} \times \vec{\boldsymbol{F}}) \cdot \hat{\boldsymbol{n}} dA$$

for a general smoothly differentiable vector field is called *Stokes' Theorem*, which you have just (sort of) proved!

Part I

Optics

Week 10: Light

- The speed of light in a medium is:

$$v_{\text{medium}} = \frac{c}{n} \tag{10.1}$$

  n is called the *index of refraction* of the medium. You need to know the following *approximate* indices of refraction to work problems: Air: $n_a \approx 1$. Water: $n_w \approx 4/3$. Glass: $n_g \approx 3/2$. Any others needed will be given in the problem in context.

- The index of refraction is not constant – it varies with the *frequency* of the light: $n(\omega)$, a phenomena known as *dispersion*.

- In the visible range, for most common transparent materials (e.g. normal glass, water, plastic) $n($**red**$) < n($**violet**, that is, the index of refraction ***increases with frequency*** across the visible spectrum. One can, however, engineer glasses where the opposite is true. Dispersion curves in general have distinct ranges where the index of refraction increases ***or*** decreases with frequency across the entire range of electromagnetic radiation frequencies.

- The Law of Reflection:

  The angle of incidence equals the angle of reflection,

$$\theta_i = \theta_\ell \tag{10.2}$$

- Snell's Law:

$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2) \tag{10.3}$$

- Fermat's Principle:

  Light takes the path that minimizes the time of flight between any two points. Both the law of reflection and Snell's law can be derived from Fermat's principle.

- Critical Angle, Total Internal Reflection:

  Light passing from a dense medium n_2 to a less dense medium $n_1 < n_2$ is *totally internally reflected* if the angle of incidence is greater than:

$$\theta_c = \sin^{-1}\left(\frac{n_1}{n_2}\right) \tag{10.4}$$

- Polarization:

  We describe the orientation and phase of the two components of the *electric* field component for a given fixed harmonic frequency as the *polarization* of the harmonic wave.

- Unpolarized Light:

  Unpolarized light is light for which the polarization vector is constantly shifting its direction around. On average, unpolarized light has its energy/intensity equally distributed between the two independent directions of polarization.

- Linear Polarization:

  Linear polarization occurs whenever the electric field vector oscillates consistently in a single vector direction in the plane perpendicular to propagation.

- Circularly Polarized Light:

  Circularly polarized light has the same electric field magnitude in the two independent polarization directions but the waves in these directions are $\pi/2$ out of phase:

$$\begin{aligned}\vec{\boldsymbol{E}}(z,t) &= \frac{\sqrt{2}}{2}E_0\hat{\boldsymbol{x}}\sin(kz-\omega t\pm\pi/2)+\frac{\sqrt{2}}{2}E_0\hat{\boldsymbol{y}}\sin(kz-\omega t)\\ \vec{\boldsymbol{E}}(z,t) &= \frac{\sqrt{2}}{2}E_0\left(\pm\hat{\boldsymbol{x}}\cos(kz-\omega t)+\hat{\boldsymbol{y}}\sin(kz-\omega t)\right)\end{aligned} \tag{10.5}$$

  There are two independent *helicities* of circularly polarized light: right (clockwise/+) and left (anticlockwise/-) when facing *in* the direction of propagation).

- Elliptically Polarized Light:

  If the amplitudes of the two waves are (potentially) different *and* the two waves are (potentially) out of phase, the most general polarization state is that of *elliptical* polarization:

$$\vec{\boldsymbol{E}}(z,t) = E_{0x}\hat{\boldsymbol{x}}\sin(kz-\omega t+\delta_x)+E_{0y}\hat{\boldsymbol{y}}\sin(kz-\omega t+\delta_y) \tag{10.6}$$

  In this expression, E_{0x} and E_{0y} may or may not be equal, and the phases δ_x and δ_y may or may not be zero *or* equal.

- Polarization by Absorption (Malus's Law):

  For an ideal polaroid filter that is otherwise fully transparent:

$$I_{\text{transmitted}} = \frac{I_{\text{incident}}}{2} \tag{10.7}$$

  The transmitted light is fully linearly polarized in the direction of the **transmission axis** of the filter.

  If the light that is incident on the filter is already polarized, then only the *component* of the electric field vector that is *parallel* to the transmission axis is transmitted:

$$E_{\text{transmitted}} = \vec{\boldsymbol{E}}\cdot\hat{\boldsymbol{T}} = E_{\text{incident}}\cos(\theta) \tag{10.8}$$

where θ is the angle between the direction of linear polarization of the incident light and a unit vector along the transmission axis. This implies that the transmitted intensity is given by:

$$I_{\text{transmitted}} = I_{\text{incident}} \cos^2(\theta) \tag{10.9}$$

This result is known as **Malus's law**.

- Polarization by Scattering:

  Rays scattered more or less at right angles to an atom, molecule, or speck of dust are linearly polarized **perpendicular to the plane of scattering.**

- Polarization by Reflection:

  Light that is reflected at a non-normal angle from a dielectric surface is (partially or completely) polarized **parallel to the surface**, which is also **perpendicular to the plane of reflection**. Light transmitted into the new medium is partially polarized the opposite way (by subtraction).

  The reflected light is *completely* polarized when the light is incident at the *Brewster angle*, where the reflected and refracted rays are perpendicular to each other, given by:

$$\tan(\theta_b) = \frac{n_2}{n_1} \tag{10.10}$$

- Polaroid Sunglasses:

  Reflected glare from any smooth surface and scattered glare at midday are both likely to be at least partially polarized *parallel to the ground*. Both are thus blocked by a pair of polaroid sunglasses with a **vertical transmission axis**.

- Doppler Shift, Moving Source:

  In a non-relativistic setting ($v_s \ll c$):

$$f' = \frac{f}{(1 \mp \frac{v_s}{c})} \tag{10.11}$$

  for an approaching (-) or receding (+) source describes the general moving source doppler shift in the frequency/color detected by the receiver.

- Doppler Shift, Moving Receiver:

  Again in a non-relativistic setting ($v_r \ll c$):

$$f' = f(1 \pm \frac{v_r}{c}) \tag{10.12}$$

  for a receiver moving towards (+) or away from (-) the source.

- Moving Source and Moving Receiver:

  Ditto:

$$f' = f\frac{(1 \pm \frac{v_r}{c})}{(1 \mp \frac{v_s}{c})} \tag{10.13}$$

- Cerenkov Radiation:

  The "light boom" given off by a charged particle moving faster than the speed of light *in a medium* is called *Cerenkov radiation*.

10.1: The Speed of Light

We just learned that the speed of light in a vacuum, derived from Maxwell's Equations, is $c = 1/\sqrt{\epsilon_0\mu_0} = 3 \times 10^8$ meters/second. However, we have *also* learned that the permittivity and permeability of bulk polarizable matter are not equal to their vacuum equivalents. The conclusion is inescapable. The speed of light is not c in a medium.

We *expect* it to be $v = 1/\sqrt{\epsilon\mu}$ where e.g. $\epsilon = \epsilon_r\epsilon_0$ (scaled by the dielectric and diamagnetic constants of the material). It turns out for many reasons that the polarization of the medium always *slows down the wave* – in free space it just sweeps along, but in the medium it has to move all of that bulk charge too, which has mass and cannot respond as quickly. For most transparent materials, $\mu \approx \mu_0$ so:

$$v \approx \frac{1}{\sqrt{\epsilon_r\epsilon_0\mu_0}} = \frac{c}{\sqrt{\epsilon_r}} \tag{10.14}$$

To keep life simple, we take all of the contributing properties of the material and roll them into a single relation:

$$v_{\text{medium}} = \frac{c}{n} \tag{10.15}$$

n is called the *index of refraction* of the medium and is roughly equal to $\sqrt{\epsilon_r}$ (which is dimensionless, recall).

However, there is a problem with this. ϵ_r is defined in the *static limit* of $\omega = 0$. Visible light has a frequency range of (roughly!) 4×10^{14} Hz to 8×10^{14} Hz (see tables below), and the charges in a dielectric material simply don't have *time* to reach their peak polarization before the wave points the other way!

Indeed, it turns out that the index of refraction is a ***function of frequency*** – $n(\omega)$ – a phenomenon known as ***dispersion***. This means (as we shall see) that different frequencies are bent by different amounts via ***Snell's law*** at an interface between two dispersive media, splitting white light up into a ***spectrum*** of colors, with the highest frequency (shortest wavelength) light usually getting bent the *most* although this is very much dependent on the particular medium in question.

This is why water droplets break up light into a rainbow! Note well that this means that – as far as we can tell examining the world around us or looking back into the remote past as we look up at the stars – water droplets have *always* broken up light into rainbows when backlit by a local source of light, just as they do if you spray water in a fine mist away from the sun in your back yard.

This has profound religious and philosophical consequences. At one time there was a rather extensive argument concerning the "frangibility of light" where Biblical literalists argued that this process could not have occurred before the Flood in Genesis, as it clearly states therein that the rainbow was first created *at a specific antediluvian time* as a sign that God wouldn't try to drown the world ever again.

It is worth noting that if light wasn't "frangible" before this (mythical) Flood, there would have *been no light* as the processes that produce it are the same as the processes that break it up in interaction with matter into colors in rainbows and everywhere else. Nor would there have been

any *normal matter* – as we have just learned in considerable detail, the electromagnetic forces that hold atoms and molecules together *are* the forces that are responsible for polarizability, which in turn is responsible for dispersion.

In much of the text below, we will *idealize* the index of refraction and assume that it is "constant" and "simple" for certain well-known materials. Basically this amounts to taking its average value in the middle of the visible spectrum as its value, and then picking a convenient nearby rational number as "the value of the index of refraction" for that medium. Be aware that this is a pure and simple simplification for the sake of rendering the arithmetic finger-and-toe easy while still preserving the entire conceptual idea and algebraic structure. We will also do just enough stuff with dispersion and more realistic $n(\omega)$ for us to see how this goes as well.

10.2: The Spectrum

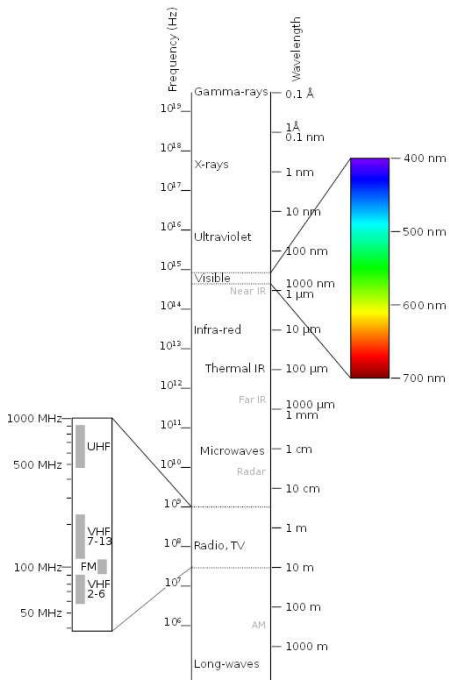


Figure 10.1: Graphical Representation of the Electromagnetic Spectrum

The sources of light can *classically* be viewed as charges bound into an electrically neutral atom in some sort of *equilibrium*. There were various classical models of stable neutral atoms that were tried out in the late 19th and early 20th century but they all failed. However, we can borrow one of the ideas – the idea that stable systems that are perturbed from equilibrium tend to *harmonically oscillate*, and (as we just saw in the chapter on Maxwell's Equations) *if* that system is an electric dipole, that oscillator will radiate away electromagnetic energy in the form of ***harmonic travelling waves*** – dipole radiation! Although the description of atoms that explains the full span of experimental observations ended up being quantum mechanical, the experimental observations themseles were indeed consistent with light being predominantly a harmonic travelling wave with a more or less definite frequency produced by oscillating electric (or magnetic) dipoles.

Harmonic electromagnetic radiation, however, is more than *just visible light!* It is *all* electromagnetic waves with *all* possible frequencies. Many ranges of electromagnetic frequency were independently identified and named, typically on the basis of both the wavelength/frequency as well as the mechanism used to generate the radiation or its *use*. Figure 10.1 above illustrates the ranges in a graphical format, while table 4 presents a short list of named ranges of the ***electromagnetic spectrum***, indexed by the *decreasing wavelengths* of the radiation from the longest to the shortest (alternatively, in the order of *increasing frequency*).

Name	Wavelength λ	Source(s)
Long Wavelength Radio	$>$ 1000 m	Electronics/Antennae
Radio	3-1000 m	Electronics/Antennae
FM/VHF/UHF Radio/TV	30 cm-3 m	Electronics/Antennae
Microwaves/Radar	1 mm-30 cm	Electronics/Antennae
Infrared Light	700 nm-1 mm	Thermal Sources (Hot Matter), Electronics
Visible Light	380-750 nm	Atoms, Molecules, LEDs
Ultraviolet Light	10-380 nm	Atoms, Very Hot Matter (Plasma)
X-Rays	$<$ 10 nm	Inner Shell Atomic Transitions
Gamma Rays	$<$ 0.001 nm	Nuclear Transitions

Table 4: The Electromagnetic Spectrum

Note that this list is not really complete, nor is it precise, as in many cases the spectral range of two independent means of generating radiation overlap, or here are multiple ways of generating the same "kind" of radiation, or particular bands with different uses are located within a more broadly named category. It is, however, important to know which of the *principle* named bands of waves have longer wavelength (smaller frequency) than which others, and to know at least approximate boundaries for the most important ranges.

In addition you are *required* to know the range of wavelengths and frequencies of visible light. You don't need to know these specifically indexed by color, but it is interesting to look over a table of colors and their associated frequencies and wavelengths to get a bit of a feel for it as well. I assume that most of you know the venerable mnemonic device "ROY G BIV" – standing for the colors in the visible part of the spectrum in the order of increasing frequency/decreasing wavelength: Red Orange Yellow Green Blue Indigo Violet. Note that many books and tables now omit Indigo as a separate color; this practice is continued in table 5 in *this* book.

Color	Frequency f	Wavelength λ
All	400-789 THz ($\times 10^{12}$ Hz)	380-750 nm
Red	400-484 THz	620-750 nm
Orange	484-508 THz	590-620 nm
Yellow	508-526 THz	570-590 nm
Green	526-606 THz	495-570 nm
Blue	606-668 THz	450-495 nm
Violet	668-789 THz	380-450 nm

Table 5: The Visible Light Spectrum

Both tables assume waves propagating in a vacuum (so the frequencies can easily be determined by using $f = c/\lambda$ where λ is the wavelength). Note well that again this table exaggerates the precision of the boundaries between colors. It is not the case that a wave with wavelength 621 nm is clearly red, but one with wavelength 619 is clearly orange. Different books specify slightly different "ends" of the range – 370-760 nm, for example. Personally, I think you will be just fine if you can remember the *approximate* ranges:

$$\boldsymbol{\lambda} = \textbf{400-700 nm},\ \boldsymbol{f} = \textbf{400-800 THz} \quad \text{(visible light)}$$

Then just remember that "red" light can be seen for wavelengths a bit longer than 700 nm, and "violet" light can be seen for wavelengths a bit smaller than 400 nm. Good enough.

10.3: The Law of Reflection

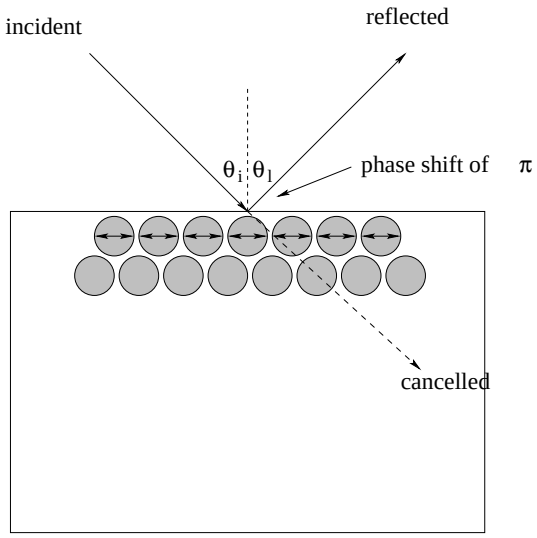


Figure 10.2: When light is incident on a perfectly reflecting surface, it creates little antennas/sources that radiate the *opposite* field in the direction of the incident field. These antennas cause the light to be reflected at the same angle and with the opposite phase from the surface.

A perfect conductor in electrostatic equilibrium, we recall, cancels the electric field inside by arranging charges on its surface to effect the cancellation. Similarly, it creates surface currents that oppose and cancel magnetic fields. In the dynamical case this is still true for good conductors and optical frequencies. An incoming *light* wave strikes the conductor, and its electric field *polarizes* the surface atoms so that they become little antennae that oscillate along with the electric and magnetic field of the light. However, the fields produced *flip over* (the way a dipole field does) and hence propagate in the leading direction with the *opposite phase*, cancelling the forward directed field quite rapidly at the surface (often within a few layers of atoms).

Since the conductor is good, very little energy is lost to eddy current heating during this cancellation. The oscillating surface currents must reradiate their energy, and the only direction they can do so that conserves energy and momentum is to *reflect* the incident energy. However, the reflected wave (in order to achieve the cancellation at the surface) must have

the *opposite phase* from the incoming wave. The situation is very much like the reflection of a wave pulse on a string from a fixed point on the wall – the reflected wave flips so it is upside down for precisely the same reasons (energy and momentum conservation).

In an elastic collision with the conductor, the component of the momentum of the light *along* the surface is unchanged, but the perpedicular component inverts (becomes minus itself). The only way this can be true is for the light to bounce off of the surface, with its phase inverted, at an angle of reflection θ_r (measured relative to the normal at the surface at that point) equal to the angle of incidence θ_i as drawn above.

So that's it:

$$\theta_i = \theta_\ell \tag{10.16}$$

is the Law of Reflection. The polarization properties of the reflected light will be discussed later below.

Note well that for this to be strictly true requires that the surface in question be extremely smooth – "shiny" as it were. Otherwise neighboring rays would be reflected at different angles because of small differences in the direction of a normal at different point on a rough surface. Many (even most) surfaces of real materials are indeed rough on a microscopic scale (compared to the wavelengths of the incoming light) and hence are diffusely illuminated ty light instead of perfectly reflecting it according to this rule. Many materials also differentially absorb light and only "reflect" particular wavelengths and hence colors.

We will assume that the law of reflection holds, more or less perfectly, for shiny smooth good conducting (e.g. metal) surfaces, such as a polished piece of silver or aluminum. This in turn will help us understand how *mirrors* work to form images of objects next week.

10.4: Snell's Law

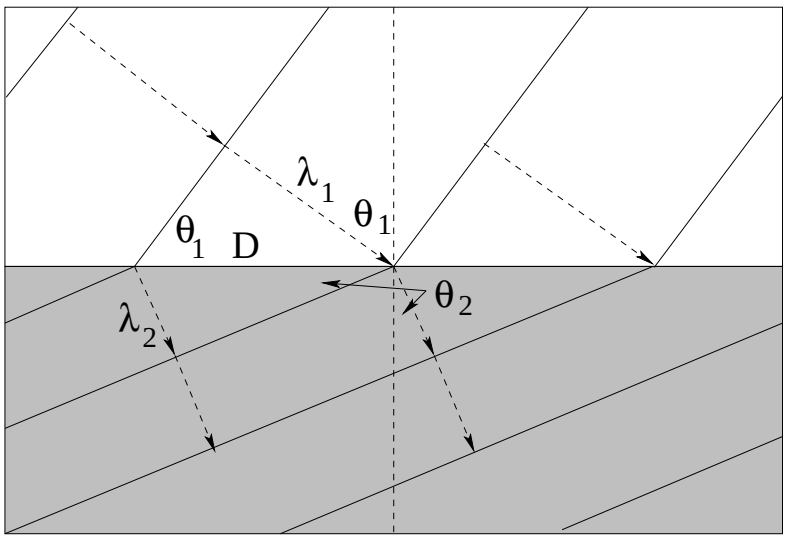


Figure 10.3: When light is incident on a transparent dielectric surface, it is partially transmitted and partially reflected. Since its *speed* changes, however, the light must *change direction* at the surface as shown.

Light is incident on a surface that separates two transparent media with different indices of refraction n_1 and n_2 (where we assume for the moment that $n_1 < n_2$ although that isn't necessary in the end). This is illustrated in figure 10.3 above.

It should be fairly obvious that the *frequency* of light in the two media cannot change. If the same number of wavefronts per second do not pass each point in either medium, wavefronts must be building up in between. This in turn means that energy (associated with the wavefronts) must be building up. This simply does not happen.

It should also be less obvious that the wavefronts themselves – the places where the waves reach their maximum amplitudes – should be the same just inside and just outside the media interface. For it to be otherwise would require a very strange charge distribution on the surface itself, one that one cannot easily imagine arising.

Since the wave must *change speed* across the media interface, and since the speed of the wave is given by:

$$v = \frac{c}{n} = f\lambda \tag{10.17}$$

with the same frequency on both sides, it is clear that the wavelength

$$\lambda = \frac{c}{nf} \tag{10.18}$$

must *also* change, being longer where the speed of light is greater (and n is smaller).

Simple geometry based on these simple ideas requires that the wave will also change direction. We can compute this change and direction from the figure above. If we look at the top triangle with angle θ_1 and hypotenuse D and the bottom triangle with angle θ_2 and the *same hypotenuse* (the distance between wavefronts on the interface between media), we note that:

$$D = \frac{\lambda_1}{\sin(\theta_1)} = \frac{\lambda_2}{\sin(\theta_2)} \tag{10.19}$$

or (substituting from above and cancelling c/f):

$$\frac{1}{n_1 \sin(\theta_1)} = \frac{1}{n_2 \sin(\theta_2)} \tag{10.20}$$

Inverting, we obtain **Snell's Law**:

$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2) \tag{10.21}$$

Since the geometry is exactly the same going from n_2 to n_1, we conclude that it doesn't matter which medium has the greater or the lesser index of refraction.

10.4.1: Fermat's Principle

In figure 10.4, we note that any curved path such as S_1 is longer than the path S_0 (something that can be proven using the calculus of variations, which we will not introduce here). The time required to traverse S_1 is $t_1 = S_1/v$ while $t_0 = S_0/v$. The minimal time path is therefore clearly the minimal distance path, the straight line. Fermat's principle thus correctly describes this case.

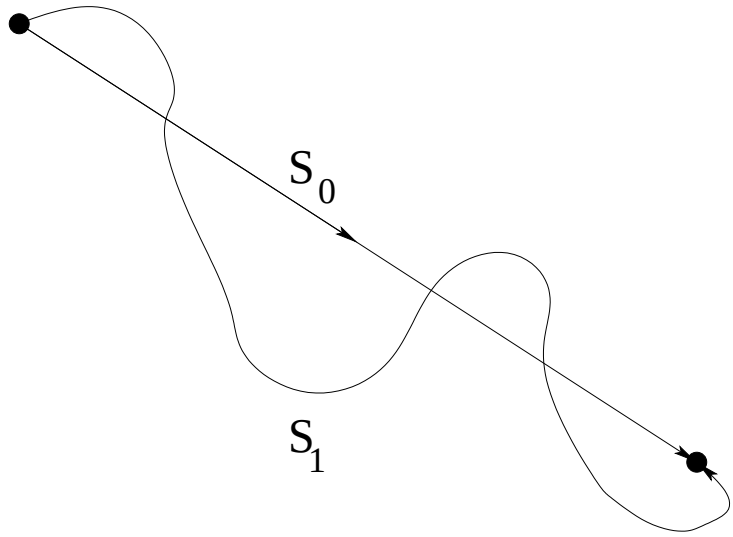


Figure 10.4: For constant speed, the straight line path between A and B takes the least time.

Fermat noted that a straight line is the path along which it takes the *least time* to travel between two points A and B at constant speed in ordinary space. Any other path is longer in distance than the straight line path, and hence takes longer to traverse at the same speed. This is illustrated in figure 10.4 – the curved path is longer, so it takes more time to traverse it if you have to move at exactly the speed of light (or the same speed along both trajectories).

Thus when we say that light travels a constant speed (the speed of light) in a straight line between A and B, it is *also* true that the path that it follows is the one that takes the least time.

Now consider the Law of Reflection above. It is equally easy to see that any reflective path between A and B that doesn't have $\theta_i = \theta_l$ is longer, and hence takes more time. We will examine and prove this below using calculus.

What happens when the speed is *not* constant? In that case, one has to solve an *optimization* problem, a problem in *economy*. It seems that one might be able to obtain some benefit from *going further* where the speed is greater and thereby reduce the amount of distance one has to travel at the slower speed, and actually go between A and B in *less* time than the straight line trajectory.

Fermat, observing that light must speed up or slow down as it passes between distinct physical media, hypothesized that the trajectory followed by light between point A in medium 1 and point B in medium 2 would *not* be a straight line; it would instead be the path that takes the minimum time. This, as we shall see, is *another* way to get Snell's law, but this time in a ray description of the light that is altogether independent of the wavelength or wave properties of the light.

Although Fermat was not the first person to propose a variational/minimum principle for optics (that honor belongs to Ibn al-Haytham in 1021, over 600 years earlier) he was the first to do so post Descartes, with an analytic geometry capable of fully exploiting the idea. Although Fermat's principle puts the cart a bit in front of the horse by making it the *cause* of the trajectory followed by light instead of a *feature* of the trajectory followed by light (that can be derived from other principles) variational principles based on his original statement proved to be essential to a formulation of classical mechanics that would translate, with minimal changes, into a

formulation of quantum mechanics. It is therefore worth looking at in a bit of detail, especially for physics majors or minors.

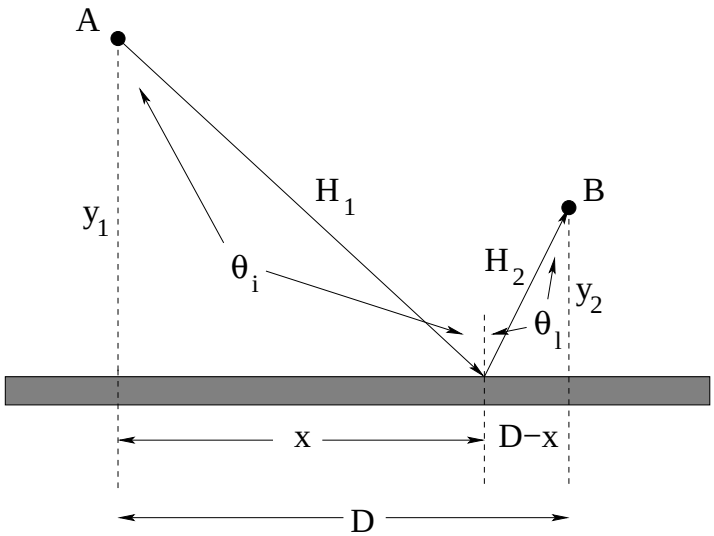


Figure 10.5: The path with $\theta_i = \theta_l$ is the one with the minimal time when the entire trajectory is otherwise in a single medium with a constant speed.

In figure 10.5 illustrate and prepare to prove the law of reflection from Fermat's requirement that the time required to go between points A and B on a path that reflects off of the mirror is a minimum. From the result above we can ignore all trajectories that are not straight except where they strike the reflecting surface. The total distance between the two points A and B is therefore the sum of the two hypotenuses:

$$\begin{aligned} H &= H_1 + H_2 \\ &= \left\{y_1^2 + x^2\right\}^{\frac{1}{2}} + \left\{y_2^2 + (D-x)^2\right\}^{\frac{1}{2}} \end{aligned} \tag{10.22}$$

We need to find a condition that produces the minimum of this function. We therefore differentiate with respect to x, set the result to zero, and solve for (say) x or θ_1. y_1, y_2 and D are all constant, so (using the chain rule, note well):

$$\frac{dH}{dx} = \frac{\frac{1}{2}2x}{\left\{y_1^2 + x^2\right\}^{\frac{1}{2}}} - \frac{\frac{1}{2}2(D-x)}{\left\{y_2^2 + (D-x)^2\right\}^{\frac{1}{2}}} = 0 \tag{10.23}$$

or

$$\sin(\theta_i) = \frac{x}{\sqrt{y_1^2 + x^2}} = \frac{x}{H_1} = \frac{D-x}{H_2} = \frac{(D-x)}{\sqrt{y_2^2 + (D-x)^2}} = \sin(\theta_l) \tag{10.24}$$

If the speed of light is a constant, this condition minimizes both distance and hence time $t = H/v$. Thus $\theta_i = \theta_l$, and we see that the Law of Reflection can be derived from Fermat's principle. What about Snell's Law?

To derive Snell's Law, we need a figure like that one drawn in figure 10.6. As was the case for reflection, we only need consider straight line trajectories *in* a given medium, but we allow x to (again) be a variable that we adjust to find the trajectory with the minimum time.

The major difference this time is that the speeds in the two media are *different*. When we right down the times required for the trajectories in media 1 and 2, we have to include the

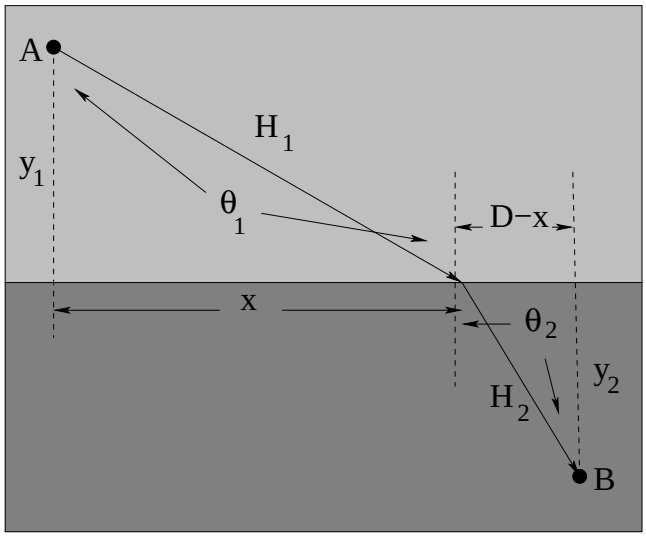


Figure 10.6: The path with $n_1 \sin(\theta_1) = n_2 \sin(\theta_2)$ is the one with the minimal time when the trajectory goes between media n_1 and n_2 where light has distinct speeds. As suggested, one minimizes the time by choosing a trajectory that trades off more distance in the faster medium against less distance in the slower one.

indices for refraction for those media, that is:

$$t_1 = \frac{\sqrt{y_1^2 + x^2}}{v_1} = \frac{n_1\sqrt{y_1^2 + x^2}}{c} \tag{10.25}$$

and

$$t_2 = \frac{\sqrt{y_1^2 + (D-x)^2}}{v_2} = \frac{n_2\sqrt{y_1^2 + (D-x)^2}}{c} \tag{10.26}$$

as the times it takes for the light to travel in a straight line 1) from A to x and 2) from x to B.

The total time is thus:

$$t = t_1 + t_2 == \frac{n_1\sqrt{y_1^2 + x^2}}{c} + \frac{n_2\sqrt{y_2^2 + (D-x)^2}}{c} \tag{10.27}$$

Differentiating and setting the result equal to zero recapitulates the *same algebra as used above* to derive the law of reflection, except that there is an extra factor of n_1 and n_2 on each side. The details are thus left as a (simple) exercise that you should attempt without looking back; the result is:

$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2) \tag{10.28}$$

and we see that Snell's law can be derived from Fermat's principle as well!

Variational principles prove to be of great use in more advanced physics, as nature appears to be intrinsically "economical" and choose extremal paths, usually ones that minimize a quantity called the *action*. Newton's laws themselves can be derived in a generalized form from a suitable variational principle of a quantity called the "action", and this proves to be a useful way to derive and understand parts of quantum theory as well!

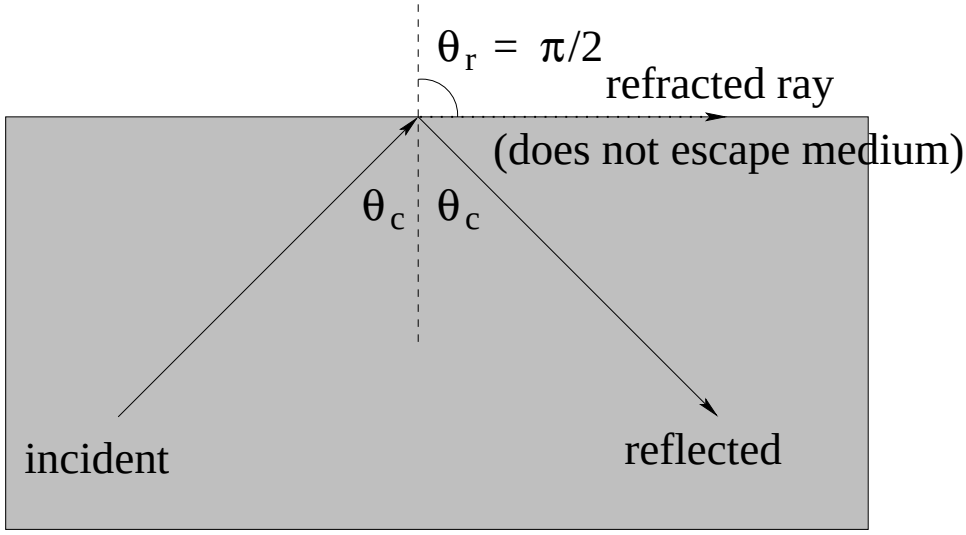


Figure 10.7: Light travelling from a denser medium to a lighter one is totally internally reflected if $\theta_i \geq \theta_c = \sin(\frac{n_1}{n_2})$, corresponding to an angle of refraction of $\pi/2$, where the refracted ray *fails to escape the medium*.

10.4.2: Total Internal Reflection, Critical Angle

If a ray is travelling from a denser medium to a lighter one, one quickly observes a curious thing. Since the ray is bent *away* from the normal, there exist angles for which Snell's law has no solution!

In fact, it is easy to identify an angle of incidence such that the angle of refraction is $\theta_r = \pi/2$. If we assume that $n_2 > n_1$ and we are going from medium n_2 (the heavier/denser) to medium n_1 (the lighter/less dense):

$$n_2 \sin(\theta_2) = n_2 \sin(\theta_c) = n_1 \sin(\pi/2) = n_1 \tag{10.29}$$

or

$$\theta_c = \sin^{-1}\left(\frac{n_1}{n_2}\right) \tag{10.30}$$

If we increase $\theta_2 > \theta_c$, we make the left hand side of Snell's law *bigger than* n_1 but we cannot find any angle θ_r for which $\sin(\theta_r) > 1$!. We conclude that at all angles θ_c and greater the ray *fails to escape the medium!*

Since it is not absorbed by the interface, and is not transmitted into medium n_1, the only place the energy in this ray can go is into the *reflected* ray. The ray is thus *totally internally reflected*.

Total internal reflection is extremely useful in our modern society. It is the basis of *fiber optics* where (laser) light signals are "trapped" inside a "light pipe" that transmits the light down the fiber and around sufficiently gentle bends without allowing the light to escape through the sides of the optical fibers that have an index of refraction greater than that of the surrounding air or other media.

It is also pretty! Diamonds and the diamond-like compound C3 (Moissonite) have *extremely*

large indices of refraction, roughly $n_d = 2.4$. This makes its critical angle:

$$\theta_{cd} = \sin^{-1}\left(\frac{1}{2.4}\right) = 24.6^\circ \tag{10.31}$$

Light incident on the facet of a diamond at any angle *greater* than this (rather small) angle is *trapped* by the diamond. Diamonds are cut so that light entering through any given facet is reflected many times without escaping, so that dispersion splits the light up into many colors until it escapes either through the sides or at corners or edges. This gives diamond (or Moissanite) its "bright and sparkly" appearance. Cut crystal prisms and lesser clear gemstones have much the same properties on a lesser scale, trapping light and splitting it up into a rainbow of colors to brighten an otherwise drab existence.

10.4.3: Dispersion

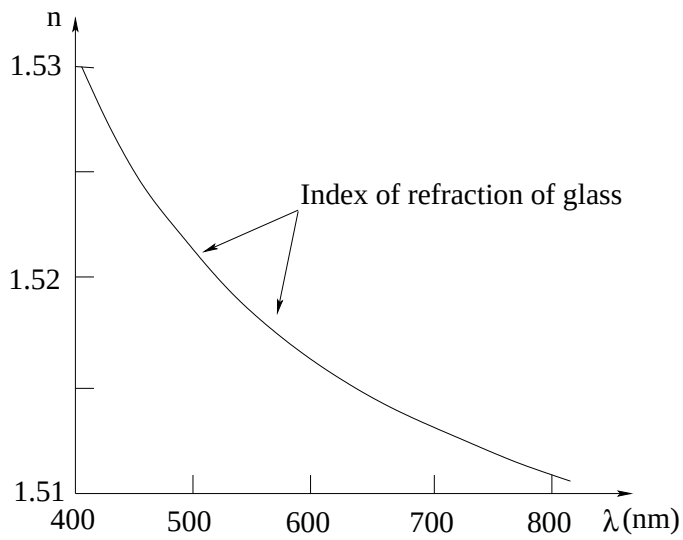


Figure 10.8: An approximate dispersion curve $n(\lambda)$ for "ordinary" glass. However, distinct glass mixtures can have very different dispersion curves, including ones where n *increases* with increasing wavelength λ (decreases with frequency).

To better understand the *colors* produced by diamond, or the colors in a rainbow, or the color band produced from white light by a prism, we have to consider refraction from a medium with ***dispersion***. Dispersion, recall, describes the fact that the index of refraction for most materials isn't really a constant, it *varies* with frequency/wavelength. Most transparent materials have a dispersion in the visible range that ***decreases*** (increases) the index of refraction with ***wavelength*** (frequency). A typical dispersion curve for the kind of glass one might find in a drinking glass or prism is shown across the range of visible wavelengths in figure 10.8. Note well that violet light (400 nm) has an index of refraction that is a percent or two higher than the index of refraction of red light (700 nm).

This is sufficient to cause white light incident at some nonzero angle to *split up* into its distinct component wavelengths in beams that gradually spatially separate as the light travels. The band of colors produced by any given source of incident light, sorted out by wavelength from ***longest to shortest*** is called the ***spectrum*** of the incident light. White light is a mixture

of all visible colors, and its spectrum is the familiar "rainbow" of colors, Red Orange Yellow Green Blue Indigo Violet, or "ROY G BIV" (a common mnemonic for the order). Note well that the frequency order is opposite – from ***smallest to largest***.

One familiar way to get a good spatially separated band of colors is to use two refractive surfaces, each of which helps to further bend the resolved colors – a ***prism***[148].

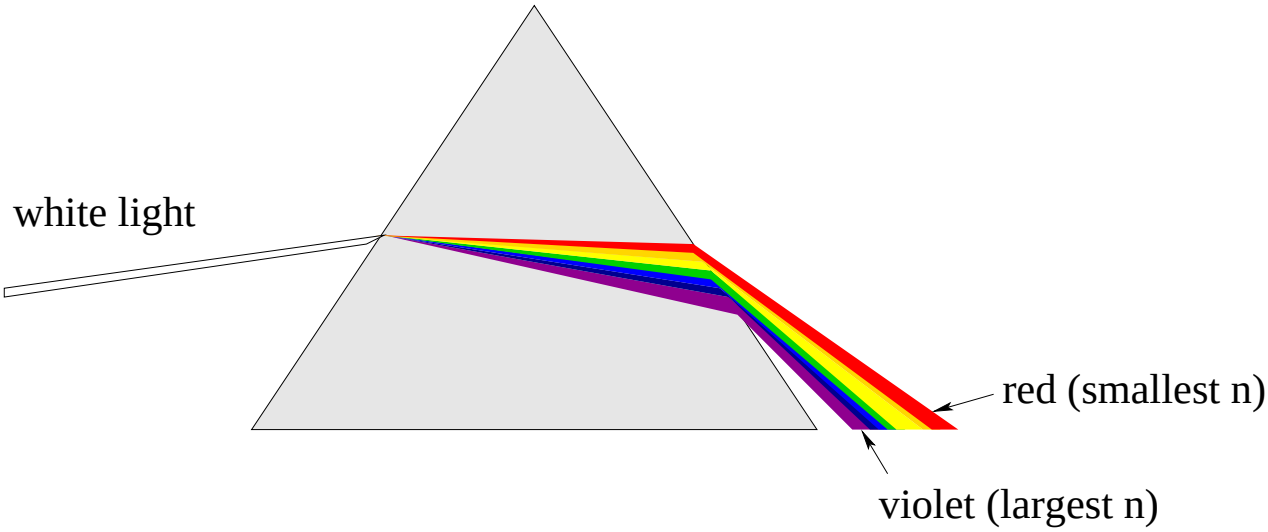


Figure 10.9: A prism causes violet light to be bent more than red light at each interface, splitting up the originally white incident light into a full spectrum.

In figure 10.9 the way a prism acts on an incident white beam of light is crudely represented. Red light, with the smallest n, is bent the least (at each of the two surfaces). Violet light, with the largest n, is bent the most.

Similarly, water droplets or ice crystals that are all roughly the same size can individually preferentially divert different colors of light into different angles, creating a ***ring*** spectrum around a white source seen through e.g. a falling rain. When the white source is sunlight shining through raindrops in the early morning or late evening (so it can come in underneath the raincloud cover) one sees only half of the ring, a *rainbow*[149]. When the white source is sunlight shining through ice crystals in light clouds in the atmosphere, one can get "sunbows", or more rarely, "sun dogs" formed from refracting/reflecting off of planar ice crystals.

10.5: Polarization

As we saw in the last chapter, the electric and magnetic field vectors can point in two independent directions perpendicular to the direction of propagation (the Poynting vector direction). These two directions/orientations are portrayed in figure 10.10. However, we don't need to

[148]This is *yet another* of the discoveries/accomplisments of Isaac Newton. He was the first person to deduce many of the properties of light from the observation that a prism would take an incoming circular ray of white light and transform it into an exiting ellipse of light with the colors of the rainbow.

[149]Or, more rarely, a *double rainbow! All the way across the sky!*

I've never seen a *triple rainbow*, but they too are possible, and I'm guessing an easy way to go viral if you ever capture one in a sappy video...

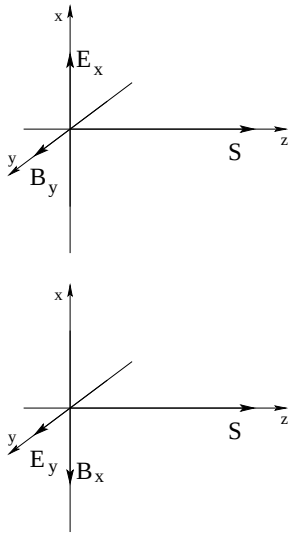


Figure 10.10: The two possible directions for the electric and magnetic vector field components to point corresponding to an electromagnetic wave propagating in the $+\hat{z}$ direction.

define *both* the electric *and* magnetic field components, because if we know (for example) $E_x(z,t)$ then $B_y(z,t) = |E_x(z,t)/c|$ is determined (and ditto for $E_y(z,t)$ and $B_x(z,t)$). As a general rule, then, we will describe the ***polarization*** of an electromagnetic wave with some given frequency and wavelength by fully specifying two independent vector (harmonic travelling wave) components of the *electric* field only that are perpendicular to the direction of propagation.

There are several ways to describe the polarization, and several physical processes produce polarized light, or take *un*polarized light and filter it into some specific polarization state. To understand this, we will start by building an understanding of what is meant by "unpolarized" light (which isn't really unpolarized as in having *no* polarization state, it is just mixed-up light with a *random, fluctuating* polarization state).

10.5.1: Unpolarized Light

Unpolarized light is light for which the polarization vector is constantly shifting its direction around. For a few tens to thousands of wavelengths the electric field vector points in some direction. Then it suddenly shifts into a new direction, as its source gets randomly interrupted. Unpolarized light is typically produced by "hot" or "random" sources such as the Sun, a hot lightbulb filament, the gas in a fluorescent bulb, a candle flame. On average, unpolarized light has its energy/intensity equally distributed between the two independent directions of polarization.

10.5.2: Linear Polarization

Linear polarization occurs whenever the electric field vector oscillates consistently in a single vector direction in the plane perpendicular to propagation. The following are all examples of linearly polarized light propagating in the z-direction with frequency ω:

Light linearly polarized in the x-direction:

$$\vec{\boldsymbol{E}}(z,t) = E_{0x}\hat{\boldsymbol{x}}\sin(kz - \omega t) \tag{10.32}$$

(The associated magnetic field *must* be:

$$\vec{\boldsymbol{B}}(z,t) = B_{0y}\hat{\boldsymbol{y}}\sin(kz - \omega t) = \frac{E_{0x}}{c}\hat{\boldsymbol{y}}\sin(kz - \omega t) \tag{10.33}$$

according to the rules derived in the previous chapter, because

$$|\vec{\boldsymbol{B}}| = \frac{|\vec{\boldsymbol{E}}|}{c} \tag{10.34}$$

and because

$$\hat{\boldsymbol{x}} \times \hat{\boldsymbol{y}} = \hat{\boldsymbol{z}} \tag{10.35}$$

in the Poynting vector.)

Light linearly polarized in the y-direction:

$$\vec{\boldsymbol{E}}(z,t) = E_{0y}\hat{\boldsymbol{y}}\sin(kz - \omega t) \tag{10.36}$$

(The associated magnetic field *must* be:

$$\vec{\boldsymbol{B}}(z,t) = -B_{0x}\hat{\boldsymbol{x}}\sin(kz - \omega t) = -\frac{E_{0y}}{c}\hat{\boldsymbol{x}}\sin(kz - \omega t) \tag{10.37}$$

according to the rules derived in the previous chapter, because

$$\hat{\boldsymbol{y}} \times -\hat{\boldsymbol{x}} = \hat{\boldsymbol{z}} \tag{10.38}$$

in the Poynting vector.)

Finally, light linearly polarized along the line at $\pi/4$ above the x-axis is::

$$\vec{\boldsymbol{E}}(z,t) = \frac{\sqrt{2}}{2}E_0\hat{\boldsymbol{x}}\sin(kz - \omega t) + \frac{\sqrt{2}}{2}E_0\hat{\boldsymbol{y}}\sin(kz - \omega t) \tag{10.39}$$

The amplitude of the electric field is E_0 (why?). What must the direction and magnitude of the associated magnetic field?

Electric dipole radiation is naturally polarized in the plane formed by the axis of the dipole moment $\vec{\boldsymbol{p}}$ (one line) and the vector $\vec{\boldsymbol{r}}$ from the dipole moment to the point of observation (a second, non-colinear line). The electric field component of the propagating light must *also* be completely perpendiculr to the direction of propagation $\hat{\boldsymbol{r}}$! We will use this connection several times below.

10.5.3: Circularly Polarized Light

There is no reason that the magnitudes of the electric polarization components in the two independent directions have to be *the same* or to be *in phase*. We start by considering the case where they have the same magnitude but are $\pi/2$ out of phase:

$$\begin{aligned}\vec{\boldsymbol{E}}(z,t) &= \frac{\sqrt{2}}{2}E_0\hat{\boldsymbol{x}}\sin(kz-\omega t\pm\pi/2)+\frac{\sqrt{2}}{2}E_0\hat{\boldsymbol{y}}\sin(kz-\omega t)\\ \vec{\boldsymbol{E}}(z,t) &= \frac{\sqrt{2}}{2}E_0\left(\pm\hat{\boldsymbol{x}}\cos(kz-\omega t)+\hat{\boldsymbol{y}}\sin(kz-\omega t)\right)\end{aligned} \tag{10.40}$$

These two components describe a vector of constant length that sweeps around in a *circle*, either counterclockwise (-) or clockwise (+). We call this *circularly polarized light*. Note that the two components must have equal amplitudes and must be $\pi/2$ out of phase to be circularly polarized. There are two independent *helicities* of circularly polarized light: right (clockwise/+) and left (anticlockwise/-) when facing *in* the direction of propagation).

Note that we might expect circularly polarized light to be produced by a *rotating electric dipole!*

10.5.4: Elliptically Polarized Light

If the amplitudes of the two waves are (potentially) different *and* the two waves are (potentially) out of phase, the most general polarization state is that of *elliptical* polarization:

$$\vec{\boldsymbol{E}}(z,t) = E_{0x}\hat{\boldsymbol{x}}\sin(kz-\omega t+\delta_x)+E_{0y}\hat{\boldsymbol{y}}\sin(kz-\omega t+\delta_y) \tag{10.41}$$

In this expression, E_{0x} and E_{0y} may or may not be equal, and the phases δ_x and δ_y may or may not be zero *or* equal. The amplitudes of the x and y limits define a rectangular box. The electric field vector rotates within that box wit the box tipped at an angle relative determined by the relative phase difference $\delta = \delta_x - \delta_y$ (where if $\delta = 0$ or $\delta = \pi$ one has linear polarization).

To see a lovely animation of the electric field vector for various flavors of polarization, visit:

http://www.nsm.buffalo.edu/∼jochena/research/opticalactivity.html

10.5.5: Polarization by Absorption (Malus's Law)

A polaroid filter is made by putting oriented conducting threads into a transparent medium in such a way that long currents in those threads created by the polarization component of light parallel to the thread heats the threads, absorbing and attenuating *only* that component of the incident polarized or unpolarized light and passing the component perpendicular to the threads (the **transmission axis** of the filter).

The rules for transmission are simple. If the incident light is unpolarized, on average half its energy is polarized in either polarization direction. Therefore (assuming that the filter is "ideal" and otherwise fully transparent):

$$I_{\text{transmitted}} = \frac{I_{\text{incident}}}{2} \tag{10.42}$$

The transmitted light is fully linearly polarized in the direction of the transmission axis of the filter.

If the light that is incident on the filter is already polarized, then only the *component* of the electric field vector that is *parallel* to the transmission axis is transmitted. That is:

$$E_{\text{transmitted}} = \vec{\boldsymbol{E}} \cdot \hat{\boldsymbol{T}} = E_{\text{incident}} \cos(\theta) \tag{10.43}$$

where θ is the angle between the direction of linear polarization of the incident light and a unit vector along the transmission axis.

To find the transmitted *intensity*, we need just remember the relation between the electric field strength and the intensity that follows from the intensity being the time-average magnitude of the Poynting vector:

$$I = \left| \frac{1}{2\mu_0} \vec{\boldsymbol{E}} \times \vec{\boldsymbol{B}} \right| = \frac{1}{2\mu_0 c} E^2 \tag{10.44}$$

The intensity is directly proportional to the electric field amplitude, squared, so that:

$$I_{\text{transmitted}} = I_{\text{incident}} \cos^2(\theta) \tag{10.45}$$

This result is known as **Malus's law**.

10.5.6: Polarization by Scattering

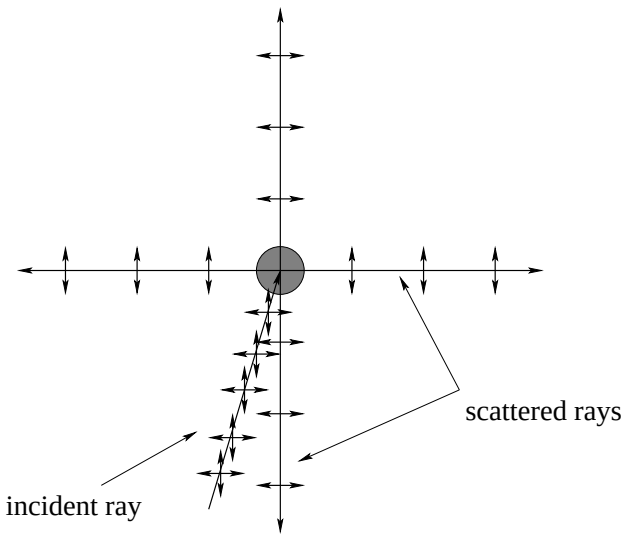


Figure 10.11: The scattering of initially unpolarized light by a molecule or dust particle. Note that the polarization is perpendicular to the *plane of scattering* for each of the possible outgoing directions.

When unpolarized light passes across an atom or molecule, it *polarizes* it in the instantaneous direction of the electric field vector (which, recall, has a definite direction at any time but which jumps around to a new direction every 10-1000 optical periods). The oscillating molecule acts like a *dipole antenna* and *reradiates* the incident electromagnetic wave. However, the reradiated electric field must be *parallel* to the dipole moment of the molecule, and there is no radiation *along* the dipole (with a clear maximum at right angles to the dipole. As a consequence we can easily see that the rule for polarization of rays scattered more or less at right angles is that they must be polarized *perpendicular to the plane of scattering!*

10.5.7: Polarization by Reflection

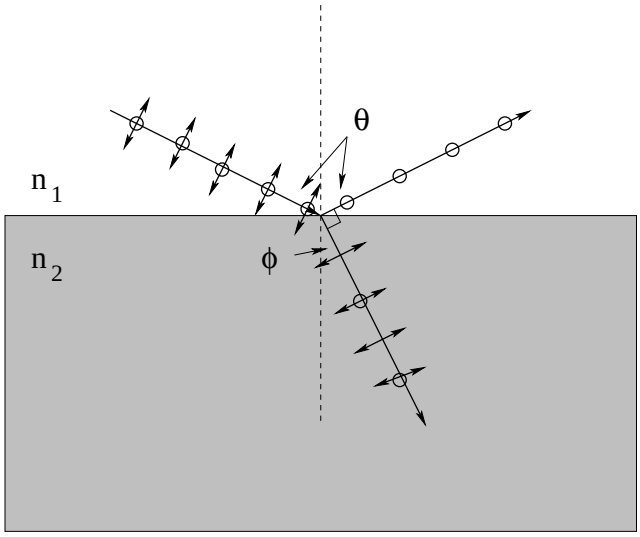


Figure 10.12: The scattering of initially unpolarized light by reflection off of a plane surface between two dielectric media at the *Brewster angle* that produces complete polarization of the reflected ray. Note that the polarization of all reflected rays incident on the surface at an angle is *parallel to the ground* even at angles other than the Brewster angle.

When light strikes a surface between two regions with differing indices of refraction, it is partially transmitted and partially reflected (with the amount of each determined by the angle of incidence and the two indices of refraction). The reflection is caused by the polarization of surface molecules in such a way that the light scattered by them adds up coherently into the reflected wave; similarly those polarized molecules create a forward propagating wave into the medium (although at a different angle according to Snell's law). As before, the polarized surface molecules (dipoles) *cannot radiate along their own axis* so that light that is reflected *parallel* to one of the polarization directions cannot contain that polarization.

This state of affairs occurs when the reflected ray is perpendicular to the refracted ray, pictured above. In this case:

$$n_1 \sin(\theta) = n_2 \sin(\phi) \tag{10.46}$$

is Snell's law, but clearly:

$$\phi = \frac{\pi}{2} - \theta \tag{10.47}$$

so that:

$$\sin(\phi) = \sin(\pi/2 - \theta) = \cos(\theta) \tag{10.48}$$

and *Brewster's formula*:

$$\tan(\theta_b) = \frac{n_2}{n_1} \tag{10.49}$$

is the condition for θ_b, the so-called *Brewster angle* of incidence (and hence reflection) where the reflected ray is completely polarized parallel to the surface (and perpendicular to the plane of reflection, just as was the case with scattered light above).

However, the polarization component in the plane of reflection is *always* reduced at angles other than $\theta = 0$ as the component of the polarization gradually lines up with the reflected ray

so reflected light is at least *partially* polarized in the plane at all angles other than 0. Note that the transmitted light is *partially* polarized *in* the plane of transmission – this is not complete because all of the perpendicularly polarized light is not reflected at the surface, some is still transmitted into the medium.

10.5.8: Polaroid Sunglasses

As we have just seen, reflected glare from any smooth surface is likely to be at least partially polarized parallel to the ground. It is thus blocked by a pair of polaroid sunglasses with a *vertical* transmission axis. Similarly, (scattered) light from the blue sky viewed near the horizon at midday is predominantly polarized parallel to the ground and is *also* blocked by a vertical transmission axis, which can make e.g. driving safer and less stressful on the eye.

10.6: Doppler Shift

Since light is a wave, the frequencies picked up by a frequency sensitive receiver (e.g. the human eye) depend on the original frequency (color) emitted by the source and *Doppler shifted* by the motion of the source and/or the receiver. A complete treatment of the Doppler shift requires relativity and is beyond the scope of this course, but an elementary treatment suffices to understand the Doppler shift at ***velocities that are small compared to the speed of light***[150].

The idea underlying the Doppler shift is very simple. If the source is moving towards the receiver, its motion foreshortens the normal wavelength, increasing the frequency observed by the stationary receiver. If the receiver is moving towards the source, its motion reduces the time between the wavefronts it receives, increasing the frequency it observes. If both motions are occurring, both shifts occur as a product. We show the picture and quick derivation of each possibility below.

10.6.1: Moving Source

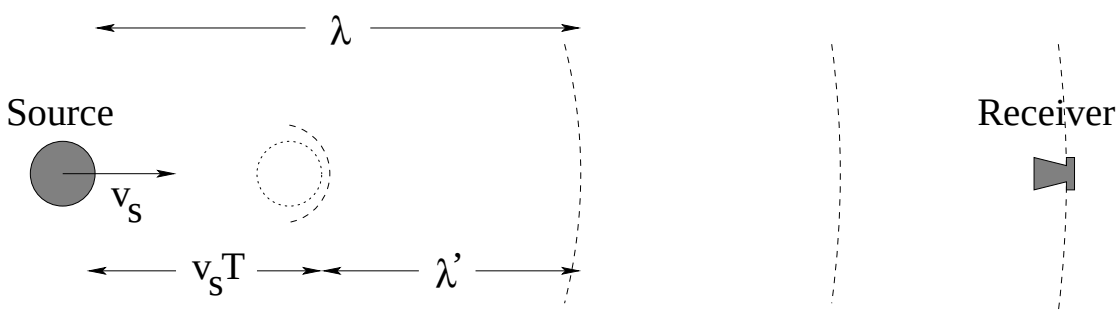


Figure 10.13: Wave geometry for Doppler shift of moving source.

[150]At higher speeds, lengths contract and times dilate, so this simple argument has to be made a bit more complicated. In this case the correct argument leads to the formula for the *relativistic Doppler shift* for moving source and/or receiver, but at low speeds the forms for the shifts are approximately (to lowest nontrivial order in v/c) the same

The source emits light waves that travel a distance $\lambda = cT$ in a single period T. However, in the time T between wavefronts, the source moving at speed v_s towards the receiver travels *in* to the wave it has emitted a distance v_sT, reducing the distance at the time of the next front to $\lambda' = \lambda - v_sT$. This in turn reduces the time T' between wavefronts that cross the receiver (e.g. an eye or camera) and hence we can solve for the frequency shift thus:

$$\begin{aligned}
\lambda' &= \lambda - v_sT \\
cT' &= cT - v_sT \\
T' &= T\left(1 - \frac{v_s}{c}\right) \\
\frac{1}{T'} &= \frac{1}{T}\frac{1}{\left(1 - \frac{v_s}{c}\right)} \\
f' &= \frac{f}{\left(1 - \frac{v_s}{c}\right)}
\end{aligned} \tag{10.50}$$

For a source moving *away* from the receiver the algebra and picture is the same, but the wavelength $\lambda' = \lambda + v_sT$ is *increased*, so that:

$$f' = \frac{f}{\left(1 \mp \frac{v_s}{c}\right)} \tag{10.51}$$

for an approaching (-) or receding (+) source describes the general moving source doppler shift in the frequency/color detected by the receiver.

Note well that visible light sources moving away from the receiver are shifted towards the *red* end of the spectrum, while sources moving towards the receiver are shifted towards the *violet* end of the spectrum. Since spectral lines produced by atoms have sharp and well-defined frequencies, this permits us to ascertain that the visible Universe is *expanding* (as all distant stars and galaxies are red-shifted). Since the velocity with which distant stars are receding from the Earth *increases with distance*, the red shift becomes a *meter stick* permitting us to measure the size of the visible Cosmos. This is a small but significant part of the physical evidence for the *Big Bang* cosmological model that so far seems best to fit the data, and that suggests that the Big Bang occurred approximately 13.5 billion years ago (give or take a billion years) so that the visible Cosmos is a sphere roughly 27 billion light years across, containing roughly a trillion galaxies containing order of a trillion stars apiece. This is around Avogadro's number of *stars*.

With no boundaries visible in any direction, there is no particular reason for us to think that we are in the exact center of the cosmos, save in the sense that every point is in the middle of an infinite line. Sometimes small pieces of physics (such as the Doppler shift of light) can have *enormous* consequences.

10.6.2: Moving Receiver

If a frequency-sensitive detector of light (such as the eye or a camera) is moving *towards* a fixed source at speed v_r, it moves into a wave that is travelling at the speed of light and "meets the oncoming wavefront half way" (not literally half way) *sooner* than it would have if it were at rest. This shortened period T' can easily be determined from the geometry above, where

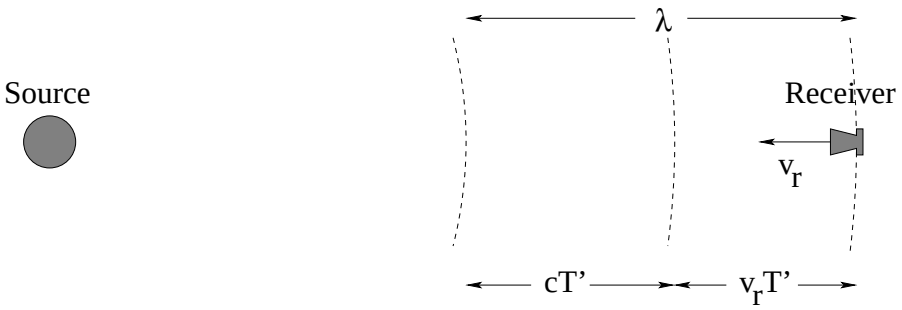


Figure 10.14: Wave geometry for Doppler shift of a moving receiver.

$\lambda = cT = (c + v_r)T'$:

$$\begin{aligned} cT &= (c + v_r)T' \\ T &= (1 + \frac{v_r}{c})T' \\ \frac{1}{T'} &= \frac{1}{T}(1 + \frac{v_r}{c}) \\ f' &= f(1 + \frac{v_r}{c}) \end{aligned} \tag{10.52}$$

As before, if the receiver is moving away, it decreases f' instead of increasing it, so that the general rule is:

$$f' = f(1 \pm \frac{v_r}{c}) \tag{10.53}$$

for a receiver moving towards (+) or away from (-) the source.

10.6.3: Moving Source and Moving Receiver

The rule is just the product of the two rules:

$$f' = f\frac{(1 \pm \frac{v_r}{c})}{(1 \mp \frac{v_s}{c})} \tag{10.54}$$

It is interesting to note that if a source is moving at the speed of light (where these expressions are no longer valid, alas, see below) the frequency f' goes to *infinity*. This divergence is the moral equivalent of a *sonic boom*, only with *light* instead of sound! Although particles cannot go faster than light in a vacuum, charged particles *can* go faster than the speed of light inside a medium with an index of refraction $n > 1$.

Consider an electron travelling at 0.99c and entering a piece of glass where the speed of light is only approximately 0.67c. The ”light boom” given off by the superluminal charged particle as it interacts with the charges in the glass is *clearly visible experimentally* and is called *Cerenkov radiation*. Cerenkov radiation is the basis of some of the high-energy particle detectors used in many of the big accelerator laboratories in high energy nuclear physics.

10.6.4: The Relativistic Doppler Shift

It is beyond the scope of this course to actually *derive* the full relativistically correct doppler shift for light waves. It differs from the one derived above because when objects are moving

fast enough, one must use ***special relativity*** to shift time intervals measured by the receiver *relative* to the source. This "relativistic" shift in the clocks attached to source and receiver actually occurs for other waves, e.g. sound as well, but when $v \ll c$ the time shift is negligible.

It is still worth looking at the result, as it is interesting in and of itself. It is called the ***Fizeau-Doppler Formula***

$$f' = \sqrt{\frac{1 \pm \frac{v}{c}}{1 \mp \frac{v}{c}}} f_0 = \sqrt{\frac{1 \pm \beta}{1 \mp \beta}} f_0$$

where we follow a common practice and define the dimensionless ***Lorentz Factor***:

$$\beta = \pm \frac{v}{c}$$

where v is the *relative* velocity of source and receiver, not the *absolute* velocity of either one in e.g. the lab frame, as in relativity theory no such thing as absolute velocity exists except for the speed of light itself. β is , positive if the source and receiver are approaching one another, negative if receding. f_0 is the frequency emitted *in the reference frame of the source* and f' is the frequency measured *in the reference frame of the receiver*.

Note well that we can no longer assert that either the source f_0 or the receiver f' are the "absolute" source frequency common to *all* frames, as time itself will be measured differently in any given inertial lab frame compared to a clock in the frame of a source or receiver moving in that frame, *independent* of the usual arguments for a moving source doppler shift!

Relativity (like Quantum Mechanics) is wild and wonderful and makes your head explode the first time you work through it, and is typically covered in a third course: Introductory Modern Physics (basically, physics as developed in the period just after and slightly overlapping the discovery of the full set of Maxwell's Equations and their relationship to light and "the present". This University-level course is typically taken by physics majors and students of other disciplines who want a glimpse "behind the curtain" of classical physics, to gain a foundational knowledge of the currently accepted beliefs about how the world *really* works and the *philosophically* interesting histor of the discoveries and experiments that force these two true paradigm shifts in the scientific worldview. ***All*** of the results we learn in the standard two-semester introduction to *classical* physics have to be reconsidered and turn out to be technically incorrect, the limit of the relativistic quantum world when considering objects that are "large" (relative to e.g. the atomic scale) and moving "slowly" (relative to light).

Don't get me wrong – classical physics works *very well indeed* for things like baseballs and most of large-scale electrodynamics – but if you apply Newton's Laws and the Galilean transformation of inertial reference frames as we've learned them to atoms, or to particles moving at speeds at all comparable to c, they *just don't work!*

The one last thing we'll do before moving on to Geometric Optics is to show that the Fizeau formula is equivalent to the non-relativistic Doppler formula in the limit that $\beta \ll 1$. Then (to leading/linear order in β) we use the binomial expansion as follows (note the step multiplying

the Fizeau formula by *one* in a form that eliminates the square root in the *numerator*:

$$\begin{aligned}
f' &= \left(\frac{(1\pm\beta)^{\frac{1}{2}}}{(1\mp\beta)^{\frac{1}{2}}}\right) f_0 \\
&= \left(1 = \frac{(1\pm\beta)^{\frac{1}{2}}}{(1\pm\beta)^{\frac{1}{2}}}\right) \times \left(\frac{(1\pm\beta)^{\frac{1}{2}}}{(1\mp\beta)^{\frac{1}{2}}}\right) f_0 \\
&= \frac{1\pm\beta}{(1-\beta^2)^{\frac{1}{2}}} f_0 \approx (1\pm\beta)(1+\cancel{\frac{1}{2}\beta^2}+...) f_0 \approx (1\pm\beta) f_0
\end{aligned}$$

This formula corresponds to the usual moving receiver, stationary source shift!

Similarly:

$$\begin{aligned}
f' &= \left(\frac{(1\pm\beta)^{\frac{1}{2}}}{(1\mp\beta)^{\frac{1}{2}}}\right) f_0 \\
&= \left(1 = \frac{(1\mp\beta)^{\frac{1}{2}}}{(1\mp\beta)^{\frac{1}{2}}}\right) \times \left(\frac{(1\pm\beta)^{\frac{1}{2}}}{(1\mp\beta)^{\frac{1}{2}}}\right) f_0 \\
&= \frac{(1-\beta^2)^{\frac{1}{2}}}{(1\mp\beta)} f_0 \approx (1-\cancel{\frac{1}{2}\beta^2}+...)\frac{1}{(1\mp\beta)} f_0 \approx \frac{1}{(1\pm\beta)} f_0
\end{aligned}$$

where we eliminated the square root in the denominator instead. This is usual moving source, stationary receiver result!

Note well that v is the *relative* speed, not the lab speed in both cases. It isn't entirely obvious that the two can still be combined into a single formula with two *distinct* source and receiver speeds, but since the shift in time relative to the lab is negligible (order v^2/c^2 and higher) in this limit, it turns out that we can shift both independently into the lab frame and it can, which is why our non-relativistic result above works for *both* light *and* sound waves. We'll defer any treatment of this, however, until that third course!

Homework for Week 10

Problem 1.

Physics Concepts

Make this week's physics concepts summary as you work all of the problems in this week's assignment. Be sure to cross-reference each concept in the summary to the problem(s) they were key to. Do the work carefully enough that you can (after it has been handed in and graded) punch it and add it to a three ring binder for review and study come finals!

Problem 2.

Derive Snell's Law. There are (at least) two ways to do so that were covered in the textbook and/or class, and either one is OK.

Problem 3.

Derive the non-relativistic Doppler Shift:

$$f' = f_0 \left(\frac{1 \pm \frac{v_r}{c}}{1 \mp \frac{v_s}{c}} \right)$$

for light sources or receivers moving "slowly" (at speeds $v \ll c$) in a vacuum, where the upper signs in both case refer to approach and the lower signs recession. Although the doppler shift is medically important in the context of sonography (sound) instead of light, the doppler shift of the frequency of light from moving sources is the basis of many everyday technologies: in doppler radar (used to detect the structure of violent storms), in the radar guns used to detect speeders by highway police, and to determine the size of the expanding Universe.

Problem 4.

Derive Malus' Law $I_t = I_0 \cos^2(\theta)$ where I_0 is the intensity of polarized light incident on a polarizing filter at an angle θ relative to the transmission axis of the filter. I'd suggest going back to the Poynting vector and expressing the intensity I_0 in terms of E_0, the E-field amplitude of the incident polarized wave.

Problem 5.

Derive Brewster's Formula (the expression for the angle of incidence for which reflected

light is completely polarized parallel to the surface).

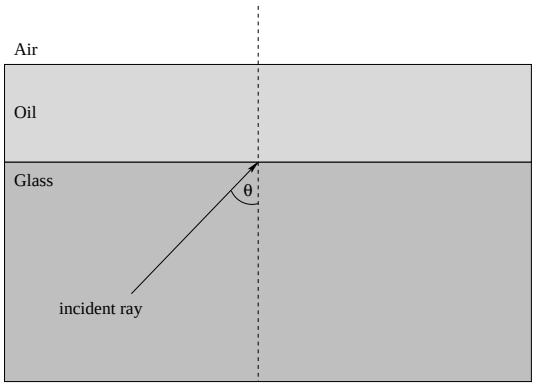
Problem 6.

Draw the figure representing ***polarization by scattering*** and provide a short explanation for why scattered light is generally polarized perpendicular to the plane of scattering.

Problem 7.

Derive the expression for the critical angle leading to total internal reflection for rays moving from a *dense* medium (high n) to a *lighter* one (with lower n). Evaluate the critical angle for water ($n_w = 4/3$) to air ($n_a = 1$) and glass ($n_g = 3/2$) to air.

Problem 8.



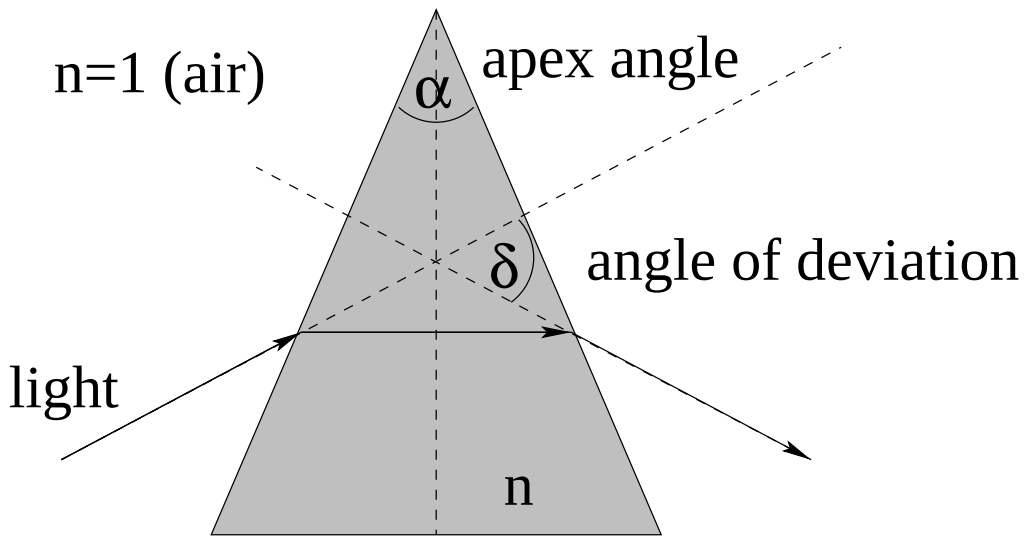
Suppose a layer of oil $n_o = 5/4$ open to the air above ($n_a = 1$) is on top of a a piece of glass $n_g = 3/2$ as shown above. Show that the critical angle for the glass in air alone is not changed by the oil; rays incident on the oil-air interface produced by rays incident on the glass-oil interface at or above the critical angle for glass-air alone do not escape the final layer of oil.

Problem 9.

Show that in spite of the occurrence of total internal reflection, one can *in principle* still see all of bottom in a shallow lake stretched out before your feet. That is, although some rays of light from a fish on the bottom are trapped and escape, there are others that will reach your eye no matter where your eye is located. (Other factors – ripples, reflections off of the surface, murkiness in the water – may limit your vision, but it isn't that any part of the bottom is *theoretically* invisible because light from there cannot escape to reach your eye, it is that the light that does reach them may be very faint and difficult to resolve from other things going on.)

Note that the "answer" to this question is likely to be a diagram or figure that illustrates the answer, not algebra per se, although one can always support the answer further with algebra.

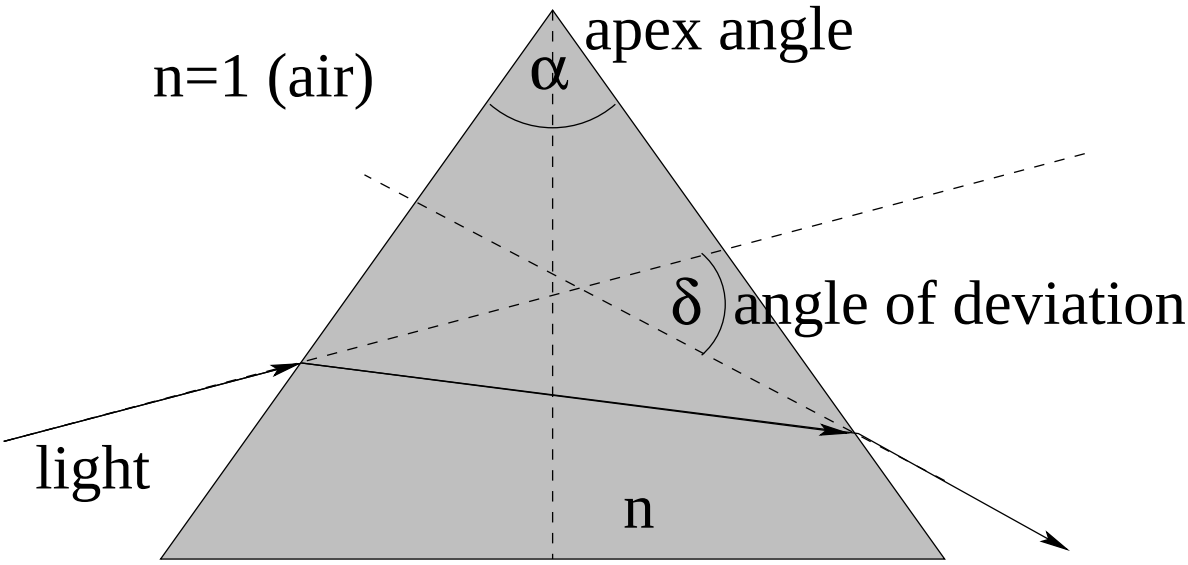
Problem 10.



In the figure above, a beam of light is incident from air onto a prism with an ***apex angle*** α. Its angle of incidence is adjusted until it refracts ***symmetrically*** across the prism, with the ray crossing the vertical bisector of the prism at right angles. Prove that the ***angle of deviation***, δ, is related to α and n by:

$$\sin\{(\alpha+\delta)/2\} = n\sin(\alpha/2)$$

Advanced Problem 11.



Prove that the ***angle of deviation***, δ, is a ***minimum*** when the light ray crosses the vertical bisector at right angles so that the figure has full reflection symmetry if one reverses the direction of the ray (as portrayed in the previous problem).

Week 11: Lenses and Mirrors

- The distance from a mirror (or lens) to an object one is viewing in (or through) it is s, the **object distance**. Object distances are positive if the object is on the side of the mirror (or lens) that the light is coming *from*. Object distances are obviously 'always' positive, unless the object is a *virtual object* formed out of the image of a previous mirror or lens, which can be either positive or negative.

- The distance from a lens or mirror to the image one is viewing is s', the **image distance**. Image distances are positive if the image is on the side of the mirror (or lens) that the light is going *to*.

- The focal length f of a mirror (or lens) is the point where incident parallel rays are focused **to** (for positive focal lengths) or appear to be defocused **from** (for negative focal lengths). f is typically measured in meters (SI) or centimeters (for convenience). However, the strength of *lenses* is usually given in *diopters*, where:

$$d = \frac{1}{f} \tag{11.1}$$

  with f in meters. This a one diopter (1.00d) lens has a focal length of 1 meter. A 10.00d lens has a focal length of 0.1 meter. A diverging lens with a focal length of one centimeter is -100.00d.

- The mirror (or thin lens) equation relating s, s', and f is:

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f} \tag{11.2}$$

- The transverse magnification of a simple mirror (or lens) is defined by the ratio of the image height y' to the object height y:

$$m = \frac{y'}{y} = -\frac{s'}{s} \tag{11.3}$$

- A **real image** is one where the rays of light that appear to the eye to diverge from a point on the image actually pass through that point. A **virtual image** is one where the rays of light that appear to the eye to diverge from a point on the image do *not* actually pass through the image.

- In addition to being real or virtual, an image can be **erect** (oriented the same way as the object) or **inverted** (oriented the opposite way from the object.

- For a spherical mirror, the focal length is given by:

$$f = \frac{r}{2} \tag{11.4}$$

  where r is positive when it is on the side of the mirror reflected light is going *to*.

- For a thin lens, the focal length is given by the **lensmaker's formula**:

$$\frac{1}{f} = (n_2 - n_1)\left(\frac{1}{r_1} - \frac{1}{r_2}\right) \tag{11.5}$$

  In this expression, n_1 is the index of the surrounding medium (typically air, $n_1 = 1$) and n_2 is the index of refraction of the lens itself. r_1 (r_2) is the radius of curvature of the *first (second) surface struck* by the ray, with the sign convention that it is positive (negative) on the side of the lens refracted light is going *to* (coming *from*).

  The advantage of using diopters as a measure of lens strength is inherent in this expression, as you can see that the combined strength of the two lensing surfaces (in diopters) is equal to the *sum* of the strength of *each* surface, in diopters. This extends to any pair of lenses placed close together – the effective strength of two lenses closely placed (relative to their focal lengths) in front of one another is the sum of their strength in diopters.

- True Facts about the Eye:

  The eye is approximately one inch in diameter. A *lens* in front casts a *real* image of objects being viewed onto its *retina*, where rods and cones transform the light into neural impulses which are then conveyed to the brain for processing by the optic nerve. Rods and cones are very sensitive to light (and easily damaged) – the light content is regulated by the *iris* of the eye, which expands and contracts the *pupil* – the aperture through which light passes as it enters the lens.

  The focal length of a relaxed lens of an eye with *normal* vision is on the retina, so distant objects are automatically in focus. Given the diameter of the eye, this means that the strength of the lens of a normal eye is approximately 40.00d. The focal length of a relaxed *farsighted* eye is *behind* the retina (too long, strength less than 40.00d) and is corrected with a *converging* lens to make up the difference. The focal length of a relaxed *nearsighted* eye is in *front* of the retina (too short, strength greater than 40.00d) and is corrected with a *diverging* lens to take away some of its strength.

  There are muscles that surround the lens of the eye in a ring that contract, making the lens bulge (to a greater radius of curvature) and thereby *shortening* the focal length (a process called *accommodation*) to bring nearby objects into focus. The nearest point one can bring an object to the eye and still bring it into focus on the retina is called the *near point* of the eye and is also the *distance of most distinct vision*, represented x_{np}. In most adults, this distance is around 25 cm (less for small children, longer for the elderly).

  A nearsighted person's lens *already* has too short a focal length to be able to focus distant objects on the retina, and accommodation only shortens the focal length still farther. A nearsighted person cannot see anything clearly at distances *greater* than some point, called the *far point* for that person's eyes. A nearsighted person is one for whom the far point x_{fp} is less than infinity.

- The simple magnifier is a converging ($f > 0$) lens placed immediately in front of the eye. An object placed at its focal point therefore forms a virtual image at infinity that is automatically brought into focus by the relaxed normal (or vision corrected) eye. The magnification of the object occurs because one can bring the object *closer* to the eye than x_{np} and still see it clearly, where it subtends a *greater* angle on the retina (angular magnification). Its magnification is given by:

$$M = \frac{x_{np}}{f} \tag{11.6}$$

  It is very important to understand the simple magnifier, as it forms the eyepiece of *both* the microscope *and* the telescope.

- A telescope is used to view a distant object by making the angle its image subtends on the retina larger. Two lenses are situated at ends of a tube such that their focal points are coincident. The first lens (with a long focal length) forms a *real image* of the distant object more or less at its focal point. The second lens (with a short focal length) is used to view this real image as a simple magnifier. This produces a virtual image at infinity that subtends a greater angle than the original object did, viewable with the relaxed normal eye.

  The overall angular magnification of a telescope is given by:

$$M = -\frac{f_o}{f_e} \tag{11.7}$$

  The eyepiece lens can be converging (regular) or diverging (Galilean). In both cases this formula for the magnification works (provided that one uses a negative f_e for the diverging lens and place the focal point f_o at the focal point on the *far* side of the diverging lens). A regular telescope inverts the image, which is inconvenient and undesireable. A Galilean telescope does not invert the image.

- A compound microscope is used to view a very small, but nearby object. It accomplishes this in two stages. Two short focal length lenses are situated at ends of a tube much longer tube. The *tube length* ℓ of the microscope is by definition the distance between the focal point of the first, or *objective* lens (which must be converging) and the second, or *eyepiece* lens. The object is placed just outside of the focal length of the objective lens in such a way that it forms a *magnified, real image* of the object more or less at the end of the tube length. The eyepiece lens is used as a simple magnifier to view this real image, and can be converging or diverging as was the case for the telescope. It produces a virtual image at infinity that subtends a greater angle than the real image formed by the objective lens alone would if viewed at the near point of the relaxed normal eye.

  The magnification of the objective is:

$$M_o = -\frac{\ell}{f_o} \tag{11.8}$$

  The magnification of the eyepiece (simple magnifier) is:

$$M_e = \frac{x_{np}}{f_e} \tag{11.9}$$

The overall magnification is therefore:

$$M_{tot} = -\frac{\ell\, x_{np}}{f_o f_e} \tag{11.10}$$

where as before, this formula for the magnification works provided that one uses a negative f_e for the diverging lens and place the real image formed by the objective on the *far* side of the diverging lens. A regular microscope inverts the image, which is inconvenient and undesireable. A "Galilean" microscope does not invert the image.

11.1: Vision and Plane Mirrors

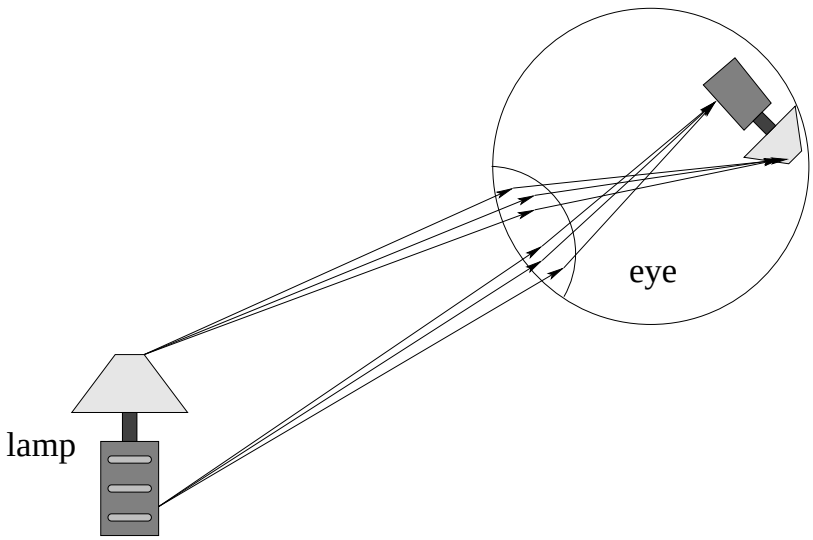


Figure 11.1: How the eye sees an object. Light diverging from points on the surface of the object are focused onto the retina of the eye, where they form an *image* of the object that the retina converts into neural impulses and your brain converts into perception.

Objects in the real world that are illuminated by diffuse light absorb the light at every point on their surface and then reradiate (selected colors/frequencies) from each point in all directions. This is why you can see something that is illuminated from all angles – every point on its surface emits light reradiated from the illuminating source in all directions so no matter where you look at it from, some of the light reaches your eye.

To *completely* understand how your eye can see the object, we have to get halfway through this week's work. On the other hand, we can't understand enough about how mirrors and lenses work to understand the eye without understanding the eye well enough to understand how lenses and mirrors work.

Hmmm, a bit of a dilemma. We have to *bootstrap* just a bit and draw a few pictures now that you won't completely understand later to help you understand what you need to understand what you need to understand later. Or something like that.

So meditate on the picture above, which shows light diffusely scattered from from a couple of points on a common object. The light goes in *all directions* from *all of the points on the surface of the object.* Some of these rays reach your eye. There the lens of your eye does

its thing, and forms a nice sharp *image* of the object cast upon the retina of the eye. Vision occurs.

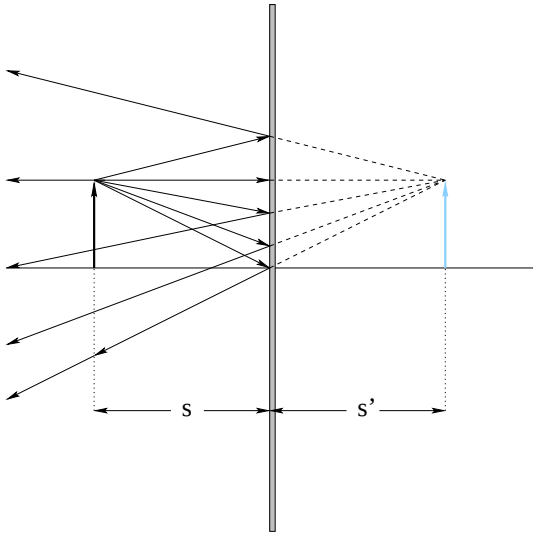


Figure 11.2: The geometry of forming an image in a plane mirror.

Now consider looking at an object in a *plane mirror*. Lamps are too hard to draw, so we consider an arrow, which we will use as a "generic object" in our diagrams.

Rays radiated from the object radiate out in all directions as shown in the figure above. When they strike the mirror they are reflected with the angle of incidence equal to the angle of reflection. As we look at the mirror, we see the rays that originated on a single point on the object *as if* they were diverging from a single point in space. That point is the *image* of the point on the object. Since every (visible) point on the object corresponds to an apparent point of divergence in space from the image, we can see the image *exactly as if* we were looking at an object.

In the case of a plane mirror (above) the image is always *behind* the mirror. The light rays you see do not actually pass through the image, they simply appear to diverge from it. We call such an image a *virtual* image.

We need to define several quantities that will be essential in our analysis of how lenses and mirrors work. The distance from a mirror (or lens) to an object one is viewing in (or through) it is s, the *object distance*. Object distances are *positive* if the object is on the side of the mirror (or lens) that the light is coming *from*. Object distances are obviously 'always' positive, unless the object is a *virtual object* formed out of the image of a previous mirror or lens, which can be either positive or negative.

The distance from a lens or mirror to the image one is viewing is s', the *image distance*. Image distances are *positive* if the image is on the side of the mirror (or lens) that the light is going *to*.

Multiple mirrors can be used to create images of images, or images of images of images (used as "virtual objects" for the second mirror). Most of us have experienced the "infinite tunnel" of images that results from standing directly in between two plane mirrors.

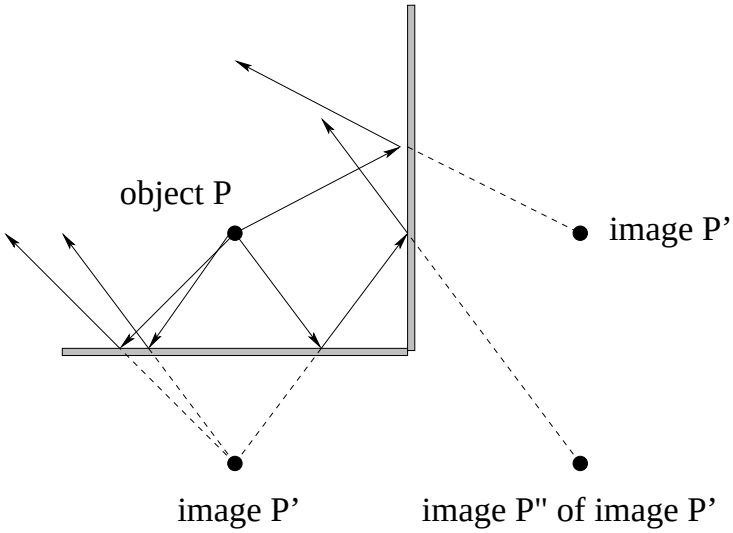


Figure 11.3: Two mirrors create an image of an image. Only a few of the many rays are drawn – copy the picture and fill in more yourself.

11.2: Curved Mirrors

Plane mirrors simply create a perfect image of everything that is in the real space reflected in the mirror. Things get more interesting if the mirrors are *curved*. Curved mirrors can create images that are systematically larger or smaller than the object, and can create a new kind of image from the one seen in figure (11.2).

In figure (11.4) we see a *concave spherical mirror*, which we will also call a *converging* mirror or a *positive* mirror[151]. The horizontal line running through the center of the mirror is very important and is called the *axis* of the mirror, which is rotationally symmetric about this axis. Even imaging an arrow is too complicated for our purpose (which is to figure out how spherical mirrors can make images at all) so we look for the image of a *single point* P, which we locate for convenience on the axis of the mirror.

The image P' occurs where two reflected rays *cross*. The two rays in question are the one that strikes a distance l up the mirror (with angle of incidence equal to the angle of reflection) and a ray that goes along the axis and is reflected directly back the way it came. This is a new kind of image – the rays don't just *appear* to come from a point in space (a point that is really in the dark of your closet or medicine cabinet, back behind the mirror) as they do with a virtual image, they *really* reach the eye after passing through a point in space. You could reach out and put your finger through the point in space they appear to be coming from. We call this kind of image a **real image**, and we need to be able to determine whether an image is real (the kind of image that can be projected on a retina, piece of film, wall, projector screen) or virtual (which cannot be projected at all, since no light actually passes through the image), so be sure you understand the distinction and can categorize images you determine from e.g. ray diagrams.

We begin by making an essential approximation. We will later talk about *aberrations* of

[151] For those who have concave/convex dyslexia, remember that con*cave* is like a cave, and curves inward, while con*vex* is nothing at all like a vex. What is a vex, anyway?

lenses and mirrors – things that prevent rays from a single point on the object from d. One of the most important ones will be *spherical* aberration – spheres have this annoying habit of not focussing parallel rays from an object point far from the axis or rays that are near the axis but that are not approximately parallel to the axis down to a single point in the image. We can't have that, so we insist that the rays we will deal with be *paraxial* – close to the axis and close to parallel. The former means that we strike the mirror close enough to its center for us to be able to pretend that the deflection occurs in a (slightly) curved plane; the latter means that small angle approximations will all work quite well.

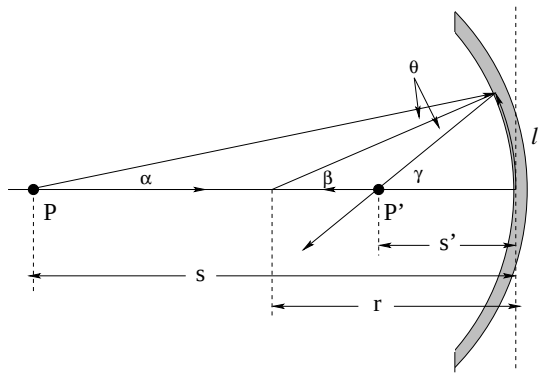


Figure 11.4: The geometry of forming an image in a concave mirror.

Three important lengths are drawn onto the figure: s, s', and r, as well as the distance l itself. Note well also the four angles: α, β, γ and the angle of incidence/reflection θ. Since the angles are all small and l is close to a straight line:

$$\alpha \approx \frac{l}{s} \tag{11.11}$$

$$\beta = \frac{l}{s} \tag{11.12}$$

$$\gamma \approx \frac{l}{s'} \tag{11.13}$$

(where the result for β, note well, is exact because l really is the length of a circular arc that is subtended by the angle β).

We now play games with the triangles in the picture. We use the following rule several times: Consider the triangle with α, θ and the angle δ (filled in to figure (11.5)). We can easily

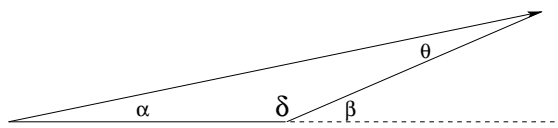


Figure 11.5: $\alpha + \theta = \beta$.

see that $\alpha + \theta + \delta = \pi$. But we can *also* see that $\delta + \beta = \pi$. Therefore:

$$\alpha + \theta = \beta \tag{11.14}$$

and similarly (considering the other triangle involving β and θ)

$$\beta + \theta = \gamma \tag{11.15}$$

If we eliminate θ, we get:

$$\alpha + \gamma = 2\beta \tag{11.16}$$

Finally, if we substitute in all of the small angle approximations and cancel l, we get:

$$\frac{1}{s} + \frac{1}{s} = \frac{2}{r} \tag{11.17}$$

As we move the object back farther and farther from the mirror (let $s \to \infty$) we note that the image distance approaches $r/2$. Rays coming from an infinitely distant object arrive at the mirror *parallel* and converge *at* $s' = r/2$. We *define* the point where a lens or mirror focuses *parallel, paraxial rays* to be the **focal point** of the lens or mirror. Thus:

$$f = \frac{r}{2} \tag{11.18}$$

and

$$\frac{1}{s} + \frac{1}{s} = \frac{1}{f} \tag{11.19}$$

This is a *very important result!* It is the equation we will use to analyze *all images formed by curved mirrors and thin lenses* (after we derive the same formula for the latter) so be sure that you have learned it and understand it.

The focal length f of a mirror (or, soon, thin lens) is the point where incident parallel rays are focused **to** (for positive focal lengths) or appear to be defocused **from** (for negative focal lengths). f is typically measured in meters (SI) or centimeters (for convenience). You may have observed that the stronger (more curved!) a mirror is, the *smaller* its focal length is. We might like to invent a quantity that expresses this strength more intuitively, so that a larger value of the quantity corresponds to a more strongly acting merror (or, shortly, lens).

The simplest way to accomplish this goal is to express the strength of a mirror by the *inverse* of its focal length, a quantity called its ***power*** (symbol P), given in new (SI) units called *diopters* (D). That is:

$$P = \frac{1}{f} \tag{11.20}$$

with f in meters. Thus a one diopter (1.00D) lens or mirror has a focal length of 1 meter. 10.00D corresponds to has a focal length of 0.1 meter. A *diverging* lens or mirror with a focal length of one centimeter is -100.00D.

Note that now the relation between power and the effect of the mirror is much more intuitive: $P = 0$ describe a ***flat*** mirror that doesn't magnify or shrink the image at all. This is much better than describing such a mirror with "$f = \infty$". There are other advantages to power expressed in diopters – so much so that we'll spend an entire section on it later – but for now let's just note that it is possible to use the same inverse length units to write the thin lens/mirror equation above. We'll define the inverses of image and object distances to be the two symbols $v = 1/s$, $v' = 1/s'$, so that[152]:

$$v + v' = P \tag{11.21}$$

[152]Note to experts: Obviously, v and v' are intended to sound like "vergences" in the Cartesian description of the lens/mirror equations in the paraxial approximation, but to avoid confusing introductory students I am not using the sign convention wherein $V = -v$ is generally negative. If you are *not* an expert, ignore this footnote!

expresses the a *direct* (instead of reciprocal) algebraic statement of the mirror equation.

However, this is not necessarily easier to use for the purposes of computation, as one still (ultimately) has to do the same algebra to e.g. actually compute s' from a knowledge of s and f. We will postpone further discussion of diopters until we reach multiple-lens systems, as this is where they really shine!

At this point we have derived a simple equation relating s, s' and f. The only rule we have used so far in deriving that equation (which you can easily see holds for plane mirrors as well) is the law of reflection. We have deduced as a *theorem* of this the rule that parallel paraxial rays are diverted by a converging mirror to an image at the focal distance from the mirror. We now need to take these two rules (and a third that is a restatement of the second) and use them to construct *ray diagrams* that permit us to visualize how a converging *or* diverging mirror forms an image out of rays diverging from an object. Constructing such diagrams, and answering a more or less standard set of questions, will constitute most of the *problems* associated with this chapter.

11.3: Ray Diagrams for Ideal Mirrors

To construct our ray diagrams, we need to begin by idealizing spherical mirrors in a way that "hides" things like the fact that many rays we might wish to image with are *not* paraxial. Later in this chapter we'll deal with many of the aberrations that are features of real lenses and mirrors as deviations from ideal behavior in the focussing elements themselves or the light that goes through them, but these will be "corrections" that should not cloud our perception of how things basically work.

First, when drawing rays in a ray diagram, one always assumes that *all deflection by the lens or mirror occurs in a single plane.* This is an idealization, to be sure – the reason mirrors and lenses focus light is because they are *curved*, not planar. But paraxial rays by definition strike close enough to the center that the deviation from planar can be ignored, and we idealize this to the entire plane.

Given this, the following three rays have rules that can be used to locate images and compute magnification for any mirror (and eventually, lens):

a) **The Parallel Ray:** A ray from the object that is parallel to the axis of the mirror is reflected by the mirror **through the focal point**.

b) **The Focal Ray:** A ray from the object that strikes the mirror either *through* the focal point or along a line that *comes from* the focal point is reflected *parallel to the axis of the mirror*.

c) **The Central Ray:** A ray from the object that strikes the mirror in the center is reflected by the mirror **with angle of incidence equal to the angle of reflection** which means that the reflected ray is symmetric across the axis from the incident one.

Now consider the following ray diagrams for various positions of our archetypical arrow object for converging (+) and diverging (-) ideal mirrors.

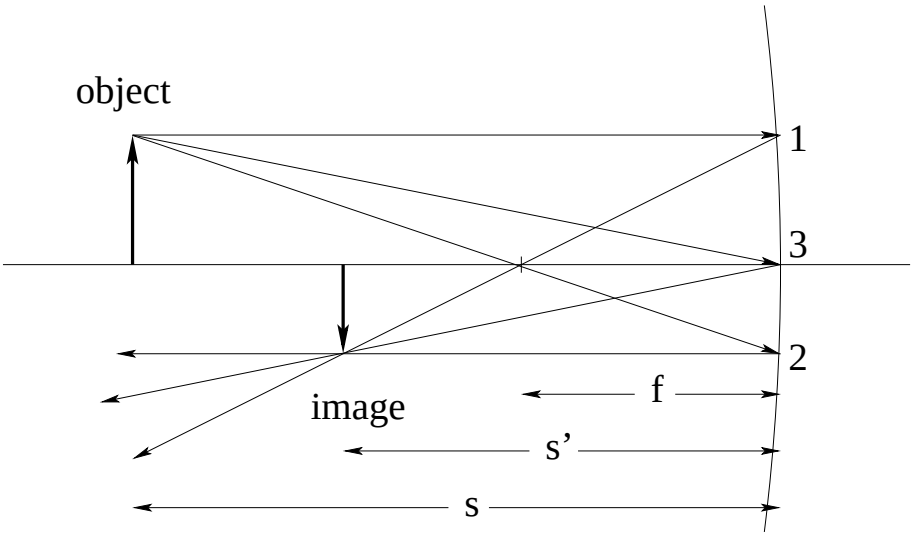


Figure 11.6: Converging mirror with $s = 25 > f = 10$.

In this figure, $f = 10$ cm, $s = 25$ cm. Therefore:

$$\begin{aligned} \frac{1}{25} &+ \frac{1}{s'} = \frac{1}{10} \\ \frac{1}{s'} &= \frac{1}{10} - \frac{1}{25} \\ \frac{1}{s'} &= \frac{1.5}{25} \\ s' &= \frac{25}{1.5} = 16.7 \text{ cm} \end{aligned} \tag{11.22}$$

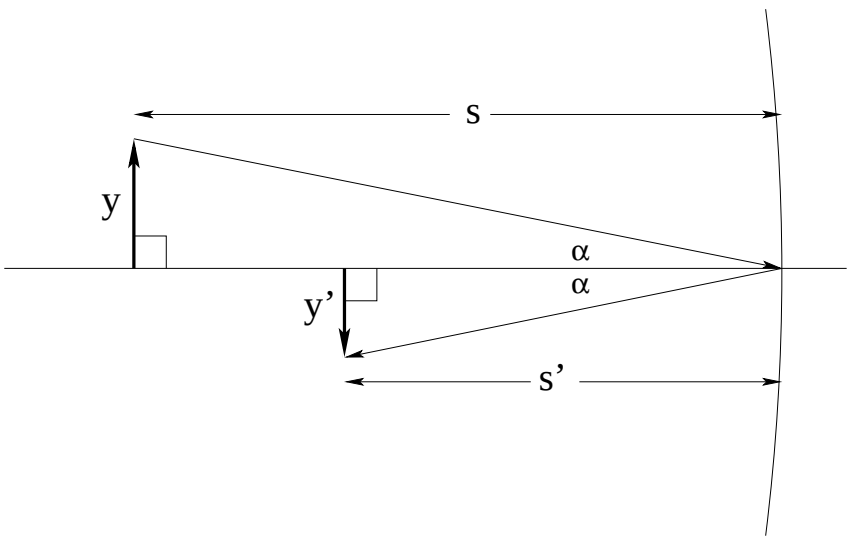


Figure 11.7: Transverse magnification can be determined from the two right triangles formed with the central ray as a hypoteneuse.

To compute the magnification of the image formed above, we note that:

$$\tan(\alpha) = -\frac{y}{s} = \frac{y'}{s} \tag{11.23}$$

(where we rigorously follow the convention that counterclockwise rotation is positive to assign the signs). We define the transverse magnification m of a simple mirror (or lens) is defined by the ratio of the image height y' to the object height y. If we rearrange the terms in this expression, we obtain:

$$m = \frac{y'}{y} = -\frac{s'}{s} \tag{11.24}$$

This expression is valid for all images obtained for any ideal lens or mirror.

Note that in this case, the image formed is real (because the light rays pass through the actual object), inverted, and that the image formed is smaller than the original object.

Let's look at two more possibilities for converging/concave mirrors. In figure (11.8), we see an (upside down) object at a position between f and $2f$. This range is the second possibility for this kind of mirror, one that leads to a *magnified* real image larger than the object.

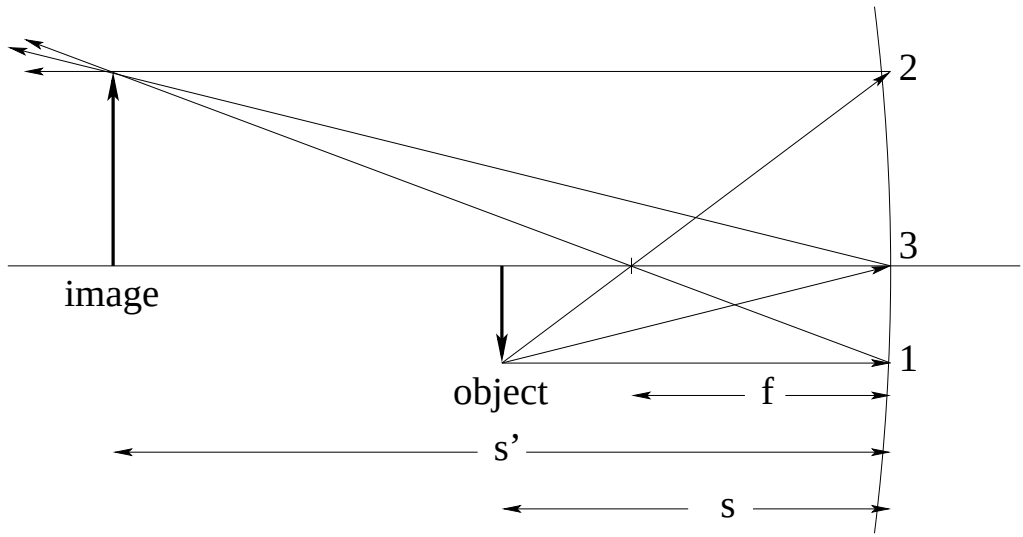


Figure 11.8: Converging mirror with $2f = 20 > s = 15 > f = 10$.

As before, $1/s' = 1/10 - 1/15 = 1/30$ so $s' = 30$ cm. The magnification is $m = -'s/s = -30/10 = -3$. The image is again real and inverted (relative to the object), but in this case the image is larger than the object.

Note that for $s > f$ there is a *symmetry* between solutions with $s > 2f > s'$ and solutions with $s' > 2f > s$, emphasized in the figure above by deliberately drawing the object upside down so that it looks very much like figure (11.8). In fact *any* ray diagram involving real images can work both ways, with s and s' (and the role of the object and image) interchanged because $1/s$ and $1/s'$ appear symmetrically in the mirror/thin lens equation.

In figure (11.9) the third and last distinct possibility for a converging mirror is drawn. In this case, the object is located *inside* the focal length at $s = 5$ cm (for $f = 10$ cm). Thus $1/s' = 1/10 - 1/5 = -1/10$ or $s' = -10$ cm. The magnification is $m = -(-10)/5 = 2$. The final image is *virtual*, *erect*, and *larger* than the object. This is the common way converging mirrors are used as "makeup mirrors" that present a magnified image of the user's face when viewed from inside their focal length.

We only need to present *one* diagram for diverging/convex mirrors, as they all have the same general diagram independent of the relative size of s and f. Note that the first and

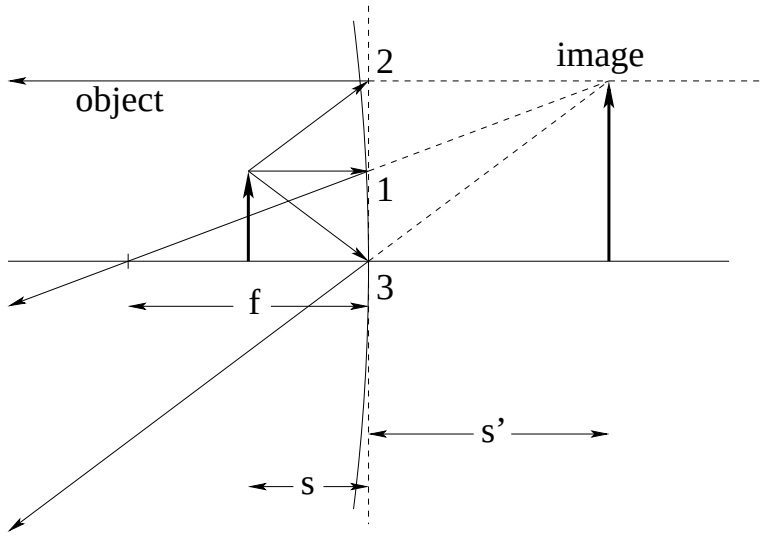


Figure 11.9: Converging mirror with $s = 5 < f = 10$.

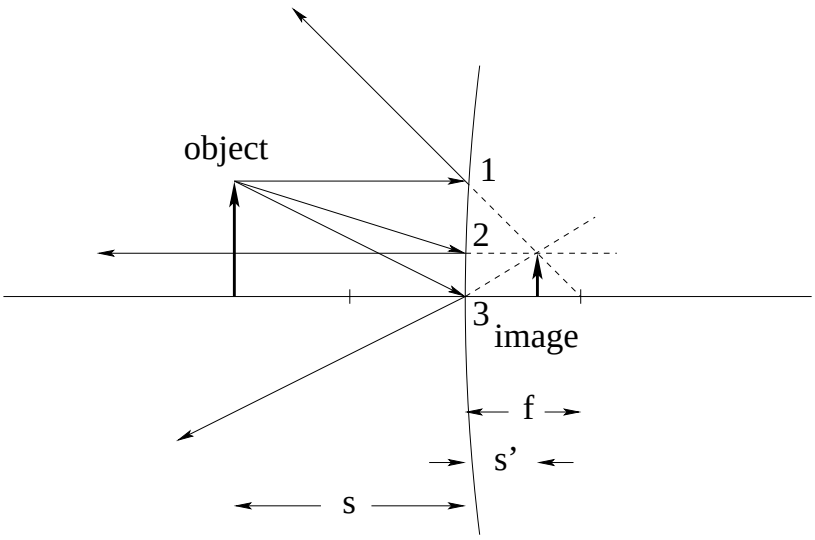


Figure 11.10: Converging mirror with $s = 20 < f = 10$.

second rules are "backwards" compared to converging lenses. A ray parallel to the axis is deflected so it appears to be *coming from* the far side focal length. A ray headed *to* the far side focal length is deflected back parallel to the axis. The central ray is drawn as before.

We apply *as always* the mirror/thin lens formula: $1/s' = -1/10 - 1/20 = -3/20$ so $s' = -6.7$ cm. The magnification is $m = -(-6.67)/20 = 0.33$. The image is erect, virtual, and smaller than the object. All of these general properties will apply (with different numbers) to *any* diverging mirror.

If you master drawing these generic diagrams (and can manage the very simple algebra associated with evaluating e.g. s' and m given s and f, you can with patience analyze any combination of mirrors (and later) lenses) you are presented with.

11.4: Lenses

A spherical lensing surface between two different media with different indices of refraction are drawn in figure (11.11).

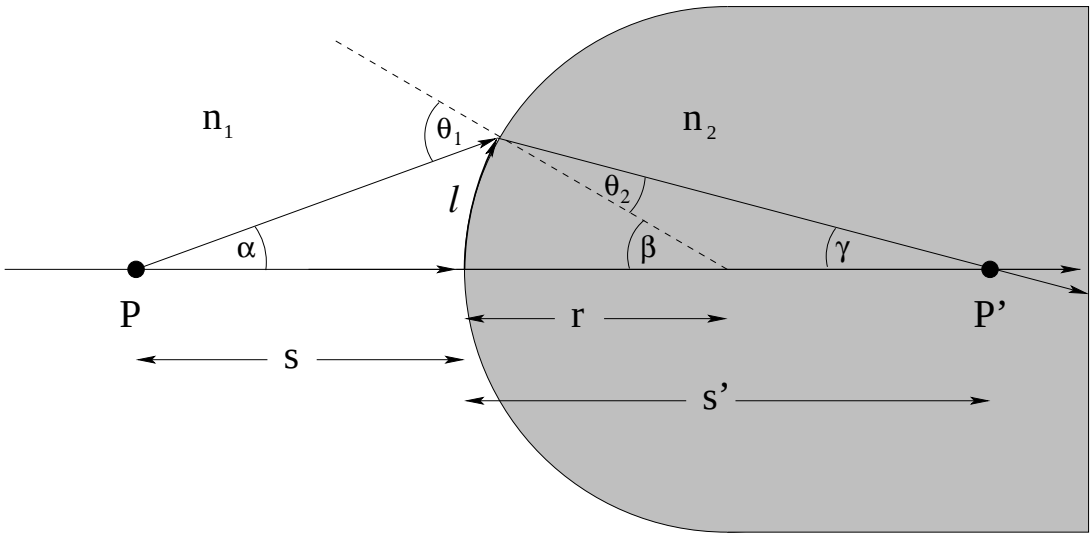


Figure 11.11: Diagram that shows how a spherical lens creates an image via refraction.

As was the case for the mirror, the three angles α, β, and γ in the small angle approximation can be written as:

$$\alpha \approx \frac{\ell}{s} \tag{11.25}$$

$$\beta = \frac{\ell}{r} \tag{11.26}$$

$$\gamma \approx \frac{\ell}{s'} \tag{11.27}$$

We also have *Snell's law* for the (small) angles θ_1 and θ_2:

$$n_1\theta_1 \approx n_1 \sin(\theta_1) = n_2 \sin(\theta_2) \approx n_2\theta_2 \tag{11.28}$$

so

$$\theta_2 = \frac{n_1}{n_2}\theta_1. \tag{11.29}$$

Using triangle rules like the ones above, we also get:

$$\theta_1 = \alpha + \beta \tag{11.30}$$

and

$$\beta = \theta_2 + \gamma \tag{11.31}$$

Eliminating θ_2, this becomes:

$$\beta = \frac{n_1}{n_2}\theta_1 + \gamma \tag{11.32}$$

If we multiply both sides by n_2 and substitute θ_1 from the first equation, this becomes:

$$n_2\beta = n_1\alpha + n_1\beta + n_2\gamma \tag{11.33}$$

or

$$n_1\alpha + n_2\gamma = (n_2 - n_1)\beta \tag{11.34}$$

We substitute in the small angle formulas and cancel l to get:

$$\frac{n_1}{s} + \frac{n_2}{s'} = (n_2 - n_1)\frac{1}{r} \tag{11.35}$$

In most cases of interest to us, the lenses in question will be made out of glass, plastic, or collagen (in the case of the eye) surrounded or faced by air, in which case this will simplify to:

$$\frac{1}{s} + \frac{n}{s'} = (n-1)\frac{1}{r} \tag{11.36}$$

If there are two lensing surfaces separated by a very small distance, we have a so-called *thin lens*. The relevant geometry of a thin lens surrounded by air is shown in (11.12). The first

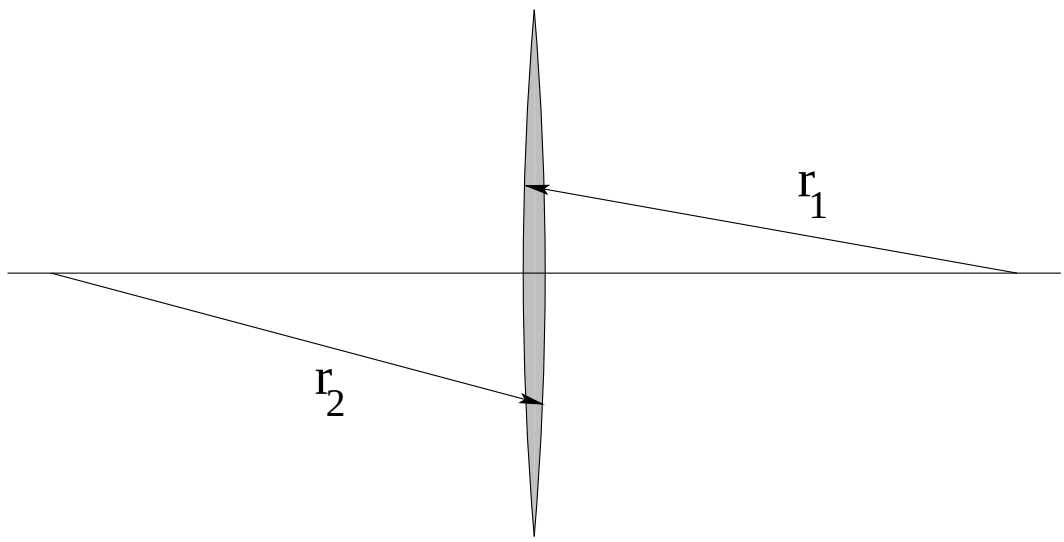


Figure 11.12: Geometry of a thin lens surrounded by air.

surface struck by light from an object (presumed coming in from the left) has positive radius of curvature r_1. The second surface has a negative radius of curvature r_2. The index of refraction of the lens is n.

Suppose we have an object on the left hand side of this lens at distance s. From the formula above, we have:

$$\frac{1}{s} + \frac{n}{s'} = (n-1)\frac{1}{r_1} \tag{11.37}$$

The image of the first lensing surface is a *virtual object* for the second lensing surface. Because it is virtual (located to the *right* of the second surface, on the side light is going *to*) and because we are going from the material with index of refraction n into air, the formula for the second lensing surface is:

$$\frac{-n}{s'} + \frac{1}{s''} = (1-n)\frac{1}{r_2} \tag{11.38}$$

If we add these two formulae, the s' term cancels and, we get:

$$\frac{1}{s} + \frac{1}{s''} = (n-1)\left(\frac{1}{r_1} - \frac{1}{r_2}\right) = \frac{1}{f} \tag{11.39}$$

This is the *thin lens formula* where s'' is the final location of the image of the entire lens. Note that this is *identical* to the formula for the mirror. The focal length is given by the **lensmaker's formula**:

$$\frac{1}{f} = (n-1)\left(\frac{1}{r_1} - \frac{1}{r_2}\right) \tag{11.40}$$

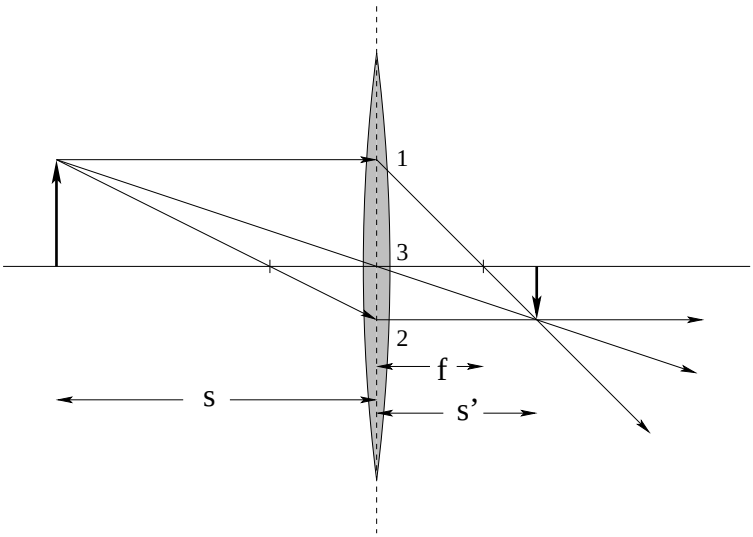


Figure 11.13: A converging lens with focal length of 10 cm and an object at $s = 30$ cm.

With the thin lens formula in hand, we can easily adapt *exactly* the same rules for drawing ray diagrams for locating images. Let's draw a simple ray diagram for a converging and a diverging lens that are similar to the ray diagrams above for mirrors. We do the usual algebra and arithmetic: $\frac{1}{s'} = \frac{1}{10} - \frac{1}{30} = \frac{2}{30}$ so $s' = 15.0$ cm, $m = -\frac{1}{2}$. The final image is inverted, real, and smaller than the object.

As before, if one puts an object inside the focal length it will make a magnified, erect, virtual image, if one exchanges the position of object and image in the example above, one will obtain an inverted, real image that is larger than the object.

A diverging lens, on the other hand, has only one generic diagram to be learned. It is basically the same as for the mirror, except that rays are transmitted through the thin lens (with all bending occurring at the thin plane representing the center plane of the lens) instead of reflected from it. In the situation represented in figure (11.14), the image is virtual, erect, and smaller than the original object. Show (from the numbers and thin lens formula) that $s' = -6.67$ cm and that $m = 1/3$.

11.5: Multiple Lenses and Diopters

We already encountered our first *compound lens system* made up of *two* lens surfaces in our derivation of the thin lens equation above. A similar idea can be used to analyze systems made up of two (or more) lenses, using the image of the first lens encountered by light as it passes through the system as the "object" of the second lens, and the image of the second as object of the third, etc. In a moment, we'll analyze a few such systems (and get some practice at using the thin lens and/or mirror equations while we are at it) but first, let's learn

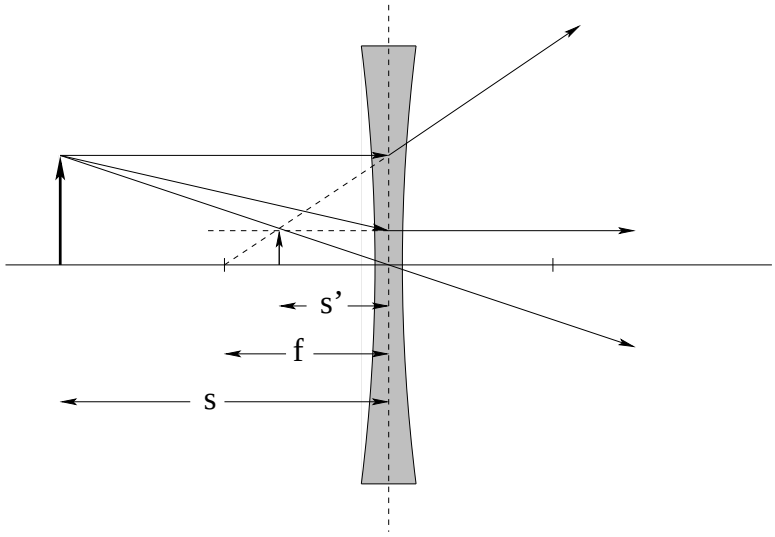


Figure 11.14: A diverging lens with focal length of -10 cm and an object at $s = 20$ cm.

more about an important concept in optics – that we mentioned briefly in our discussion of mirrors – that ***massively simplifies*** the algebra (and arithmetic!) of multiple lenses and is the one commonly used in the everyday optics of the glasses used to correct defects in vision: the ***diopter***.

11.5.1: Diopters

You will have noticed, I'm sure, that the thin lens equation and the mirror equation are both reciprocal sum equations. As a consequence, if you try to solve them algebraically, you will always find yourself doing things like (to find s' given s and f):

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f} \quad \Longleftarrow \quad \frac{1}{s'} = \frac{1}{f} - \frac{1}{s} = \frac{s-f}{fs}$$

If you are given numbers, say $s = 30$ cm, $f = 10$ cm, then:

$$s' = \frac{fs}{s-f} = \frac{300}{20} = 15 \text{ cm} \tag{11.41}$$

That's not actually *terribly* difficult, but consider the following. Recall the short discussion in the section on mirrors where we introduce the notion of the *power* of a mirror (or lens) in units of inverse length, so that a flat mirror or lens has power $P = 0$ instead of $f = \pm\infty$. In the SI system, diopters are equivalent to:

$$1 \text{ diopter (D)} = \frac{1}{1 \text{ m}} = 1 \text{ m}^{-1}$$

although one usually doesn't express other inverse length quantities (such as wave number) in diopters – the use of the unit is customary only in geometric optics applications.

Let's recall the expression the thin lens/mirror equation(s) in terms of the power of the lens or mirror (the inverse of its focal length) given in diopters. We previously defined $v = 1/s$ and $v' = 1/s'$ *with our usual sign convention* so v is positive when the object is on the side light is

coming *from* and v' or P are positive when the image distance or focal length is on the side light is going *to*. Then the thin lens equation takes the following simple form:

$$v + v' = P \tag{11.42}$$

with the units of all three in diopters[153]

This certainly seems *algebraically* simpler. Let's rework our example with the same numbers. $v \approx 3.33$ D, $P \approx 10$ D, so:

$$v' = 10 - 3.33 = 6.67 \text{ D}$$

and $s' = 1/v' = 1/6.67 = 0.15 = 15$ cm. It looks like the algebra is simpler, but the actual arithmetic itself is pretty much a tossup – we can avoid putting things over common denominators and taking an inverse, but we have to take inverse of all of the given quantities instead. So why do we bother with diopters? Why are *they* the way optometrists and ophthalmologists prescribe and describe lenses?

It is because of the way we can ***combine two lenses with different focal lengths to create the equivalent of a single lens***, at least if the two lenses are physically very close together relative to their focal lengths. Suppose we have two lenses that are physically very close together as shown in figure 11.15:

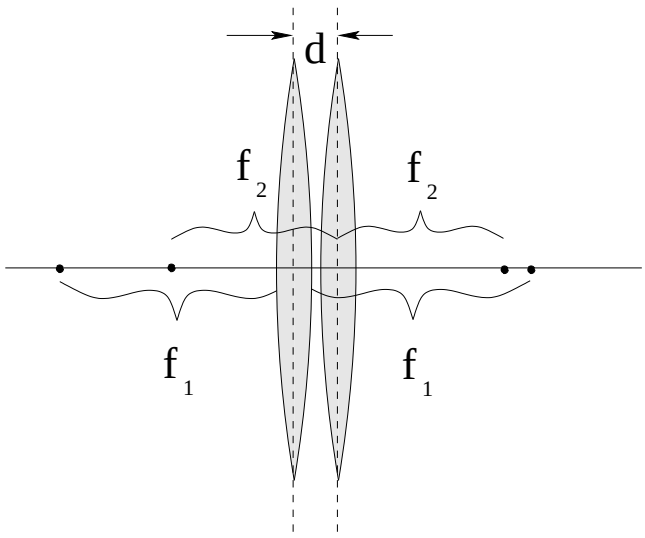


Figure 11.15: Two thin lenses almost touching (with $d \ll f_1, f_2$).

Now suppose an object is placed a distance $s = s_1$ to the left of the first lens with focal length f_1. It makes an image at position s_1' according to:

$$s_1' = \frac{f_1 s_1}{s_1 - f_1} \tag{11.43}$$

[153]For the record, the thin lens equation here should really be written as $V + P = V'$, using the object ***vergence*** $V = -nv$ with n the index of refraction of the medium between the object and the lens (in this case $n = 1$ for air or vacuum). This has the opposite sign, note well, of the one I am using for v in this textbook, while the signs of image vergence $V' = nv'$ and power P are unchanged. However, this ***Cartesian*** description of optics is usually taught in more advanced courses in optics that treat e.g. thick lenses and more general optical elements in arbitrary media in a *matrix* representation, and swapping a single sign and adding in a discussion of why it is necessary to multiply by the index of refraction is going to cause nothing but confusion in an introductory course.

as indicated above in equation 11.41. The image from the first lens becomes a *virtual object* for the second lens, with object distance:

$$s_2 = -(s_1' - d) \approx -s_1' = \frac{f_1 s_1}{f_1 - s_1} \tag{11.44}$$

Here we made a key approximation – d is "small enough", relative to s_1', that we can neglect it. We certainly *could* carry it through the next two steps, but I think you'll agree that the algebra is difficult enough *without* our doing so!

Then the final image is formed at $s' = s_2'$ according to:

$$s_2' = \frac{f_2}{\frac{f_1 s_1}{f_1 - s_1} - f_2} \frac{f_1 s_1}{f_1 - s_1} = \frac{f_1 f_2 s_1}{(f_1 + f_2)s_1 - f_1 f_2} = \frac{\frac{f_1 f_2}{f_1 + f_2} s_1}{s_1 - \frac{f_1 f_2}{f_1 + f_2}} \tag{11.45}$$

Or (letting $s = s_1$, $s' = s_2'$ for the compound lens system):

$$s' = \frac{sf}{s - f} \tag{11.46}$$

with:

$$f = \frac{f_1 f_2}{f_1 + f_2} \tag{11.47}$$

This is *more than a bit* algebraically daunting. The final expression *does* tell us how to make a single lens that is "equivalent" to the two lens system, but there were a lot of steps in between, and we actually *started* the algebra with 11.41 and skipped the steps we already put in to get it! If you look carefully, you'll see that this is pretty much *exactly* the same thing we did obtaining the thin lens equation in the first place, except that we didn't derive it in terms of f's per se, and the "thin" bit in that derivation is the same thing as "neglect d" step above.

Now let's do it with our diopter "vee" description and *power*. The first lens is:

$$v_1 + v_1' = P_1 \Longrightarrow v_1' = P_1 - v_1 \tag{11.48}$$

The second lens again uses a virtual object made of the image from the first lens (with d ignored), $v_2 = -v_1'$. Then:

$$v_2 + v_2' = -v_1' + v_2' = v_1 + v_2' - P_1 = P_2 \tag{11.49}$$

or

$$v_1 + v_2' = v + v' = \boxed{P_1 + P_2 = P} \tag{11.50}$$

The algebra (and result) are now *easy*. ***The two "touching" lenses together can be replaced by a single lens with the same total power!*** This makes it *very easy* to assemble a set of lenses that have any desired target power/collective focal length!

Note that we ***get the same answer either way we derive it!*** Either way leads to:

$$P = \frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} = P_1 + P_2 \Longrightarrow f = \frac{f_1 f_2}{f_1 + f_2}$$

if there is ever any real *need* to find the focal length of the composite lens system. In most cases there is not! Overall, it is simply a lot more convenient to work with a slightly more

general version of v, v' and power when working with complex systems with multiple, possibly thick, lenses separated by distances too large to neglect.

This is – within an overall sign that isn't important in this introductory discussion – the way to "add lenses" using the *Cartesian* representation of optical elements, a topic usually treated in higher-level courses in optics. It also neglects what happens when the two lenses (or lensing surfaces) aren't very close to each other (so d in the figure above isn't "negligibly small") – so-called *thick* lenses or lenses separated by *large* distances d require (a lot!) more work to handle, and are the motivation for using the Cartesian matrix representation if you are actually working for an optics company designing multilens systems. But it is plenty for our purposes in this introductory treatment.

11.6: The Eye

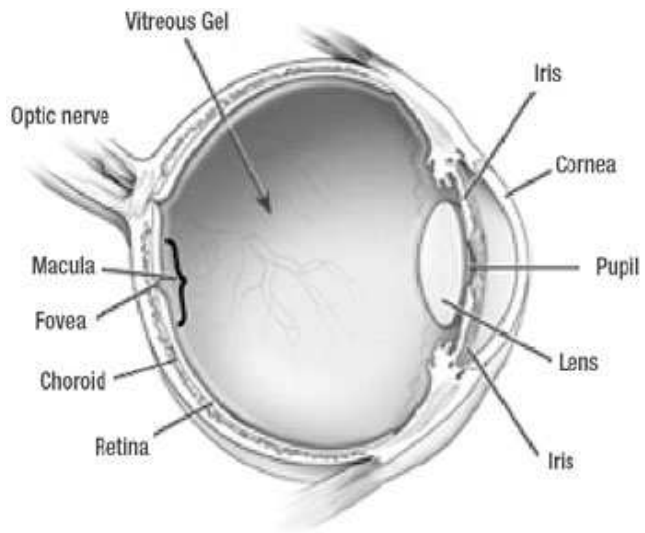


Figure 11.16: A simplified anatomical diagram of the human eye.

The eye is roughly spherical and approximately one inch in diameter. Figure (11.16) shows its essential anatomy. Here is a brief review of the components of the eye.

- **Cornea:** The cornea of the eye is the rounded, transparent structure at the front of the eye. It is strongly curved, and is responsible for *most* of the bending of light required to focus images onto the...
- **Retina:** The retina is the "film" of the eye. It consists of tight bundles of photosensitive nerves called *rods* (sensitive to light intensity) and *cones* (sensitive to intensity in specific colors. In the center of the retina is the...
- **Macula:** The macula is the most sensitive part of the retina and is where one "sees" the object of one's attention. It is more or less in front of the...
- **Optic Nerve:** which pipes all of the information transduced from the light image cast on the retina to the brain. The retina (especially the macula) is very sensitive to light and easily damaged. To control the amount of light entering the eye, the...

- **Iris:** The iris is a ring of pigmented tissue that can open or contract to let more or less light into the...
- **Pupil:** The pupil is the aperture for light into the eye. When it is dark, the iris opens and lets all the light possible into the retina (which is very sensitive and capable of seeing with remarkably little light). When it is very bright, the iris closes down to a pinpoint. This actually increases visual acuity – see the *pinhole camera* – independent of the action of the...
- **Lens:** The lens of the eye is normally in a state of tension maintained by suspensory ligaments called **zonules** that keep it flattened out, with a maximally long focal length. A ring of **ciliary muscles** surrounding the lens can be contracted, which removes a part of this tension, predictably bulging the lens and thereby reducing its focal length. This process is called **accommodation**.

It is important to understand that accommodation can only *reduce* the focal length of the lens, not increase it, as well as the fact that the cornea is responsible for most of the focal length of the combined system – the actual lens is more of a "correction" to the overall focal length already achieved by the cornea alone. We now need to understand the three common conditions that describe the eye.

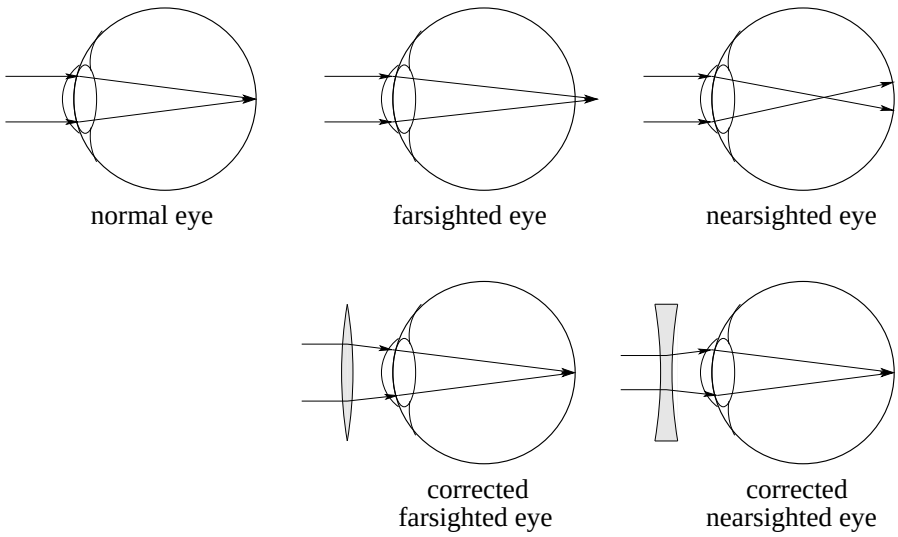


Figure 11.17: The focal length of the relaxed (combined) lensing acting of the eye for a normal eye, a farsighted eye (hyperopia), and a nearsighted eye (myopia).

The focal length of a *relaxed* lens of an eye with *normal* vision is on the retina, so distant objects (at "infinity" compared to the size of the eye) are automatically in focus (as a real image cast upon) on the retina. Given a distance from the cornea to the retina of roughly 2.5 cm, this means that the strength of the lens of a normal eye is approximately $\frac{1}{0.025} = 40.00$d. When viewing less distant objects, accomodation *shortens* the focal length to bring them into focus on the retina.

The focal length of a relaxed *farsighted* eye is *behind* the retina (too long, strength less than 40.00d) and is corrected with a *converging* lens to make up the difference. If one expresses strength in diopters, one can simply add a converging lens with a strength in diopters to the

strength of the the eye to get the "right strength" to make the combination focus distant objects on the retina with the eye's lens relaxed. Note that a hyperopic person *can* see in focus all the way out to infinity, but they have to use accommodation to shorten their lens's "too long" relaxed focal length see even distant objects, which can lead to eye fatigue and headaches.

The focal length of a relaxed *nearsighted* eye is in *front* of the retina (too short, strength greater than 40.00d) and is corrected with a *diverging* lens to take *away* some of its strength. A myopic individual simply cannot see distant objects in focus without a corrective lens because accommodation cannot *increase* the focal length of the eye's lens, it can only further decrease it.

Accommodation can shorten the focal length only so far, which limits how close an object can be and still be focused on the retina. The nearest point one can bring an object to the eye and still bring it into focus on the retina is called the *near point* of the eye and is also the *distance of most distinct vision*, represented x_{np}. In most adults, this distance is around 25 cm (less for small children, longer for the elderly).

A nearsighted person's lens *already* has too short a focal length to be able to focus distant objects on the retina, and accommodation only shortens the focal length still farther. A nearsighted person cannot see anything clearly at distances *greater* than some point, called the *far point* for that person's eyes. A nearsighted person is one for whom the far point x_{fp} is less than infinity.

A common aberration of human eyes is a condition called **astigmatism**. Astigmatism is what happens when the eye's lens is no cylindrically symmetric. That is, the focal length of the lens in the horizontal plane is not the same as the focal length in the vertical plane. One can then bring things into focus in one dimension with accommodation, but only at the expense of blurring them in the other. The solution is to wear lenses that are astigmatic in the opposite direction to add up to neutral (or to person's otherwise necessary correction).

As a person's eyes age, their ability to focus changes. People with once normal vision can become nearsighted or farsighted. After the age of roughly 50 a new condition often emerges – that of **presbyopism**. The collagen of the lens hardens over time. Its flexibility decreases, making it more difficult for the eye to accommodate and *increasing the near point*. This kind of "farsightedness" can occur even for nearsighted individuals. The solution is to correct with "reading glasses" – positive lenses that permit a presbyopic individual to read at normal distances. They can be combined into "bifocals" – reading glasses for short distances plus diverging lenses to correct myopia at long distances – for people with the latter condition.

11.7: Optical Instruments

11.7.1: The Simple Magnifier

The "size" of an object to the human eye is determined by three distinct things. Humans have binocular vision, and use *parallax* – the apparent displacement of an object seen from two slightly different positions – to get a sense of an object's distance. This is reinforced by the physiological sense of *accommodation*, which gives one a sense of relative nearness. Finally, given the distance, it is determined by the *angle* the image subtends on the retina.

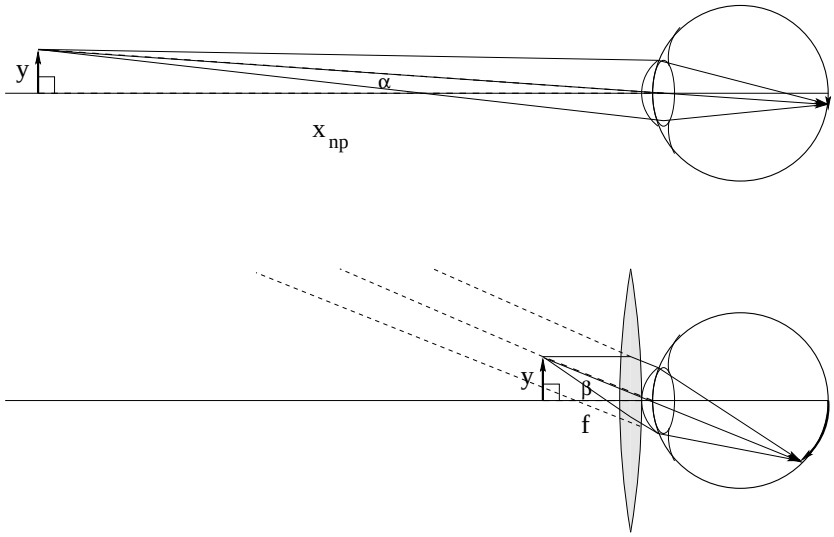


Figure 11.18: A converging lens used as a simple magnifier.

To see a small thing as clearly as possible, we naturally bring it to the closest point we can, so its details subtend the largest possible angle when our eyes are maximally accommodating. In figure (11.18) the top picture shows an object of height y viewed at the near point. When the image is focused on the retina by the maximally accomodating eye, it subtends an angle of α, where:

$$\alpha \approx \tan(\alpha) = \frac{y}{x_n p} \tag{11.51}$$

in the small angle approximation (which is entirely justified because we only "see" detail with the macula, which in turn only occupies around 0.2 radians in the center of the visual field. Even if we are examining a larger object, we do so by redirecting the eye to look at it in patches that cover it in small angle chunks.

To use a simple magnifier we place a converging ($f > 0$) lens immediately in front of the eye. The object is placed at its focal point. It therefore forms a *virtual image* at $-\infty$ that is automatically brought into focus by the relaxed normal (or vision corrected) eye. It now subtends an angle β on the retina given by:

$$\beta \approx \tan(\beta) = \frac{y}{f} \tag{11.52}$$

The magnification is therefore the ratio of the new angle (with the magnifier) to the angle without it, when the object is seen at the near point. The magnification of the object occurs because one can bring the object *closer* to the eye than x_{np} and still see it clearly (more clearly, even, than before given that one does not have to accommodate). Its magnification is given by:

$$M = \frac{\beta}{\alpha} = \frac{x_{np}}{f} \tag{11.53}$$

It is very important to understand the simple magnifier, as it forms the eyepiece of *both* the microscope *and* the telescope, our next two optical instruments.

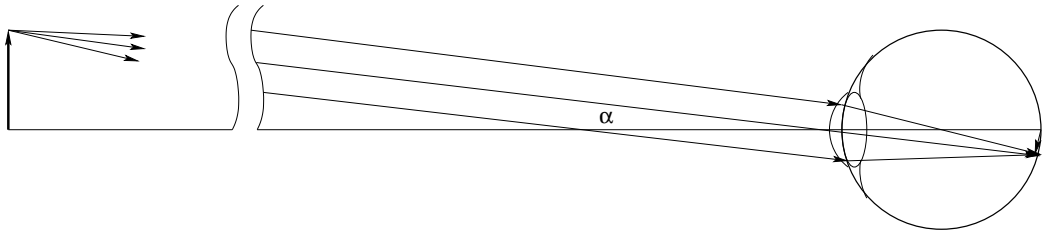


Figure 11.19: An regular (inverting) telescope.

11.7.2: Telescope

A telescope is an optical instrument used to bring *distant* objects *closer* so that you can see them magnified and much more clearly. In figure (11.19) you can see what a ray diagram looks like for light from a very distant object entering the naked human eye. The rays from the originating point, after travelling a long distance, necessarily enter the eye more or less parallel and are focused by the relaxed normal lens onto the single point on the retina determined by the central ray entering at angle α.

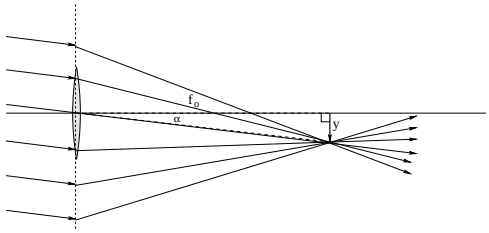


Figure 11.20: The first lens creates a real image of the distant opject.

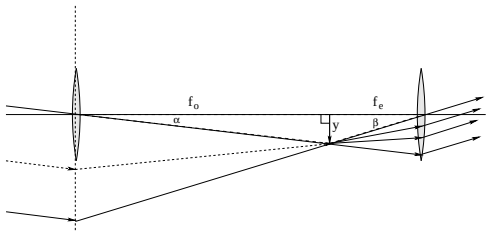


Figure 11.21: The second lens acts as a simple magnifier to allow this (tiny, inverted) real image to be viewed at infinity from a point of view much closer than the near point of the eye.

To magnify our view of this object, we begin by inserting a lens with a *long* focal length f_o into the optical path. This takes light from the (infinitely) distant object and creates an *inverted real image of it* at the focal point as shown in the first panel in figure (11.21) above. We draw many parallel rays and show them *as if* they were deflected by the *ideal* lens at its plane of refraction. This shows how we can use rays from the image the same way we would use rays from the original object when this image becomes a virtual object for the second lens, and pick any ray that is convenient for our purposes of analyzing the magnification.

This image (virtual object) is "infinitely" smaller than the original object but it has the advantage of being *right there in space* in front of the eye, not infinitely distant. We can therefore examine it quite closely. To do so, we use a second lens as a *simple magnifier*, placing it so that the virtual object is at *its* focal point. This is shown in the second panel, figure **??**.

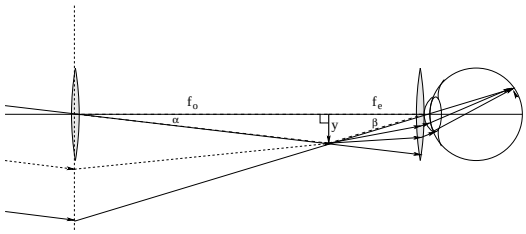


Figure 11.22: This ultimately creates an angular magnification $M = -\beta/\alpha = -f_o/f_e$.

Since the virtual object is at the focal point f_e, rays diverging from the virtual object exit the second lens parallel to the central ray, shown entering at angle β. This bundle of parallel rays corresponds to a virtual image at (negative) infinity but deflected so that their angle relative to the central axis if much steeper. We can easily compute the angular magnification of this telescope by noting that:

$$\alpha \approx \tan(\alpha) = -\frac{y}{f_o} \tag{11.54}$$

and

$$\beta \approx \tan(\beta) = \frac{y}{f_e} \tag{11.55}$$

so that

$$M = \frac{\beta}{\alpha} = -\frac{f_o}{f_e} \tag{11.56}$$

In the final panel figure 11.22, we show what happens when this final image at infinity coming in at angle β looks like when closely viewed by a human eye. Since the image is infinitely distant (the rays enter the eye parallel) it can be comfortably viewed with the relaxed normal lens, which will focus the bundle down to a single point on the retina determined by the central ray at angle β. Obviously the total angle subtended on the retina is much larger – the object being viewed appears much larger to the eye and senses. The major disadvantage of this telescope is that it *inverts* the image – everything viewed is upside down and backwards. This makes it a bit tricky to find objects as they move the *opposite* way one thinks that they should when viewing them through the telescope.

Interestingly, this final disadvantage can easily be eliminated by using a ***diverging*** lense for the eyepiece. Ordinarily one thinks of a diverging lens as making something smaller, but because we can place the image from the first lens anywhere we wish, we can turn it into a virtual object at the *far* focal point of a diverging lens. One obtains the same formula for the magnification, but now $f_e < 0$ and the overall angular magnification is *positive*.

This kind of telescope is called a **Galilean telescope** and is much more convenient to look through than a regular telescope. As you can see from figure (11.23), the angular magnification of a Galilean telescope is still:

$$M = \frac{\beta}{\alpha} = -\frac{f_o}{f_e} \tag{11.57}$$

(where now $f_e < 0$ is *negative*) but parallel rays from the distant object enter the eye after passing through the telescope in the *same* angular sense that they enter it when viewed without the telescope. As before, note that we used a ray that *would* have passed through the center of the second lens (and the eye, if the eye were drawn into the figure) in order to determine

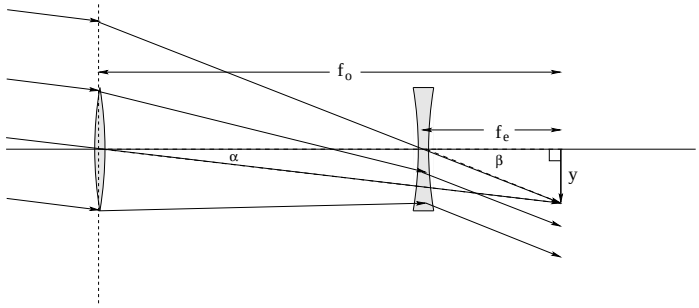


Figure 11.23: A "Galilean" telescope uses a *diverging* lens for the eyepiece. This does not affect the formula for the magnification, but it ensures that the eye sees the distant objects *erect* instead of inverted.

the angle all of the parallel rays leave the eyepiece lens before entering the (normal) eye and being focused on the retina.

Telescopes (in the hands of Galileo and others) were an instrument that ushered in the Enlightenment in the seventeenth century, putting an end to several thousand years of human history where mythology and inexact observations prevented the systematic development of a consistent theory of physics. Let's look at another instrument that had a revolutionary impact on human society, the microscope.

11.7.3: Microscope

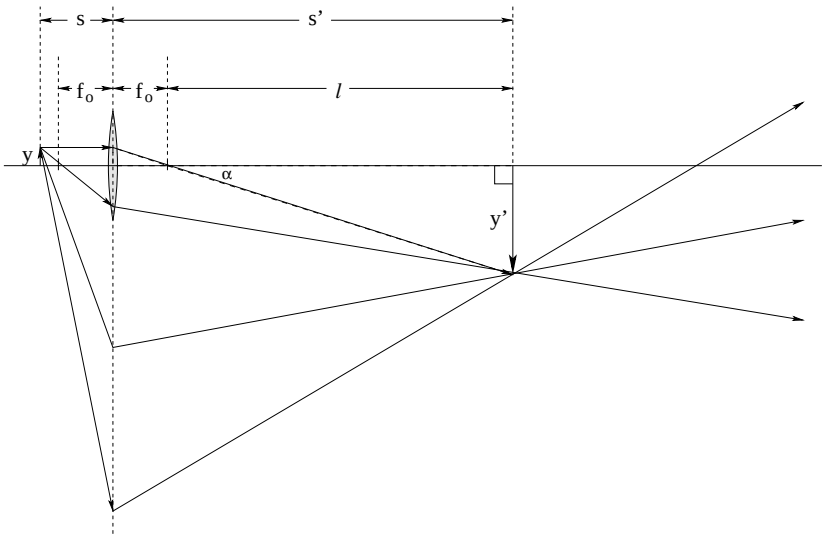


Figure 11.24: The first magnification stage of a compound microscope brings a *small* object just outside of the focal point of the objective lens into focus as a *real, magnified image* at the end of the **tube length** *l*. By comparing the two dashed similar triangles, one can see that the first stage magnification is $-\frac{l}{f_o}$.

A compound microscope is used to view a very small, but nearby object. It accomplishes this in two stages. Two short focal length lenses are situated at ends of a tube much longer tube. The **tube length** l of the microscope is by definition the distance between the focal point of the first, or *objective* lens (which must be converging) and the second, or *eyepiece* lens.

The objective stage of the magnification occurs as the the object is placed on a movable platform just outside of the focal length of the objective lens of the microscope. The platform is raised or lowered (altering s, the object distance) until the objective lens forms a *magnified, real image* of the object at the end of the tube length as shown in figure (11.24).

The magnification of the objective stage is:

$$M_o = -\frac{\ell}{f_o} = -\frac{f_o + l}{s} \tag{11.58}$$

where the first relation is the one actually used, but the second one (based on the observation that $s' = f_o + l$) can be used to find the correct object distance s that will accomplish this.

This real, magnified image can be viewed with the naked eye, but of course the naked eye can view it no *closer* than x_{np}. The second stage of a compound microscope consists of an eyepiece lens is used as a *simple magnifier* to view this real image in precisely the same way we used it for the telescope, and can be converging or diverging as was the case for the telescope. It produces a virtual image at infinity that subtends a greater angle than the real image formed by the objective lens alone would if viewed at the near point of the relaxed normal eye.

The magnification of the eyepiece used as a simple magnifier is therefore:

$$M_e = \frac{x_{np}}{f_e} \tag{11.59}$$

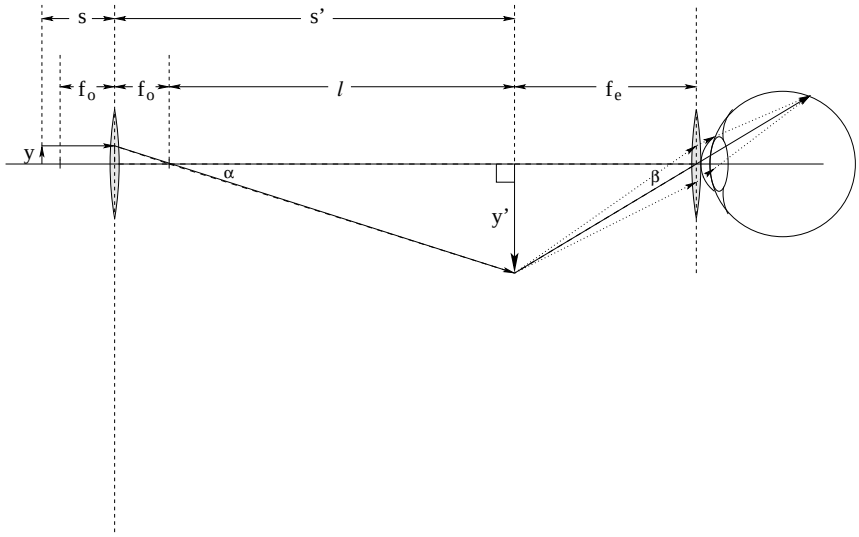


Figure 11.25: The second magnification stage of a compound microscope brings the *highly magnified* image from the objective stage close to the eye by functioning as a simple magnifier. By bringing the virtual image in from x_{np} to f_e it magnifies it by an addtional factor of $\frac{x_{np}}{f_e}$.

which yields an overall magnification for the two stages working together of:

$$M_{tot} = -\frac{\ell\, x_{np}}{f_o f_e} \tag{11.60}$$

As we noted and can see in figure (11.26) above, one can use a diverging lens for the eyepiece by placing the real image formed by the objective on the *far* side of the diverging lens to form a **"Galilean" microscope**. As before (for the telescope) this microscope does not invert the image (inversion is inconvenient and undesireable) but otherwise the same formula works for the magnification provided that one uses a negative f_e for the diverging lens. It has the further advantage of having a slightly shorter overall length.

Typical numbers for a compound microscope this might be $f_o = f_e = 1$ cm, $l = 10$ cm, for a total magnification of 250 (inverting or non-inverting). 250x microscopes are *more than adequate* to observe e.g. blood cells, bacteria, the cellular structure of plant an animal tissue, amoeba, paramecium, and a host of microorganisms and cellular structures. For example, amoeba can range in size from 10-1000 μm (where the latter, note well, is roughly a millimeter and barely visible to the naked eye). A 250 power microscope can make an amoeba appear to the eye as large as a 25 cm object, clearly revealing its nucleus and vacuoles. Even small amoeba or bacteria will appear several millimeters in size at this magnification.

Just as the telescope caused a revolution in our vision of cosmology and the structure of the Universe at large distances and over long times, the microscope caused a revolution in our vision of the world of biology. Disease, which had long been thought of as being caused by demons or by a curse afflicted on sinners by God, was seen to be caused by living organisms too small to be seen by the naked eye. Where before the only possible cure for most diseases was believed to be divine intervention, miracles brought about by repentance and prayer, the microscope enabled the discovery of antiseptic medicine – that heat, soap and water, alcohol, and eventually antibiotics kill off disease-causing microorganisms to prevent or cure disease quite independent of "magic" such as miracles or prayer. The two together brought about the

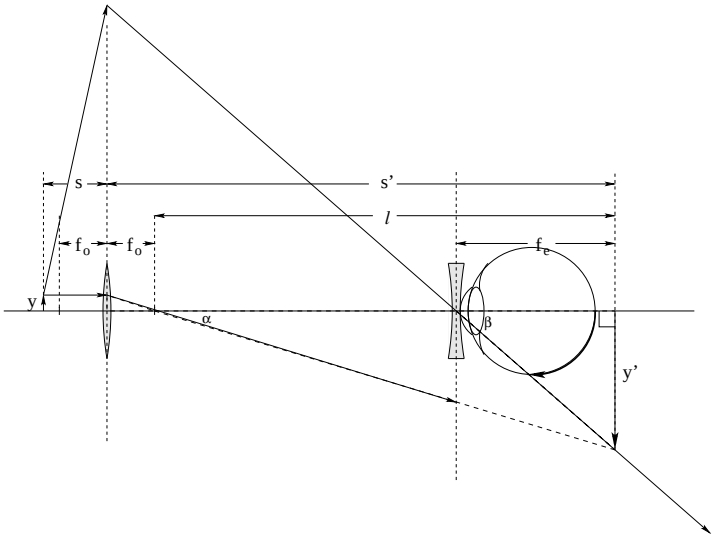


Figure 11.26: A “Galilean” microscope uses a *diverging* lens for the eyepiece. This does not affect the formula for the magnification, but it ensures that the eye sees the tiny objects *erect* instead of inverted. As always, we use a “central” ray for the second lens that is deflected at the plane of the first lens *as if* it passes through both lenses to find the location and size of the final image.

Enlightenment, a time of intense discovery and invention that ultimately ushered in the rational modern world of today.

Homework for Week 11

Problem 1.

Physics Concepts

Make this week's physics concepts summary as you work all of the problems in this week's assignment. Be sure to cross-reference each concept in the summary to the problem(s) they were key to. Do the work carefully enough that you can (after it has been handed in and graded) punch it and add it to a three ring binder for review and study come finals!

Problem 2.

Derive the "mirror" (and thin lens) equation:

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f} = \frac{2}{r}$$

for a ***spherical concave mirror*** as seen in class. Remember, this involves drawing a picture of an object that is a point on the axis of the mirror and the rays that local its point-image, then doing some work with triangles and the small angle approximation.

Problem 3.

"Solve" the following four cases involving mirrors. In all cases ***draw a standard 3-ray diagram*** using an erect arrow as an object, ***solve for*** s', ***find the magnification*** m, and ***indicate whether the image is erect or virtual!***

a) $f = 10$ cm, $s = 5$ cm.

b) $f = 10$ cm, $s = 15$ cm.

c) $f = 10$ cm, $s = 25$ cm.

d) $f = -10$ cm, $s = 10$ cm.

Problem 4.

"Solve" the following four cases involving thin lenses. In all cases ***draw a standard 3-ray diagram*** using an erect arrow as an object, ***solve for*** s', ***find the magnification*** m, and ***indicate whether the image is erect or virtual!***

a) $f = 10$ cm, $s = 3$ cm.

b) $f = 10$ cm, $s = 15$ cm.

c) $f = 10$ cm, $s = 30$ cm.

d) $f = -10$ cm, $s = 20$ cm.

Problem 5.

The human eye is the primary optical instrument. Draw a normal eye, a nearsighted eye, and a farsighted eye, showing the location of the relaxed-eye focal length in all three cases. Draw them a second time with the appropriate corrective lenses, showing with simple rays how they work to fix the problem(s).

Problem 6.

a) Suppose that the relaxed eye of a great white shark has a focal length of 5 cm when the shark is underwater (which is just the distance from the lens to the shark's retina, of course). Several inane movies portray great white sharks hunting prey on land, with their eyes entirely in air.

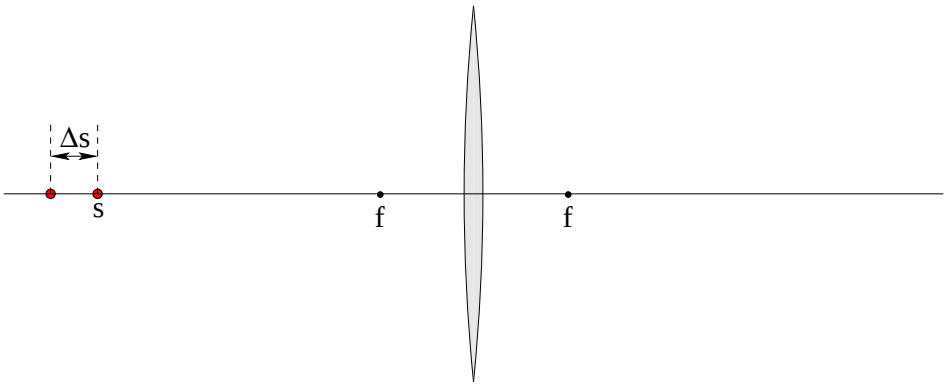
   Is the relaxed focal length of the shark's eye *in air* longer or shorter? In other words, is the fish nearsighted or farsighted in air? Would we need to equip the sharks with special glasses if we *really* expected them to be dangerous if they fell to earth in sharknadoes or came along swimming through the corn?

b) Suppose that the relaxed eye of a particular human has a focal length of 2.5 cm in air. The human falls into the water and opens their eyes (to check for nearby great white sharks)! Is the focal length of the relaxed human eye *in water* longer or shorter? Can a person with myopia (nearsightedness) or presbyopia (farsightedness) see better than a normal human underwater?

Don't just answer with guesses for either question – you need to make a complete argument based on the lens-maker's formula or Snell's law directly, possibly supported by pictures and a bit of discussion.

Problem 7.

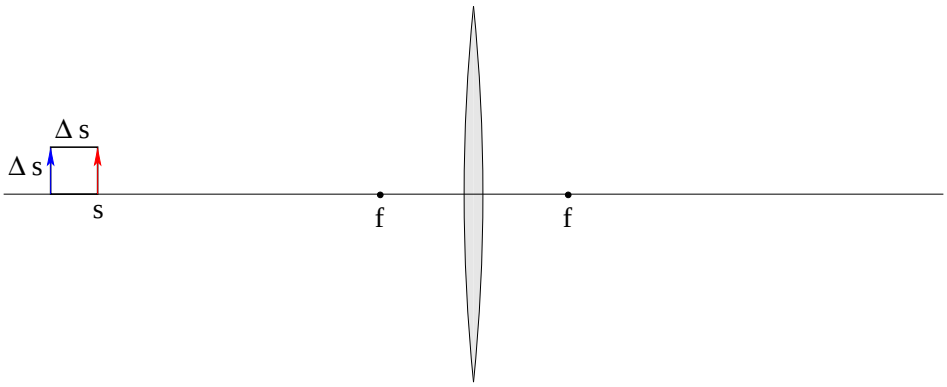
So far we have only looked at one aspect of magnification: the transverse magnification. However, thin lenses and mirrors *also* alter the apparent *depth* of an extended object – the distance between details on its visible face closest to the lens or mirror and detail on its visible face farthest from the lens or mirror. The change in apparent depth is called the *lateral* magnification, and it is *nonlinear* in s and s', unlike the transverse magnification. Let's explore this idea.



a) Consider the two "light sources" separated by a short bar located on the central axis of the thin lens drawn above. Note that the one closest to the lens has an object distance of s, the farther one is at a distance of $s + \Delta s$ where $\Delta s \ll s$. Prove that the *lateral magnification* of this object is:

$$m_l = \frac{\Delta s'}{\Delta s} = \frac{s'^2}{s^2} \tag{11.61}$$

where $\Delta s' \ll s'$ is the length of its image along the central line.



b) Consider the square with height Δs in the figure drawn above. Imagine that the front and rear sides of the square are independent objects, and draw a careful ray diagram for the front vertical arrow and a second one for the rear vertical arrow onto your figure to locate and draw the resulting image of the square. You may wish to use different colors for the two diagrams. Is it still square?

You will likely need to use your friend, the *binomial expansion*, when solving this problem.

Problem 8.

a) ***Draw the standard ray diagram(s) for the simple magnifier*** and use them to derive its (angular) magnification when it is used to view a (virtual, erect) image of the object located at $-\infty$ (in focus with the relaxed lens of a normal eye).

b) One *can* adjust the object position and view a virtual image that is *closer* than $-\infty$ and a bit larger at the cost of using accommodation to view it. Solve for object distance s such that the (virtual, erect) image is ***at the near point of the eye*** x_{np} when viewed through the magnifier. Draw a ray diagram for this case as well.

c) What is the overall (angular) magnification of the image now (with the image located at x_{np})? Is the improvement worth the eyestrain?

Problem 9.

For a physics lab, you are given the following lenses:

$$f_a = 2.5 \text{ cm} \qquad f_b = -2.5 \text{ cm} \qquad f_c = 25 \text{ cm}$$

and some PVC pipes with just the right diameter to permit the lenses to be affixed to the ends that can be cut to any desired length.

a) Design an *inverting* pirate telescope with fixed focus at "infinity" using lenses from your kit. Sketch the lenses in place in the ends of the pipe, and indicate the length of the pipe you plan to have cut. ***Draw the correct ray diagram*** used to minimally compute the angular magnification of your telescope in the small angle approximation and enter its magnification below.

b) You decide that real pirates hate looking at their future prizes upside down. Design a *non-inverting* pirate telescope with fixed focus at "infinity" using lenses from your kit. Sketch the lenses in place in the ends of the pipe, and indicate the length of the pipe you plan to have cut. ***Draw the correct ray diagram*** used to minimally compute the angular magnification of your telescope in the small angle approximation and enter its magnification below.

$M_a =$ $\boxed{}$ $M_b =$ $\boxed{}$

Problem 10.

For a physics lab, you are given several each of the following lenses:

$$f_a = 2 \text{ cm} \qquad f_b = -2.5 \text{ cm} \qquad f_c = 2.5 \text{ cm}$$

and some PVC pipes that can be cut to any desired length and have just the right diameter to permit the lenses to be glued into the ends.

a) Design an *inverting* microscope with overall magnification of $M_1 = -50$ using lenses from your kit. Sketch the lenses in place in the ends of the pipe, and indicate the length of the pipe you plan to have cut.

b) Determine the correct object distance s for a (small) object to be placed in front of the objective (first) lens of your microscope to form a real image at the correct place.

c) ***Draw the correct ray diagram*** for a suitable small object placed at this location that is most easily used to minimally compute the total magnification of your microscope in the small angle approximation and prove that it equals your target value M_1.

d) You get tired of trying to have the reverse-view *chaos carolinensis* ameoba you are looking at on a slide ooze out of your field of view *in* the direction you then move the slide to try to keep them there and decide to fix it so you won't have to. Design a *non-inverting* microscope with overall magnification $M_2 = +50$ using lenses from your kit and a new-cut tube. Sketch the lenses in place in the ends of the pipe, and indicate the length of the pipe you plan to have cut.

e) ***Draw the correct ray diagram*** used to minimally compute the angular magnification of your telescope in the small angle approximation and demonstrate that it equals your target value of M_2.

Week 12: Interference and Diffraction

- **Huygen's Principle:** Each point on a wavefront of a propagating harmonic wave acts like a ***spherical source*** for the future propagation of the wave. This is the basis of our understanding of interference and diffraction of waves through slits, circular holes, and around other kinds of obstacles.

- Note well that ***waves do not travel in straight lines*** when the pass around or through obstacles or holes through obstacles that are ***of the same general order of size as the wavelength or less!*** Waves are perfectly happy travelling around corners (as anyone who has ever watched water waves in a lake or the ocean will attest).

- **Coherence:** A wave is said to be ***coherent***[154] if it has a single frequency over a long enough distance (time)) that path difference (time difference) equals phase difference. The **coherence time** of a wave is the largest such time where this is true, and the **coherence length** is similarly the largest such path difference, typically c times the coherence time.

- The coherence time $\tau_{\rm coh}$ of a typical hot source (such as a light bulb) is anywhere from few tens or hundreds of periods

- The coherence length of a laser can be as long as meters.

- **Two Slit/Point Source Interference:** If one has two coherent, monochromatic sources that are within one another's coherence length (typically very narrow slits that are illuminated by a single source of plane waves) then the intensity received by a *distance* (compared to slit spacing and wavelength) screen is given by:

$$I(\theta) = 4I_0 \cos^2(\delta/2)$$

where

$$\delta = kd\sin(\theta)$$

is the phase difference between the light waves from the two slits. In this expression, I_0 is the central maximum light intensity from either of the two slits/sources alone.

[154] Wikipedia: http://www.wikipedia.org/wiki/Coherence (physics). This is a lovely review article on coherence times and lengths that goes far beyond the remarks below.

- One can easily find the angles θ where maxima and minima in this interference pattern occur.

  Heuristically: The maxima occur where the path difference between the two slits, $d\sin(\theta)$, equals an integer number of wavelengths (so the light from the two slits/sources arrives at the screen ***in phase***. The minima occur where the path difference contains a half integral number of wavelengths, so the light arrives at the screen exactly ***out of phase***.

  By Inspection or Calculus: By inspection, the maxima in the expression for $I(\theta)$ above occur when $\cos(\delta/2) = \pm 1$ and the minima occur when $\cos(\delta/2) = 0$. Alternatively, one can differentiate it with respect to δ and set the derivative equal to zero and solve for δ for max's or min's that way.

  Either path leads one to:

$$d\sin(\theta) = m\lambda \qquad \text{Maxima}$$

$$d\sin(\theta) = (m + \frac{1}{2})\lambda \qquad \text{Minima}$$

  with $m = 0, \pm 1, \pm 2, \pm 3....$

- **N-slit Interference:** When there are multiple slits, they will all arrive in-phase at the screen when:

$$\delta = kd\sin(\theta) = \frac{2\pi}{\lambda} d\sin(\theta) = m(2\pi)$$

  or

$$d\sin(\theta) = m\lambda$$

  for $m = 0, 1, 2....$ At these ***principle intensity maxima*** the field amplitude is N times the amplitude of a single slit, so that the intensity is:

$$I = N^2 I_0$$

  where I_0 is the intensity produced by a single slit.

- If we use phasors to search for heuristic minima and ***secondary maxima***, we find that we get (zero) minima when the phasors form a closed N-gon. This occurs when:

$$\delta = n\frac{2\pi}{N}$$

  for (**note well!**) $n = \cancel{0}, 1, 2, ..N-1, \cancel{N}, N+1, N+2, ..2N-1, \cancel{2N}, 2N+1...$ The crossed out numbers represent places where δ is an integer multiple of 2π, but those are where the ***principle maxima*** occur, not another minimum! Secondary maxima will occur *approximately* half way in between these minima, when:

$$\delta = (n + \frac{1}{2})\frac{2\pi}{N}$$

for (**note well!**) $n = \cancel{0}, 1, 2, ..N-1, \cancel{N}, N+1, N+2, ..2N-1, \cancel{2N}, 2N+1...$ Finding the exact angles for the maxima, however, requires solving a transcendental formula as there is a small trade off between unwinding the phasors a bit and the resultant length.

- **Rayleigh's Criterion for Resolution:**

  > *Two (principle) maxima produced by diffraction (or interference using a misnamed N-slit* diffraction grating*) are considered resolved if the angle for the maximum of either one is separated from maximum the other by at least the angle to the other's first minimum.*

  If this criterion is satisfied, there is a resolvable dip in intensity in between the two separate maxima. If the two maxima are any closer, there is just one broad central maximum and one cannot tell that the images of the two source points or wavelengths are distinct (that is, one cannot tell that there are two source points there *at all* from the image).

- **The Diffraction Grating:** If one illuminates N slits with the distance between adjacent slits d (such that all N slits are within the coherence length of the light) then different wavelengths in the light source have principle maxima at different angles for any given order. This can be used to perform experimental spectroscopy and invert the observation as a *measurement* of the wavelengths of the light in the source. From the discussion of N-slit interference, we know that the principle maxima are brightened by a factor of N^2 relative to the light from a single slit and that these maxima occur at the angle(s) where:

  $$d\sin(\theta) = m\lambda$$

  for $m = 0, 1, 2....$

- The resolving power R of a diffraction grating depends on the order of the maximum. In the small angle approximation,

  $$R = mN = \frac{\lambda}{\Delta\lambda_{\min}}$$

  where $\Delta\lambda_{\min}$ is the ***minimum separation in wavelength that can be resolved*** according to the Rayleigh criterion from the wavelength $lambda$, at any given order m. Inverting this:

  $$\lambda_{\min} = \frac{\lambda}{mN}$$

  so that resolution improves (closer wavelengths can be resolved) with both the number of slits and the order of the maxima being resolved.

- **Single Slit Diffraction:** The intensity of light of wavelength λ passing through a single slit of width a to strike a distant screen is:

  $$I(\theta) = I_0 \left(\frac{\sin(\phi/2)}{\phi/2}\right)^2$$

where the phase angle $\phi = ka\sin(\theta)$ and where θ is, as usual, the angle from the center of the slit to the point on the screen. The phase angle ϕ can be thought of as the phase difference between light from the first Huygens radiator on one side of the slit and light from the last Huygens radiator on the other side of the slit, the difference accumulated across the *width* of the slit.

- A simple heuristic (described in the text) can be used to show that *minima* occur in this "diffraction pattern" (the intensity function given above) when:

$$a\sin(\theta) = m\lambda$$

for (**note well!**) $m = 1, 2, 3....$ Note the omission of $m = 0$. This is because $\theta = 0$ (corresponding to $m = 0$) is always the position of the ***central maximum*** of the diffraction pattern, with peak intensity I_0.

- In between the minima given at these *exact* angles are secondary maxima of strictly descending intensity at the *approximate* angles:

$$a\sin(\theta) = (m + \frac{1}{2})\lambda$$

As was the case for N-slit intereference secondary maxima, however, the exact angles of the secondary maxima requires the solution of a transcendental equation and not a formula as simple as this.

- **Combined Interference and Diffraction:** If one takes (e.g.) two slits, each of width a, separated by a distance $d > a$ and illuminated by light with wavelength λ, the intensity on a distant screen is given by:

$$I(\theta) = 4I_0 cos^2(\delta/2)\left(\frac{\sin(\phi/2)}{\phi/2}\right)^2$$

The resulting intensity is the usual two slit interference pattern, *modulated* by the so-called "diffraction envelope" of each slit indepedently.

- **Diffraction Through Circular Apertures and Optical Instruments:** A circular aperture produces a diffraction pattern that qualitatively resembles that of a single slit with an *axially symmetric* central maximum surrounded by rings of minima and ever-fainter secondary maxima. In many cases it is this diffraction of light from small or distant source points as it passes through the *objective lens* of a microscope or telescope (respectively) that limits the resolution of optical instruments. One can, of course, magnify objects almost without bound as far as geometric optics is concerned, but at some point diffraction makes further magnification pointless because neighboring source points in the field of view are no longer resolvable according to the Rayleigh criterion at any greater magnification.

The angle of the ***first minimum*** (dark ring around the central maximum) produced by a given wavelength of light is determined by the formula:

$$D\sin(\theta_{\min}) = 1.22\lambda$$

where D is the diameter of the circular aperture of the optical instrument. It is beyond the scope of this course to derive this, but it is "reasonable" as an approximation of the single slit result above. In almost all cases, we are only interest in using this when the angles involved are very small, in which case we can write:

$$\theta_{\text{min}} = 1.22\frac{\lambda}{D}$$

- The Rayleigh criterion for wave-optic resolution with an optical instrument is then simply that the angle between the two source points as they enter the first lens of the microscope or telescope must exceed the angle to the first minimum of either one, or:

$$\alpha_{\text{incidence}} > \theta_{\text{min}} = 1.22\frac{\lambda}{D}$$

- **Thin Film Interference:** Light that strikes a thin transparent partially reflective film on top of a second reflective medium can interfere with *itself* provided that the film is thin enough that the total path difference between light reflected from the first versus the second surface is inside the coherence length of the light. Thin film interference is what makes soap bubbles and a drop of oil on water on dark pavement swirl with odd pastel colors.

- To understand this, note that when light reflects from an interface between a medium with a lower index of refraction (source) and a medium with a higher index of refraction (destination) the reflected wave ***inverts*** (shifts its phase by π or a half-wavelength). When light refects from an interface between a medium with a lower index (source) moving towards a higher index (destination) the reflected wave does ***not*** invert its phase. Note that we learned precisely these rules for wave pulses reflected from the interface between light string and heavier string or vice versa in the first part of this course.

- Second, the *transmitted* light that is partially reflected and partially transmitted at the first surface of the thin film has to travel to the second surface through the film (typically a distance given as d, not to be confused with the distance between two slits above) and then back to the first surface again, where the wave that is partially transmitted here recombines with the original reflected wave. The light that went into the film thus travels an (approximate) additional distance of $2d$, and we can use the heuristic rule above to determine whether or not we get constructive interference (brightening of some given wavelength) or destructive interference (partial cancellation and dimming of some given wavelength), ***if*** we also account for the discrete phase shift(s) at the interfaces.

- Let $n_1 < n_2 < n_3$ **or** $n_1 > n_2 > n_3$, where by convention we will use 123 to indicate the order of the media in the direction of the incoming light. Then there are either two phase shifts of π (first case) or no phase shifts of π (second case) at the two reflecting surfaces of the middle layer, and the phase difference is due ***only*** to the path difference ***in the film medium with index of refraction*** $\boldsymbol{n_2}$. The heuristic rule is then:

$$2d = m\lambda' = m\frac{\lambda}{n_2} \qquad \text{Maxima}$$

$$2d = (m + \frac{1}{2})\lambda' = (m + \frac{1}{2})\frac{\lambda}{n_2} \qquad \text{Minima}$$

with $m = 0, \pm 1, \pm 2, \pm 3...$ as usual. **Note Well:** the use of $\lambda' = \lambda/n_2$, the path difference ***in the medium*** must contain an integer number of wavelengths for the reflected light that emerges back into n_1 to be in phase.

- A special result occurs when $d \ll \lambda$. In this case there is "no" path difference, and the waves emerge in phase for *all* wavelengths. The surface becomes "shiny". You can observe this when a drop of oil spreads out on water on dark pavement – at first there are many colors and then the surface takes on a silvery grey sheen.

- Let $n_1 < n_2 > n_3$ **or** $n_1 > n_2 < n_3$. Then there is only one phase shift of π at the first surface (first case) *or* one phase shift of π at the second surface (second case), and the total phase difference is that from the path difference plus an additional phase of π. This is equivalent to half a wavelength difference. The heuristic rule then ***reverses:***

$$2d = (m + \frac{1}{2})\lambda' = (m + \frac{1}{2})\frac{\lambda}{n_2} \qquad \text{Maxima}$$

$$2d = m\lambda' = m\frac{\lambda}{n_2} \qquad \text{Minima}$$

with $m = 0, \pm 1, \pm 2, \pm 3....$

- A second special result occurs when $d \ll \lambda$. In this case there is "no" path difference, and the waves emerge exactly ***out of phase*** by π for all wavelengths. The surface becomes perfectly non-reflective, hence transparent. You can observe this when a soap bubble has persisted long enough for most of its water to evaporate – as it becomes thinner than the wavelengths of visible light, it becomes almost perfectly transparent and invisible. This is also used to make nonreflectiving coatings for glass and lenses to maximize their light transmission.

12.1: Harmonic Waves and Superposition

Several weeks ago we learned about **harmonic waves**, solutions to the wave equation of the general form (in one dimension):

$$\vec{\boldsymbol{E}}(x,t) = E_0\hat{\boldsymbol{e}}\sin(kx - \omega t) \tag{12.1}$$

where $\hat{e}$ is a unit vector in the direction of the wave's polarization. Waves spreading out spherically symmetrically in three dimensions from a source with radius a have a similar form:

$$\vec{\boldsymbol{E}}(r,t) = E_0\frac{a}{r}\hat{\boldsymbol{e}}\sin(kr - \omega t) \tag{12.2}$$

(where $|\vec{\boldsymbol{E}}(a,t)| = E_0$ is the field strength at the surface of the source for this component of the polarization). Recall also that we only need to write the electric field strength because the

associated magnetic field has an amplitude of $B_0 = E_0/c$, is in phase, and is perpendicular to the electric field so that the Poynting vector:

$$\vec{\boldsymbol{S}} = \frac{1}{\mu_0}\vec{\boldsymbol{E}} \times \vec{\boldsymbol{B}} \tag{12.3}$$

points in the direction of propagation. Finally, don't forget that the (time averaged) intensity of the wave is:

$$I_0 =< |\vec{\boldsymbol{S}}| >_{\text{av}}= \frac{1}{2\mu_0}E_0B_0 = \frac{1}{2\mu_0 c}E_0^2 \tag{12.4}$$

We also learned about ***Huygen's principle***, which states that each point on a wavefront of a propagating harmonic wave acts like a ***spherical source*** for the future propagation of the wave. This will prove to be a key idea in understanding interference and diffraction of waves that pass through slits, the **superposition principle**, which says that to find the total field strength at a point in space produced by waves from several sources we simply add the field strengths from all the sources up, and one of the ideas underlying Snell's law, that the wavelength of a wave of a given fixed frequency depends on the index of refraction of the medium through which it propagates according to:

$$\lambda' = \frac{\lambda}{n} \tag{12.5}$$

where λ is the wavelength in free space; the wavelength of a wave is *shorter* in a medium with an index of refraction greater than 1 so that the wave slows down. All of these things that we have already learned will be important in our development of interference and diffraction.

In addition to these old concepts, we will require one or two new ones. One is the idea of a *hot source*. A hot source is something like the hot filament of a light bulb, the hot flame of a candle, the hot gasses on the surface of the sun, all *so* hot that they glow and give off light. Even the gasses in a relatively cool fluorescent tube are "hot" in the sense we wish to establish, as the atomes that are giving off the light are very weakly correlated with one another.

12.1.1: Hot Sources and Wave Coherence

Although we've see that Maxwell's equations in free space become the electromagnetic wave equation (so that light is plausibly and electromagnetic wave) we haven't spent much time considering how light arises in the first place, how charges can end up *emitting* electromagnetic waves. The bulk of our understanding came from thinking about a Lorentz model atom – an electric dipole moment that harmonically oscillates, producing an electric field that propagates and oscillates, inducing its companion magnetic field as it goes to produce a wave.

That's pretty much how it (classically) goes, so this isn't a bad thing. We also get electromagnetic radiation (usually at radio frequencies) if we make a magnetic dipole moment oscillate in time, for example by putting an alternating current into an antenna consisting of N circular turns of wire, but radiation from atoms is predominantly electric dipole radiation. The only "catch" is that the radiation is a *quantum* process and hence only comes out of the atoms in particular frequencies and "all at once" instead of continuously and at varying frequencies as we might expect classically.

There are two general *kinds* of sources we need to be concerned with when dealing with electromagnetic waves and superposition leading to interference and diffraction: ***Coherent*** and ***Incoherent***. These are both relative terms – no causal, periodic source of electromagnetic waves is perfectly coherent or perfectly incoherent (it would have be periodic over an infinite amount of time to manage this, which seems infinitely unlikely in a "messy" Universe), and ultimately source coherence is thus described by a real number that can vary over some range.

A source is said to be coherent if:

a) It is (approximately) monochromatic (or at least, a fixed mixture of frequencies that are independently otherwise coherent).

b) The waves emitted by these source are *ideally harmonic*, that is, their phase temporally accumulates as ωt for the fixed frequency ω and with a constant additional phase, if any.

The latter implies the former, as you can see.

Coherence, we see, is implicit in our writing down (an x-polarized harmonic wave propagating in the z direction):

$$\vec{\boldsymbol{E}}(z,t) = E_{0x}\hat{\boldsymbol{x}}\sin(kz - \omega t) \tag{12.6}$$

An ordinary monochromatic harmonic wave is perfectly coherent.

To understand why coherence is important to us, let us consider what a "harmonic" wave might look like that is *not* coherent[155]:

$$\vec{\boldsymbol{E}}(z,t) = E_{0x}\hat{\boldsymbol{x}}\sin(kz - \omega(t)t + \phi(t)) \tag{12.7}$$

In this wave I have illustrated two common sources of incoherence. One is a frequency that isn't really constant in time but e.g. slowly varies in such a way that it has some constant *average* value, e.g.

$$\omega_{\rm avg} = \lim_{T\to\infty}\int_0^T \omega(t)dt \tag{12.8}$$

that is, it might be *approximately* constant over a time that is long compared to a period of the wave, perhaps several thousands or millions of those periods, but on shorter times it might vary within some range. This variation might be caused by e.g. thermal fluctuations in the source, by thermal doppler shifting of a sharp natural frequency in a gas, or by still other things (including humans, who amplitude or frequency modulate a carrier wave to encode information).

In nature, not even quantum sources have infinitely sharp frequencies, so even "monochromatic" light is only *approximately* monochromatic or monochromatic within some bandwidth or range[156], and the variation over longer time scales may be sufficient to cause *temporal* interference (beats) instead of the *spatial* interference we will examine in this chapter when waves that follow different paths from a common source are recombined.

[155] And hence, of course, not perfectly harmonic or monochromatic! Students who have taken more advanced math can understand this in terms of the ***Fourier transform*** of the wave above, which will ***not*** be a Dirac delta function of any single frequency but rather will involve a ***band*** of frequencies around a peak at $\omega_{\rm avg}$. This in turn takes us back to the discussion of amplitude modulated waves from the AC Circuits chapter above compared to ***frequency modulated*** waves that can also be used to carry encoded information. Deep waters underlie these simple concepts.

[156] We speak of "line broadening" and the "natural width" of spectral lines to acknowledge or quantify this.

The other source of incoherence is the phase angle $\phi(t)$. We recall that when we solved the wave equation we could add an arbitrary phase constant to the argument of the harmonic wave and we'd still have a harmonic wave. Basically, that constant simply indicated when we "started our clock", and we could more or less choose to use a sine wave or cosine wave with no phase at all by starting our clock appropriately when examining or describing the wave.

The problem is that for many sources, especially *hot* sources, this clock gets *reset* whenever the oscillators that are producing the wave are physically disturbed or re-energized (the oscillation necessarily damps out over time as the energy in the oscillator is radiated into the electromagnetic field). There is no reason to expect that the phase of the oscillator producing the light will be constant over time indefinitely. Indeed, we rather expect the opposite!

The simplest model for "hot source" incoherence is that of *phase interruption*. We imagine a sample of some element that is hot enough so that when an atom collides with a neighbor it excites some particular oscillator state with a fixed frequency and a phase determined by the time of the collision. It then oscillates monochromatic light with a phase and polarization direction determined by the time and angle of that collision. Eventually, however, the atom collides again, and although the same oscillator state is re-excited and light of the same frequency emerges, it has a (discretely) different phase and direction of polarization!

In this (most common) case, the hot "monochromatic"[157] source is temporally phase coherent only for the mean time between collisions, which in turn depends on things like the density of the material and its temperature. Although our mental picture of "collisions" is simplest to envision for a fluid like a liquid or gas, related (e.g. phonon based) events also phase interrupt the wavetrains emitted by hot solids, and again there is a characteristic *average time* between such phase interruption events.

The effect of these phase interruptions is such that when adding the electric fields of two completely incoherent sources, no interference or spatial diffraction is observed to occur – the intensities of the different sources simply add because the fields themselves add for a few cycles, then cancel for a few cycles, then add, then cancel, in such a way that the average energy transmitted smooths out and just adds. Temporal incoherence over long time scales ***destroys spatial interference patterns and replaces them with mere average intensity addition***[158]***!*** This is ***very important*** – it is the reason we don't see interference patterns all the time, e.g. why windowpanes and drinking glasses don't exhibit ***thin*** film interference like that discussed below! Whenever we add two harmonic waves to get a harmonic wave as a result, we are implicitly assuming *coherence*.

Hot sources are thus coherent, but only over a *comparatively short time*. We use the heuristic arguments above to *define* the time over which a hot source (or any source) will remain coherent – the *coherence time*: $\tau_{\rm coh}$. For most hot sources in the visible band of frequencies, the coherence time is on the order of a few tens to hundreds of optical periods. A reasonable round number might be:

$$\tau_{\rm coh} \approx 10^{-12} \text{ seconds} \tag{12.9}$$

(given frequencies in the range of 10^{14} to 10^{15} cycles per second).

[157] In quotes because the fourier transform of a harmonic wave with random phase interruption is no longer sharp or monochromatic.

[158] All of this is proven in more advanced mathematical treatments.

Light, of course, doesn't travel very far in such a short time. We can define the *coherence length* of light as the distance light travels in the coherence time:

$$L_{\rm coh} = c\tau_{\rm coh} \approx 10^{-4} \text{ meters} \tag{12.10}$$

In all of the text below, we will therefore assume that all of the relevant length scales (such as the maximum path difference in interference problems) is smaller than 0.1 millimeter, or 100 microns. For slit separations or film thicknesses much larger than this, interference will generally be washed out by the random phase shifts associated by hot sources.

Coherent sources in the range of frequences that we might generally call "radio waves" of all sorts are common as dirt in our society. Every device that transmits energy and information over a carrier frequency to a remote receiver relies on the coherence of the transmitted wave to permit information to be encoded on top of that wave.

Coherent sources in the optical regime are correspondingly *rare* and for all practical purposes there is just one source of coherent optical radiation – the laser. The laser is nearly unique as a source of monochromatic coherent light. Lasers typically have coherence lengths measured in *meters*. Lasers are so coherent that light from two *different* lasers produces a stable interference pattern. Laser light can be split and sent along two very different path lengths and still interfere. This is the basis of laser holography [159] , the ring laser gyroscope[160] and laser interferometry[161] .

All other sources of visible light generally rely on *atoms* to produce the actual light, most often atoms that are *hot*, hot enough to glow as they thermally bounce off of each other at high speed, exciting various electric "oscillators" in their quantum structure. The sun is a very hot source (surface temperature around 5778 °K). Incandescent bulbs produce light from a hot tungsten filament that is joule heated to some 3600 °K. Fluorescent bulbs operate much cooler – the optimum bulb temperature is around 313 °K (40 °C or 104 °F) but are still "hot" in the sense of thermally random and chaotic.

Finally, one of the most recent developments in electrical lighting is the increasing prevalence of light emitting diodes (LEDs) as commercially important sources of light. LEDs actually operate at room temperatures and are so efficient that their temperature generally doesn't greatly exceed the ambient temperature – nearly all of the energy delivered to them emerges as light. LEDs are usually more or less monochromatic, emitting light at particular wavelengths determined by the quantum properties of the semiconductors that make up the diode. In this they are *almost* identical to solid state diode-based *lasers*, except in the one important regard – they are still "hot" incoherent sources.

Pay careful attention to coherence as you work through interference and diffraction below. Remember, even hot (monochromatic) sources will usually produce interference when the light being summed is within the mutual coherence time/length of the light source in question, and even white light from hot sources – as a mixture of many frequencies that are *all* coherent over

[159] Wikipedia: http://www.wikipedia.org/wiki/Holography. This is actually a fascinating topic and a great thing for someone seeking an extra credit project to try out. It does, however, require a laser, film and a darkroom, and a very, very solid/motionless lab bench to use as a base, and probably won't work the first time you try it.

[160] Wikipedia: http://www.wikipedia.org/wiki/ring laser gyroscope.

[161] Wikipedia: http://www.wikipedia.org/wiki/interferometry.

similar $L_{\rm coh}$ – can be locally sufficiently coherent to support e.g. thin film interference in all of the colors/frequencies independently.

12.1.2: Combining Coherent Harmonic Waves

The unifying idea of this entire chapter is then: Monochromatic coherent light from some source follows two (or more) different paths to reach a detector (e.g. – an eye, a screen observed by an eye, a piece of film, a photoelectric detector). Along the way it accumulates *phase differences* between the waves due to the different path lengths that they follow (and possibly other things such as reflection that introduce phase shifts discretely along the way). The electric (and magnetic) fields then *recombine*, and the intensity of the resulting electromagnetic field is registered by the detector.

Provided that the maximum path differences involved are less than the coherence length $L_{\rm coh}$ of the light, we will then have to repeatedly evaluate below sums such as (for a single polarization component of the wave):

$$E_{\rm tot} = E_1 \sin(kx - \omega t + \delta_1) + E_2 \sin(kx - \omega t + \delta_2) + E_3 \sin(kx - \omega t + \delta_3) + ... \qquad (12.11)$$

where the phase shifts δ_i are all determined by the path differences plus discrete shifts.

It is too difficult to solve this equation generally. Instead we will make a variety of simplifying assumptions that are all reasonably valid in the context of the following specific topics. The primary ones will be that we will generally assume that all of the field amplitudes are the same (although we could certainly deal with specific cases where they are different in some simple way using the methodology we develop). We will usually set *one* of the phases e.g. δ_1 to be zero (setting our clock, as it were, by the first source). The other phase differences δ_i will usually be assumed to be constant in time (the light from all of the paths is perfectly coherent at the time of recombination).

With those assumptions, we can usually reduce the *algebraic* problem of adding the harmonic waves to the simpler *geometric* problem of adding two or more make-believe *vectors*, called *phasors*. Phasor addition will simplify the problem of finding the interference and diffraction patterns produced by idealized slits and apertures to where it is straightforward, if not quite easy.

Along the way we will also endeavor to establish some very simple *heuristic rules* that enable one to determine where interference or diffraction patterns are *maximum*[162] or *minimum*.

The heuristic rules are worth stating here, although we'll repeat them many times below. One will generally get interference *maxima* when the waves arrive at the detector *in phase*, which in turn means that the path difference will contain an *integer number of wavelengths* (and still be less than $L_{\rm coh}$). One will get interference *minima* when the path difference contains an *odd half-integer number of wavelengths* so that the waves arrive exactly *out of phase*.

Tres simple, no?

[162]Since this is a common enough point of confusion, let me make it clear that the term **maximum** in interference or diffraction problems already refers to a **maximum in intensity at the point of observation of the e.g. interference pattern**, not "maximally interfering" and hence of *minimum* intensity. Similarly a **minimum** refers to the minimum (usually zero) in the interference or diffraction intensity at the receiver, on the screen, to the eye.

Let's start with the simplest of interference problems: Two Slit Interference.

12.2: Interference from Two Narrow Slits

The first, and simplest, example of interference is monochromatic (constant wavelength) light falling upon two extremely narrow (slit width less than the wavelength of the light) separated by a distance d that is order of a few wavelengths in size. Because the slits are so close together, they are within the correlation length even of most (monochromatic) hot sources, so that two slit interference patterns can easily be produced.

To compute the interference pattern produced by two slits, we begin by examining figure (12.1), wherein light of fixed wavelength λ falls normally onto a blocking screen through which two narrow slits have been cut. Each slit is so narrow that it acts like a "point" Huygens radiator. Light from one slit (the upper) travels a long distance and falls on a distant screen. Light from the lower slit travels this distance plus the *additional* distance $d\sin(\theta)$ to arrive at the same point.

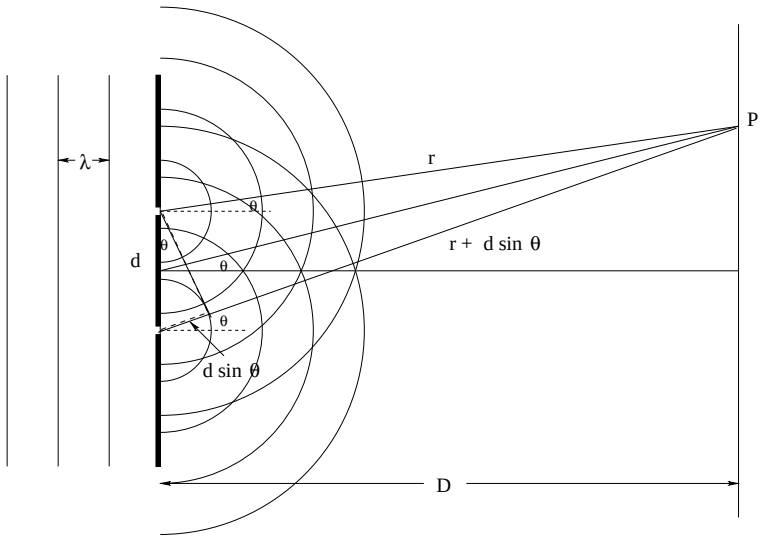


Figure 12.1: Two narrow slits act as Huygens radiators when indident plane wavefronts fall upon them. Light from the two slits is *coherent* and *in phase* as it leaves the slits, but arrives at P with a phase difference that depends on the path difference.

As long as the distance D between the two slits and the screen is much larger than d the distance between the slits themselves then the angle θ between the horizontal line shown and *both* paths to the point of observation P is the *same* (although this is not visibly the case in the figure, where D is not sufficiently large compared to d). The condition $d \ll D$ is called the **Fraunhofer condition** and must be compared to the **Fresnel condition** which evaluates interference patterns "close to" the slits where the simplifying Fraunhofer condition does not hold. Fresnel patterns can "easily" be evaluated as well, but the evaluation requires methodology that is beyond the scope of this course.

Light from the top slit travels a distance r to arrive at point P. Light from the bottom slit travels a distance $r + \Delta r = r + d\sin(\theta)$ to arrive at the point P. $r \geq D$ and $d\sin(\theta) \leq d$, so

$r \gg \Delta r$. We can therefore find the total electric field at P by adding the electric fields produced by each slit. Let us call the amplitude of the electric field produced by a single source in the center of the screen E_0. Then the total field at point P is:

$$\begin{aligned}
E_{\rm tot}(P) &= E_0 \frac{D}{r}\sin(kr-\omega t) + E_0 \frac{D}{r+\Delta r}\sin(kr+k\Delta r-\omega t) \\
&= E_0 \frac{D}{r}\sin(kr-\omega t) + E_0 \frac{D}{r}\left(1+\frac{\Delta r}{r}\right)^{-1}\sin(kr+k\Delta r-\omega t) \\
&= E_0 \frac{D}{r}\sin(kr-\omega t) + E_0 \frac{D}{r}\left(1-\frac{\Delta r}{r}+...\right)\sin(kr+k\Delta r-\omega t) \\
&= E_0 \frac{D}{r}\sin(kr-\omega t) + E_0 \frac{D}{r}\sin(kr+k\Delta r-\omega t) + \mathcal{O}\left(\frac{\Delta D}{r}\right) \\
&\approx E_0 \sin(kr-\omega t) + E_0 \sin(kr-\omega t+\delta) \qquad (12.12)
\end{aligned}$$

The last step follows because *for a small angle* θ:

$$r = \frac{D}{\cos(\theta)} \approx \left(\frac{D}{1-\frac{\theta^2}{2}+...}\right) \approx D\left(1+\frac{\theta^2}{2}+...\right) \approx D \qquad (12.13)$$

so $E_0 D/r \approx E_0$ for *both* sources. Obviously this will not hold for large θ (angles pointing out at the edges of a large screen stretching to infinity on the horizon), nor will it hold if the screen is close to the two slits (where *Fresnel* interference or diffraction must be considered, which is a lot more work and beyond the scope of this course although answers there are certainly computable). In the last equation we also introduce the *phase shift produced by the path difference*:

$$\delta = k\Delta r = kd\sin(\theta) = \frac{2\pi d}{\lambda}\sin(\theta) \qquad (12.14)$$

To add these two waves, we could use a trigonometric identity for $\sin A + \sin B$. Unfortunately, nobody can ever remember the trig identities for things like this (supposedly memorized back in high school), including me. For those of us who find it impossible to remember arbitrary things we memorized out of any context where they would be useful to us for more than busy work, it behooves us to learn how to *derive* the answer in simple ways from things we *can* remember and that make sense in context. We therefore eschew the use of a trig identity and *derive* the result from a geometric picture, a *phasor diagram* just as we did before for e.g. LRC circuits.

In figure (12.2) we see the requisite phasor geometry. The light from the first slit has a field amplitude of the y-component of a "vector" (phasor) of length E_0 at angle $kr-\omega t$ with respect to the x-axis. The light from the second slit is the y-component of a phasor of length E_0 at angle $kr-\omega t+\delta$. The field amplitude of the sum is the y-component of the phasor that is the vector sum of these two phasors, added by putting the tail of the second at the head of the first. Since the triangle representing this sum is isoceles it is easy to see that the two acute angles must both be $\delta/2$[163]. The total amplitude is thus the sum of the adjacent side lengths

[163]The argument goes as follows: "δ plus the obtuse angle at the vertex of the triangle form a straight line and hence add up to π. The sum of the angles in the triangle also add up to π. Therefore the sum of the two acute angles have to add up to δ. The triangle is isoceles, so they must be equal, hence they are each $\delta/2$." This is why geometry is *better* than algebra or trig – proving this algebraically is nearly impossible without the use of complex variables and with trig identities it is difficult and requires knowing the relevant identity.

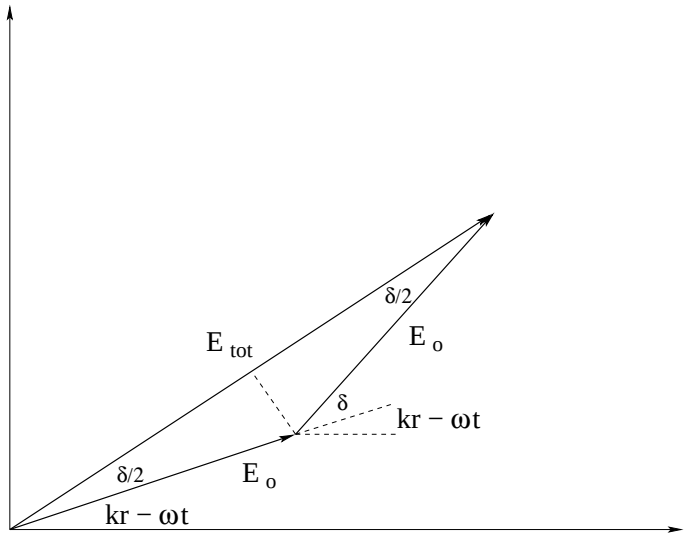


Figure 12.2: Phasor diagram for the addition of the electric field components of two slits.

of the two right triangles formed by dropping a normal as shown:

$$|E_{\text{tot}}| = 2E_0 \cos(\delta/2) \tag{12.15}$$

and the full time dependent electric field is given by:

$$E_{\text{tot}} = 2E_0 \cos(\delta/2) \sin(kr - \omega t + \delta/2) \tag{12.16}$$

We don't actually care about the *field strength*, of course – we care about the *intensity*. The time-averaged intensity of light from a *single* slit at the point P is equation 9.134 obtained in the section studying the Poynting vector above:

$$I_0 = \frac{1}{2}\epsilon_0 c |E_0|^2$$

The total intensity from the pair of slits is therefore:

$$I_{\text{tot}} = 4I_0 \cos^2(\delta/2) \tag{12.17}$$

While this expression is entirely correct, it is not very general – it works *only* for two slits! In a short while, we'll solve the N-slit problem exactly and obtain a solution that (for $N = 2$ slits) can be written as:

$$I_{\text{tot}} = \left(\frac{\sin(\delta)}{\sin(\delta/2)}\right)^2 I_0 \tag{12.18}$$

One of your (very short) homework problems will be to use a trig identity to prove that these two forms are equal to one another. This is the equation I actually use to plot out the two slit interference pattern in figure 12.3.

While this is the completely general solution for the two slit problem (within the approximations made above) we are often most interested in finding the specific angles θ where the interference is *maximum* and/or *minimum*. Clearly the **minima** occur where $\cos^2(\delta/2) = 0$, which are the phase angles:

$$\delta/2 = \pm\pi/2, \pm 3\pi/2, \pm 5\pi/2, ... \tag{12.19}$$

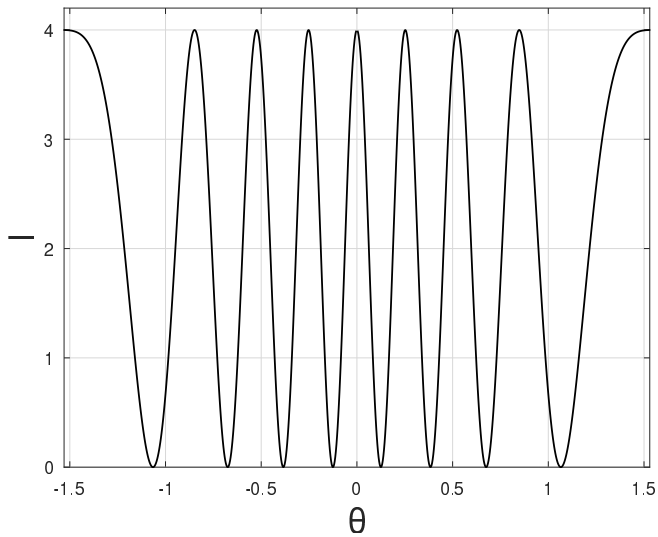


Figure 12.3: The "ideal" interference pattern produced by two narrow slits for $d = 4\lambda$ as a function of θ (***not*** $\sin\theta$, note well, so they are not equally spaced). The minima and maxima occur precisely at the angles predicted above. In order to plot it, the intensity of a single slit was arbitrarily set to be $I_0 = 1$ on the vertical scale in no particular set of intensity units.

or

$$\delta = \frac{2\pi d}{\lambda}\sin(\theta) = \pm(2m+1)\pi \tag{12.20}$$

or the actual angles θ where:

$$d\sin(\theta) = \pm\frac{2m+1}{2}\lambda \tag{12.21}$$

The intensity is zero at the minima.

The maxima occur at the angles where:

$$\delta/2 = 0, \pm\pi, \pm 2\pi ... \tag{12.22}$$

or

$$\delta/2 = \frac{2\pi d}{2\lambda}\sin(\theta) = m\pi \tag{12.23}$$

or the actual angles θ where:

$$d\sin(\theta) = \pm m\lambda \tag{12.24}$$

The intensity is $4I_0$ at the maxima.

The minima and maxima occur at precisely the angles that agree with our heuristic rule from above. We heuristically expect a constructive interference maximum when the path difference $d\sin(\theta)$ contains an integer number of wavelengths, and this is exactly what we get. We heuristically expect a minimum of light from the lower slit travels half a wavelength farther than light from the upper one, or three half wavelengths farther, or five half wavelengths farther, and that's exactly what we get. It's always nice when our intuitive, heuristic expectations are confirmed by the actual algebra of the solution. It gives us confidence that the latter is correct.

12.3: Interference from 2, 3, ... N Narrow Slits

Let's set up 3 slits and sketch a phasor diagram for the problem, use this to guess how e.g. $4, 5, ...N$ slits would look, and use the general phasor diagram for 5 slits (why not?) and inductively magic it into a quantitative result for N slits. Then we'll examine a *general* heuristic for determining maxima, minima, secondary maxima, and more.

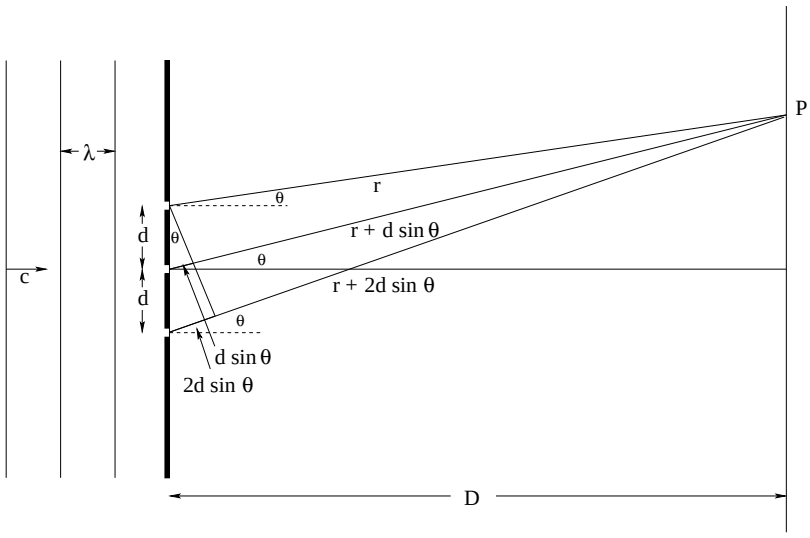


Figure 12.4: Three narrow slits, equally spaced a distance $d > \lambda$ apart, are illuminated by monochromatic light that is coherent over distances long with respect to both d and λ to produce an interference pattern on a distant screen. Note well that the path difference between any adjacent pair of slits is $d\sin(\theta)$.

Three narrow slits, each separated by the same distance $d \gtrsim \lambda$ and centered on the midline running perpendicularly to the (cylindrical) screen a large distance $D \gg d$ away is illustrated in figure 12.4. As you can see by inspection, in this limit the angles from all of the slits to the point of observation P are almost exactly the same, the paths the waves follow to P from the slits are almost perfectly parallel, and hence the path difference between any two adjacent slits is still $d\sin\theta$ just as it was for (only) two slits in the previous section.

The light will arrive at the screen with almost exactly the same amplitude from each slit so we need to evaluate the simple sum of the electric field of the three monochromatic waves, each with a *phase difference determined by the path difference*:

$$E_{\rm tot} = E_0 \sin(kr - \omega t) + E_0 \sin(kr - \omega t + \delta) + E_0 \sin(kr - \omega t + 2\delta) \tag{12.25}$$

where (as before) $\delta = kd\sin(\theta)$ is the phase difference between *any two adjacent slits*. The phasor diagram that represents this sum is drawn in figure 12.5.

Examining the geometry of the phasors in this figure (where I've drawn in most of the clever tricks needed to sum it up), one can see that *one* way of writing the general solution for $E_{\rm tot}(P)$ is:

$$E_{\rm tot} = E_0(1 + 2\cos(\delta))\sin(kr - \omega t + \delta) = |E_{\rm tot}|\sin(kr - \omega t + \delta) \tag{12.26}$$

In this particular case we could then write the ***time averaged*** intensity of of the interference

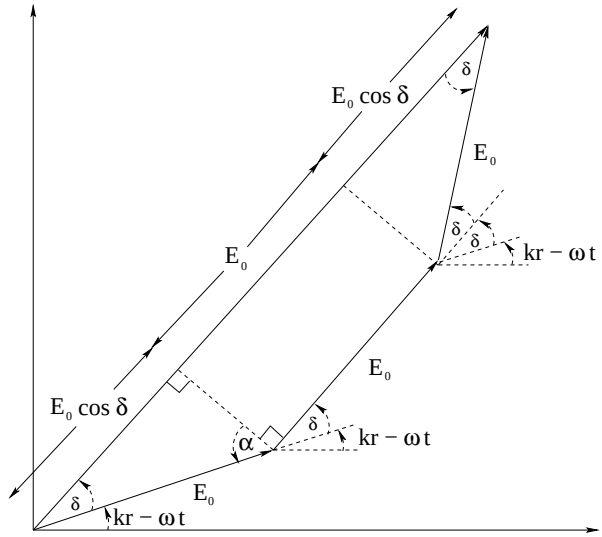


Figure 12.5: The phasor diagram for the total field at point P from three slits.

pattern as:

$$I_{\text{tot}} = \frac{1}{2}\epsilon_0 c|E_{\text{tot}}|^2 = I_0\left(1 + 4\cos(\delta) + 4\cos^2(\delta)\right) \tag{12.27}$$

where:

$$I_0 = \frac{1}{2}\epsilon_0 c|E_0|^2$$

is the intensity of any single slit at P with the other slits closed. As usual, the time average of the $\sin^2(kr - \omega t + \delta)$ piece is $\frac{1}{2}$[164]. With this in hand, we could even go on to find maxima and minima and plot out the result as a function of θ, but...

...there is something *unsatisfying* about this result. Sure, this simple approach worked for *two* slits, and we even made it work for *three* slits (largely because the phasor diagram is a symmetric trapezoid) so we can work plane geometry triangle devil magic at the apices, but what are we going to do for four slits? Five? A *thousand*? This one-off approach simply won't *scale* to an arbitrary number N of identical, thin slits, each separated from its neighbors by d! At the very least, we need to analyze different triangles in a way that scales!

Still, we've made enough progress that we can try a bit of inductive reasoning. We know enough to generalize what we have to add up for an arbitrary number N slits illuminated as usual in such a way that the light transmitted through all of the slits is intially in phase and eventually reaches a cylindrical screen a distance $D \gg d \gtrsim \lambda$ from the center of the set of slits (that is, a set up that satisfied the Fraunhofer condition). I'm not even going to bother drawing the *specific* diagram with N slits and the screen – at this point the idea should be easy to visualize and besides, drawing too many slits in a diagram where D isn't *that* much bigger than d will actually make the large D limit *more* difficult to heuristically visualize.

Hopefully it is now obvious to you that the total field through the N slits arriving at P must

[164]Being appropriately diligent and then lazy, we don't even need to write it out and evaluate after writing it out and explicitly evaluating it enough times, as surely we just *know* this at some point, right?

be given by:

$$\begin{aligned} E_{\rm tot} &= E_0 \sin(kr - \omega t) + E_0 \sin(kr - \omega t + \delta) + E_0 \sin(kr - \omega t + 2\delta) + ... \\ &\quad + E_0 \sin(kr - \omega t + (N-2)\delta) + E_0 \sin(kr - \omega t + (N-1)\delta) \\ &= |E_{\rm tot}| \sin(kr - \omega t + \gamma) \end{aligned} \tag{12.28}$$

where $\delta = kd\sin\theta$ is now and for *any* N in the future the phase angle produced by the path difference between any two adjacent slits, $|E_{\rm tot}|$ is the amplitude of the resultant wave at P, and γ is the difference in phase of the total time dependent wave relative to the phase of light from the first slit. So far, so good, but (sigh) *I*, at least, don't know any trig identity learned in high school that would give me the slightest bit of traction in finding $|E_{\rm tot}|$, the amplitude of the resultant field, from which we can find the resultant intensity on the screen.

Fortunately, there is a *much better, completely general* way of solving this using the geometry of its phasor diagram, a way where the number of slits appears in a simple, countable way. To derive this result, it will make sense to draw this phasor diagram for some N larger than 2 or 3 but still small enough that we can add things up on our fingers. Since we have *five* fingers, let's try five slits!

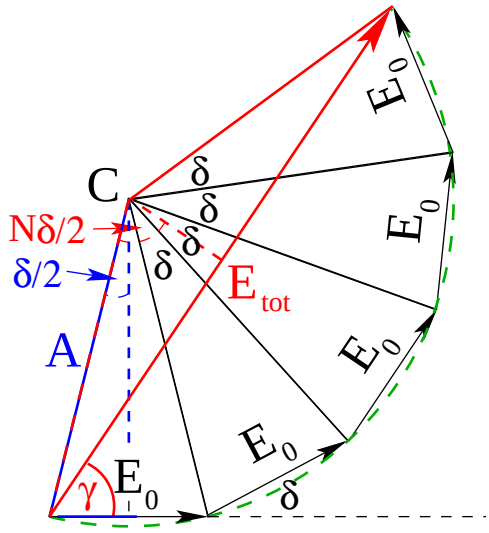


Figure 12.6: The general phasor diagram for $N = 5$ slits. Note that I label the figure in terms of N per se, not the specific value 5! The construction should make it obvious that the relations we deduce from this diagram work for *any* value of $N \geq 2$!

The phasor diagram for the $N = 5$ case is shown in figure 12.6. Let's examine this figure. First, I've set the harmonic angle $\alpha = kr - \omega t = 0$, so that the *first* phasor lines up with the x-axis. This is perfectly fine as it is a common phase to *all* of the individual phasors and the *entire diagram* including especially the (red) $E_{\rm tot}$ phasor will rotate clockwise around the base of the first phasor as t increases. In fact, we can already see what the harmonic term in the total field will look like:

$$E_{\rm tot} = |E_{\rm tot}| \sin(\alpha + \gamma) = |E_{\rm tot}| \sin(kr - \omega t + \gamma)$$

as indicated above, and you can verify that this is what we got for $N = 2$ and $N = 3$ above, for $\gamma = \delta/2$ and δ respectively. At the end of the day ***we're going to time average*** $\boldsymbol{\sin^2(kr - \omega t + \gamma)}$ to get $\frac{1}{2}$ so that the $\alpha = kr - \omega t$ will not appear in the final intensity.

In this figure, I've identified a common "center" C such that the blue dashed vertical line that bisects the first isoceles triangle with C at its apex intersects the similar line bisecting the other N isoceles triangles with long sides A and base E_0. C is their *mutual* apex such that start and end of the E_0 phasors lie on a (green dashed) circle of radius A.

If you imagine rotating the first such triangle through an angle δ clockwise, it ***turns into the second triangle***. *This* triangle, rotated by an additional δ, becomes the third, and the third becomes the fourth, etc. Clearly we could do more slits by simply wrapping around more rotated isoceles triangles! This construction thus *scales*.

Also note that this means that the total angle in each apex must be δ, as the (blue) line A rotates through the same angle that the first E_0 phasor rotates through as drawn. This explains the entire "black" line drawing of the $N = 5$ isoceles triangles, which we can extend to $N = 6, 7...$ whatever positive integer[165].

The *resultant* phasor obtained by *adding* the 5 E_0 phasors is drawn in red as $E_{\rm tot}$. This is what we wish to obtain. To find it, we have to observe two things. First, looking at only the blue right triangle on the far left with hypoteneuse (red and blue dashed, as this contributes to two things in our reasoning) A and blue side $E_0/2$ opposite to the angle $\delta/2$ we see that:

$$\frac{E_0}{2} = A\sin(\delta/2) \quad \Rightarrow \quad A = \frac{E_0}{2\sin(\delta/2)} \tag{12.29}$$

Second, note that the total angle in the apex of the *big* isoceles triangle with side A and base $E_{\rm tot}$ is just $N\delta$ with (in this specific case drawn) $N = 5$. It's just the sum of the apex angles of the $N = 5$ identical isoceles triangles! Obviously (to belabor the obvious) if we had three, or four, or six slits, the total angle in the red *large/collective isoceles triangle* that contains $E_{\rm tot}$ would be $N = 3$ or $N = 4$, or $N = 6 \times \delta$.

Now it is easy! If we bisect this isoceles triangle to drop the (red, dashed) perpendicular from C to $E_{\rm tot}$, *half* of this angle is then $N\delta/2$ (for any value of N slits) and hence from the right triangle with red and blue dashed hypoteneuse A we see that:

$$E_{\rm tot} = 2 \times A\sin(N\delta/2) \quad \Rightarrow \quad E_{\rm tot} = \frac{\sin(N\delta/2)}{\sin(\delta/2)}E_0 \tag{12.30}$$

and we're (almost) done! All that remains is to form the intensity, time average away the harmonic waveform $\sin(kr - \omega t + \gamma)$ (so that we don't even case what γ turns out to be), and get as our final result the average intensity:

$$I_{\rm tot}(\theta) = \frac{1}{2}\epsilon_0 c E_{\rm tot}^2 = \frac{1}{2\mu_0 c}E_0^2\left(\frac{\sin(N\delta/2)}{\sin(\delta/2)}\right)^2$$

or:

$$\boxed{I_{\rm tot}(\theta) = \left(\frac{\sin(N\delta/2)}{\sin(\delta/2)}\right)^2 I_0} \tag{12.31}$$

As usual:

$$I_0 = \frac{1}{2}\epsilon_0 c E_0^2 \tag{12.32}$$

is the intensity of any *single* slit on the cylindrical screen at the angle θ in the Fraunhofer limit where $D \gg d$.

[165] At least, we can if we are space aliens with $6, 7...$ fingers or humans possessed of a good imagination...

This expression is general! It solves the N-narrow-slit interference problem for any value N as long as all N slits are within the lateral coherence length for the plane-wave light illuminating them and being transmitted to the screen. It even works for *two* slits because:

$$2\cos(\delta/2)\sin(\delta/2) = \sin(\delta) \quad \Rightarrow \quad \Big(2\cos(\delta/2)\Big)^2 = 4\cos^2(\delta/2) = \left(\frac{\sin(\delta)}{\sin(\delta/2)}\right)^2$$

or:

$$I(\theta) = 4I_0\cos^2\delta/2 = I_0\left(\frac{\sin(\delta)}{\sin(\delta/2)}\right)^2$$

Let's use this to plot the interference patter for $N = 2, 3, 4, 5$ slits in figure 12.7 for $d = 4\lambda$ This will help us inductively understand how the pattern heuristically *changes* as N is increased.

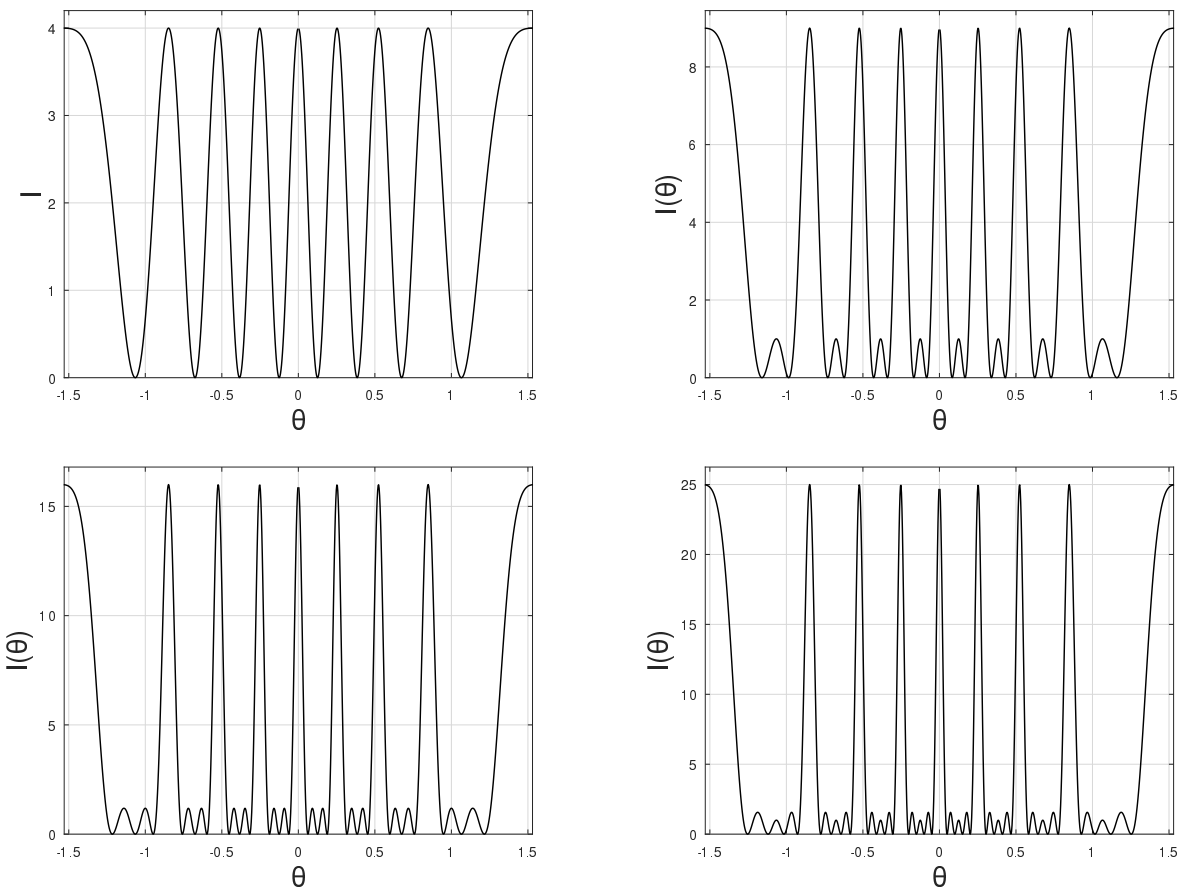


Figure 12.7: The interference patterns for $N = 2, 3, 4, 5$ narrow slits evaluated using the explicit formula derived above and plotted with $I_0 = 1$. The principle maxima are clearly of height $N^2 I_0$ and appear *at the same angles* for all four figures! The rest of the (secondary) maxima are "in between" the ordered minima and are "order of unity" compared to the N^2 height of the principle maxima.

From this figure we see a number of interesting things, things we'd like to understand and even be able to compute! For example, the principle maxima apparently ***occur at exactly the same angles*** independent of the number of slits, suggesting that their location depends only

on d, the slit separation, not N. However, as we add slits one new *minimum* and one more *secondary maximum* appear in between the principle maxima for each bump in N! Finally, if you look closely, the secondary maxima are not constant in height, although they seem so far to be "order unity", smaller than the height of the principle maximum by a factor somewhere between N and N^2, with those closest to the principle maxima larger than those in the middle.

In the next section, we'll find a *new, more intuitive use* for the N slit phasor diagrams that will work well enough for N in the finger range to allow one to draw a quite accurate graph of the intensity without evaluating a single function. Then we'll bite one more algebraic bullet and find *explicitly* where the principle maxima and minima occur using algebra instead of intuition and visualization, and obtain a nasty but useful transcendental formula from which the angles where the secondary maxima occur can be found.

12.3.1: Principle Maxima, Minima, and Secondary Maxima

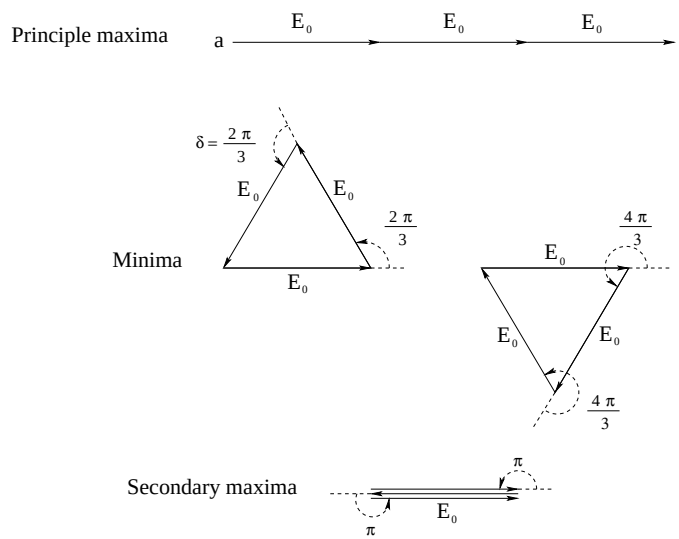


Figure 12.8: Phasor diagrams illustrating principle maxima, minima, and secondary maxima in the interference pattern. Note that we get minima when the three phasors *close* to get a three-sided polygon or 3-gon (a.k.a. an equilateral triangle in this case). In between the minima we get maxima, but the secondary maxima are much weaker than the principle maxima that occur when the light from all three slits arrives in phase at P.

As it turns out, it is quite easy to develop heuristic rules from which we can obtain the angles (in terms of δ) where the principle maxima and minima occur based on the phasors themselves and some refreshingly simple geometry. We can even use this approach to get an *approximate* idea of where the *secondary* maxima occur, although as we'll see finding exact solutions is beyond the reach of heuristics and even beyond the reach of traditional algebra – as a general rule, if they are ever needed they must be obtained using numerical methods to some precision.

Let's start by looking at a series of *specific* phasor diagrams for $N = 3$ slits that correspond to these extrema. Consider the first of the four phasor diagrams drawn in figure 12.8. Clearly, we get a principle maximum whenever the three phasors line up (for simplicity the figures

are again shown at a time that $\alpha = kr - \omega t = 0$) for a *total* field amplitude of $3E_0$. This obviously occurs when $\delta = 0$, but it can *also* correspond to $\delta = 2\pi, 4\pi, 6\pi...$ – rotating any field phasor through any integer multiple of 2π puts it back where it started. We conclude that this arrangement leads to *maxima* in intensity with $I_p = 9I_0$, which we've already been (correctly) referring to as the *principle maxima* of the interference pattern.

We can reduce this rule to an even more intuitive and hopefully (from the heuristic analysis of 2 slits above) familiar condition on $d\sin\theta$ as follows:

$$\delta_{\text{principle max}} = \frac{2\pi}{\lambda} d\sin(\theta) = 0, \pm 2\pi, \pm 4\pi... = \pm 2\pi\, m \qquad m = 0, 1, 2... \tag{12.33}$$

If we divide by 2π and multiply by λ, we see that this corresponds to:

$$d\sin(\theta) = \pm m\lambda \quad \Rightarrow \quad \boxed{\theta_m^{\text{principle max}} = \sin^{-1}\left(\frac{\pm m\lambda}{d}\right) \qquad m = 0, 1, 2...} \tag{12.34}$$

This is *exactly what we already observed* in the plots above! **The locations of the principle maxima of N slits are determined by the slit separation d, not by the number of slits N!** The two signs just mean that the pattern obtained is symmetric, with maxima at the same angles above and below the horizontal $\theta = 0$ line. We will (from now on) ignore this and just present positive m and find positive θ's, and remember that the intensity pattern is symmetric for negative θ.

There is an extremely simple and intuitive explanation for this result. When the path difference between any two adjacent slits contains an integer number of wavelengths, light from those two slits arrives at P in phase. But the same condition holds for *all* pairs of slits separated by the common distance d, so the light from *all* the slits arrives in phase! The field amplitude is thus just $3E_0$ for 3 slits, so $I_{\text{principle max}} = 9I_0$!

Now let's consider the minima, obtained from the triangles in the middle row of diagrams in figure 12.8. First, note that the intensity cannot be negative – I don't even know what a negative intensity would *mean* for light – the Poynting vector can have a sign relative to some coordinate frame, but the intensity is just the absolute power per unit area that flows past any given point in space. The smallest it can possibly be is zero.

For this problem it will *be* zero when the phasors for the field *add up* to zero! Our next job, then, is to figure out the geometries for which this occurs. In our $N = 3$ example, given three equal field strengths, the phasors will add up to zero when the phasors form a ***closed, three sided figure***, that is, a *unilateral triangle*[166]. The two triangles in the figure above thus represent the two phase angles that lead to *minima* with intensity *zero* in our graphs of $I_{\text{tot}}(\theta)$.

We observe that we close these triangles when:

$$\delta_{\text{min}} = \frac{2\pi}{3} \text{ or } \frac{4\pi}{3} \tag{12.35}$$

or these angles with *any integer multiple of* 2π *added (or subtracted).* If we multiply this out

[166] To begin to get ready for the next topic, you might want to think about a unilateral triangle as a *3-gon*, a polygon with three sides.

and turn it into a rule, it becomes:

$$\begin{aligned}
\delta_{\min} = kd\sin(\theta) &= \frac{2\pi}{3}, \frac{4\pi}{3}, \cancel{\frac{6\pi}{3}}, \frac{8\pi}{3}, \frac{10\pi}{3}, \cancel{\frac{12\pi}{3}}, \frac{14\pi}{3}, ... \\
\frac{2\pi}{\lambda} d\sin(\theta) &= \frac{2\pi}{3}, \frac{4\pi}{3}, \otimes, \frac{8\pi}{3}, \frac{10\pi}{3}, \otimes, \frac{14\pi}{3}, ... \\
d\sin(\theta) &= \frac{m\lambda}{3} = \frac{m\lambda}{N} \qquad m = \otimes, 1, 2, \otimes, 4, 5, \otimes, 7, 8...
\end{aligned} \tag{12.36}$$

We get minima when δ is an integer multiple of $2\pi/3$ (where 3, recall, is N, the *number of slits*), *except* that we have to *skip* the multiples of $2\pi/3$ that are also multiples of 2π because we already know that the multiples of 2π are ***principle maxima, not minima!*** I indicate this above by putting $\otimes$'d out *holes* in the m-sequence in the final result. We'll continue this practice in the next section.

Finally, consider the last phasor diagram, which (more or less) corresponds to a *secondary maximum*[167]. If we set:

$$\delta_{\text{secondary max}} = \pi, 3\pi, 5\pi... \tag{12.37}$$

then this phasor diagram results. Although at the moment there isn't any compelling reason to see why (there will be shortly) let's write this as:

$$\begin{aligned}
\delta_{\text{secondary max}} &= \pi, \cancel{2\pi}, 3\pi, \cancel{4\pi}, 5\pi... \\
\frac{2\pi}{\lambda} d\sin(\theta) &= \pi, \otimes, 3\pi, \otimes, 5\pi, ... \\
d\sin\theta &= \frac{m\lambda}{2} = \frac{m\lambda}{N-1} \qquad m = \otimes, 1, \otimes, 3, \otimes, 5...
\end{aligned} \tag{12.38}$$

which looks like it *might* be a path difference rule involving $m\lambda/(N-1)$ with the usual "skip all m that lead to a multiple of 2π because that's where principle maxima happen" exception.

The expressions for $\theta_m^{\min}$ and $\theta_m^{\text{secondary max}}$ are now found by a bit of algebra/arithmetic and taking an inverse sine. A typical problem for multiple slits would have you build a table of angles (or sines of angles) for the principle maxima, the minima, and the secondary maxima, and then to draw a "generic" graph of the intensity using this information.

Unfortunately, just looking at two and three slits isn't *quite* enough to infer a trustworthy rule, especially for the secondary maxima. Let's jump right on up to $N = 5$ slits, and use the phasor diagrams that work to deduce quite general rules for principle maxima, minima and sort-of secondary maxima for *arbitrary* N. Then we'll proceed to get all of these results exactly using calculus to locate the extrema of the intensity function itself.

Once again we'll illustrate the *methods* involved with $N = 5$. In figure 12.9 I've drawn phasors that illustrate the principle maxima that result when light from all five slits arrives at the screen perfectly in phase. I've also illustrated the total (zero) *minima* that occur when the phase angles δ are ones that *close the* $N = 5$*-sided regular polygon*. Finally, I've drawn a *crude* version of what a secondary maximum might look like if it occurred at a specific, simple, set of angles δ that close the $(N - 1 = 4$-sided polygon (leaving one phasor exactly uncancelled).

Note the following features, described in terms of the *general* rules that they represent:

[167] For $N = 3$ this correspondance is, in fact, exact, but not for higher N.

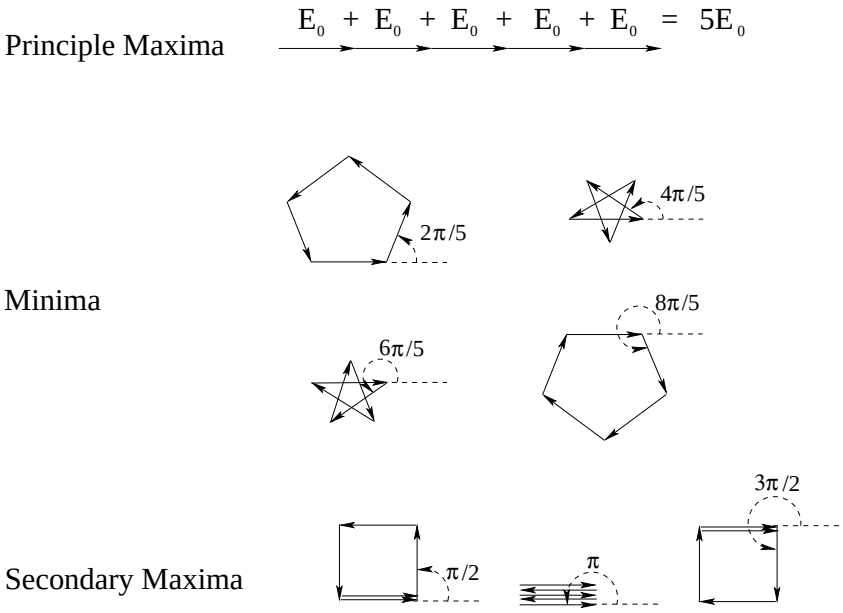


Figure 12.9: Phasor diagrams principle maxima, minima, and secondary maxima for five slits. The amplitude of the secondary maxima aren't *exactly* E_0 (or equal) and the angles aren't *exactly* at $\delta = 2\pi/(N-2)$ (for $N = 5$) but this is close enough for an excellent semi-quantitative graph of the intensities (and our heuristic understanding).

a) Principle maxima have field amplitude of NE_0 (for $N = 5$) when the field phasors "all line up". They do so whenever the phase angle δ is an integer multiple of 2π. Clearly this result (which held for $N = 2$ and 3 as well) is general. Thus for *all* N we find:

$$\delta_{\text{principle max}} = 2\pi m \qquad m = 0, \pm 1, \pm 2, \pm 3... \tag{12.39}$$

or:

$$\begin{aligned} \delta_{\text{principle max}} = kd\sin(\theta) &= 2\pi m \\ \frac{2\pi}{\lambda} d\sin(\theta) &= 2\pi m \\ d\sin(\theta) &= m\lambda \end{aligned} \tag{12.40}$$

Principle maxima occur when the light from *all* of the slits arrives at the point of observation *in phase*, which in turn happens when the path travelled by light from any two adjacent slits differs by an integer number of wavelengths. This makes perfect heuristic sense and corresponds to what we actually observed in explicit graphs of $I_{\text{tot}}(\theta)$ above.

Note well that the series doesn't continue indefinitely – the largest m that contributes is one where:

$$\theta_m^{\text{principle max}} = \sin^{-1}\left(\frac{m\lambda}{d}\right) \tag{12.41}$$

exists, so $m\lambda/d$ ***has to be less than or equal to 1***. This condition constrains all of the other series (below) as well, just as it did for 2 or 3 slits.

b) Minima occur when the N-gon formed by the amplitudes closes (forming pentagons or five pointed stars in the $N = 5$ case). The angles δ where these minima occur clearly

form the series:

$$\delta_{\min} = \frac{2\pi m}{5} \qquad m = \otimes, 1, 2, 3, 4, \otimes, 6, 7, 8, 9, \otimes, ... \tag{12.42}$$

where I've $\otimes$'d out the values $m = 0, 5, 10, ...$ We *have* to skip those in the series because e.g. $10\pi/5 = 2\pi$, and we already know that $\delta = 2\pi$ is a principle *maximum*. Clearly this generalizes to:

$$\delta_{\min} = \frac{2\pi m}{N} \qquad \text{for}$$
$$m = \otimes, 1, 2, ..., N-1, \otimes, N+1, N+2, ..., 2N-1, \otimes, 2N+1, ... \tag{12.43}$$

where we have to skip every Nth value of m.

Take a moment and verify that this rules works for $N = 2$ and $N = 3$ slits.

c) In between any pair of adjacent, isolated minima, a smooth function must have a maximum. We therefore expect that in between each adjacent pair of minima enumerated above, there must be a maximum. The principle maxima have already been enumerated, but there also exist a whole list of *secondary maxima*. These occur as the "chain" of E-field vectors twists around in between closed N-gons, and occur *close to* (but not exactly at) where the $(N-1)$-gon closes, leaving a single "dangling" E_0 at the end. If one evaluates the maxima more carefully (using calculus) one finds that they aren't exactly at the $(N-1)$-gon angles, and don't have the exact length E_0, but they are all *close to* these angles and lengths and we'll consider this to be "good enough" to help us draw a semi-quantitatively correct graph of the intensity.

This was illustrated in the 5-slit example above as:

$$\delta_{\text{secondary max}} = \frac{2\pi m}{4} = \frac{\pi m}{2} \qquad m = \otimes, 1, 2, 3, \otimes, 5, 6, 7, \otimes ... \tag{12.44}$$

where we note that we again have to skip the values of m that would lead to a δ that is an integer multiple of 2π, and generalizes to:

$$\delta_{\text{secondary max}} = \frac{2\pi m}{N-1} \qquad m = \otimes, 1, 2, ..., N-2, \otimes, N, N+1 ... \tag{12.45}$$

and so on which put the secondary maxima roughly "half way" in between the minima.

These rules are more than sufficient to allow us to draw by hand graphs of the expected intensity in *excellent semi-quantitative agreement* with the exact graphs for $N = 2, 3, 4, 5$ above, and of course we can extend it to $N = 6, 7...$ as long as we have the patience to proceed. We won't get the exact heights or angles of the secondary maxima quite right, but can we do better? Sure we can! We can use *calculus* to find the angles for the extrema of the intensity!

12.3.2: Finding the Maxima and Minima Exactly

We start with the expression for the intensity we derived above:

$$I = \left(\frac{\sin(N\delta/2)}{\sin(\delta/2)}\right)^2 I_0 = \left(\frac{\sin(Nu)}{\sin(u)}\right)^2 I_0 \qquad \text{where} \quad u = \delta/2 \tag{12.46}$$

We take the derivative of the intensity with respect to u (*much* easier than screwing around with $\delta/2 = kd\sin\theta/2$ at this point in the game) and set the result equal to zero to identify the maxima and minima:

$$\frac{dI}{du} = 2\left(\frac{\sin(Nu)}{\sin(u)}\right)\frac{d}{du}\left(\frac{\sin(Nu)}{\sin(u)}\right)I_0 = 0 \tag{12.47}$$

or (cancelling 2 and taking the remaining derivatives and factoring):

$$\sin(Nu) \times \left(N\frac{\cos(Nu)}{\sin^2(u)} - \frac{\sin(Nu)\cos(u)}{\sin^3(u)}\right) = 0 \tag{12.48}$$

This makes it clear that we get extrema whenever:

$$\sin(Nu = N\delta/2) = 0$$

(and $\sin u \neq 0$). This happens at the angles where:

$$\frac{N\pi d\sin\theta_m}{\lambda} = m\pi \quad \Rightarrow \quad d\sin\theta_m = \frac{m\lambda}{N} \quad \text{for} \quad m = 0, \pm 1, \pm 2$$

exactly as expected! Whenever m is an integer multiple of N, we get principle maxima (technically taking limits to correctly identify the central maximum where $u \to 0 \quad \Rightarrow \quad Nu \to 0$); all the rest we identify as the minima!

However, we *also* get extrema – the ***secondary maxima*** – when the stuff in the large parentheses equals zero! We can algebraically rearrange and rewrite this condition as:

$$N\tan(u) = \tan(Nu) \tag{12.49}$$

Solving for u in this case is hard! This is a *transcendental equation*[168] , and pretty much the only way to solve it for the specific values of u – which are not, alas, simple multiples of π – is numerically, although we can get a good heuristic picture of the solutions graphically.

Figure 12.10 illustrates the "graphical solution" – graphing both red $N\tan(u)$ and blue $\tan(Nu)$ for $N = 5$ on a single coordinate frame[169]. The two circles at $u = 0$ and $u = \pi$ are irrelevant – they correspond to ***principle maxima*** which we've already discovered. The three x's on the u axis correspond to three more points where the two functions are equal, where the red lines cross the blue lines:

$$u_1 \approx 0.91 \qquad u_2 = \pi/2 \text{ (exactly)} \qquad u_3 \approx 2.23 \tag{12.50}$$

The $u_2 = \pi/2$ is where both functions are infinite and hence "cross". These are the values of u that lead to:

$$2u_i = \delta_i = \frac{2\pi d}{\lambda}\sin\theta_i \tag{12.51}$$

corresponding to the secondary maxima! We can rearrange and evaluate (for, recall, $d = 4\lambda$):

$$\sin\theta_i = \frac{2u_i}{8\pi} \quad \Rightarrow \quad \theta_1 = 0.072 \qquad \theta_2 = 0.125 \qquad \theta_3 = 0.178 \tag{12.52}$$

[168] Wikipedia: http://www.wikipedia.org/wiki/Transcendental equation.

[169] Note that the vertical blue lines at the singularities of $\tan(Nu)$ are irrelevant as the drawing program doesn't know how to jump discontinuously from $+\infty$ to $-\infty$ the way the tangent function does.

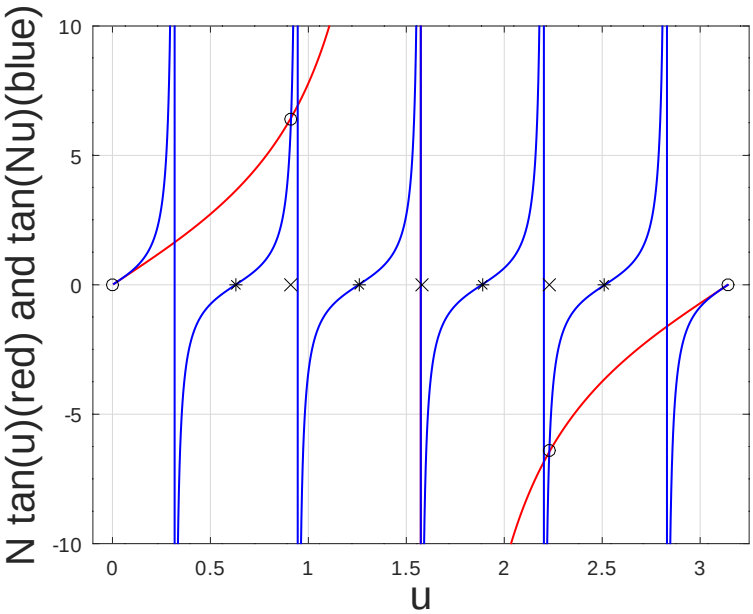


Figure 12.10: A single cycle of the transcendental relation derived above for $N = 5$, showing the locations of the values u where the intensity has extrema. The minima and principle maxima occur where the blue $\tan(Nu)$ curve is itself 0, with the (four) minima marked with $*$'s and the two principle maxima at the ends marked with o's . The remaining (three) points are *secondary maxima* and occur where the $N \tan u$ function crosses the $\tan(Nu)$ curve and neither one is zero, marked with x's on the horizontal axis.

(accurate to around 2 significant figures). One can also use e.g. Newton's Method[170] or finding the zeros of a function to find the secondary maxima to high precision. These numbers can be compared with the graph of $I_{\rm tot}(\theta)$ for $N = 5$ above, where you can see they are in excellent – indeed precise – agreement.

This also gives us insight into the heuristic rule for getting *close* to a secondary maximum given above. Note that the blue-red crossings all happen at least *close to* (for $d = 4\lambda$, $N-1 = 4$:

$$\sin\theta_i = \frac{i}{4(N-1)} = \sin\theta_i = \frac{i}{12} \quad i = \cancel{0}, 1, 2, 3, \cancel{4}...$$

or

$$\theta_1 \approx 0.063 \qquad \theta_2 \approx 0.125 \qquad \theta_3 \approx 0.189 \tag{12.53}$$

obtained from the heuristic rule suggested above. The are all pretty good estimates, and in one case (the one in the middle) the angle is more or less "accidentally" exact!

It's worth looking at the *exact* phasor diagram for at least one of the secondary maxima. In figure 12.11 you can examine the actual phasor diagram for the first secondary maximum past the first minimum. Note that the angle $\delta_1 = 1.82$ radians is just a bit larger than the heuristic angle $\delta_1 \approx \pi/2 = 1.57$ radians that is "halfway" between the first two minima. By winding a bit past it, it opens the structure just enough to lengthen the resultant.

I'm going to beg out of pursuing this any further in this (after all) *introductory* textbook. I'm pretty sure that you, dear reader, won't mind. However, this leaves us in excellent shape for

[170]Wikipedia: http://www.wikipedia.org/wiki/Newton's Method. f

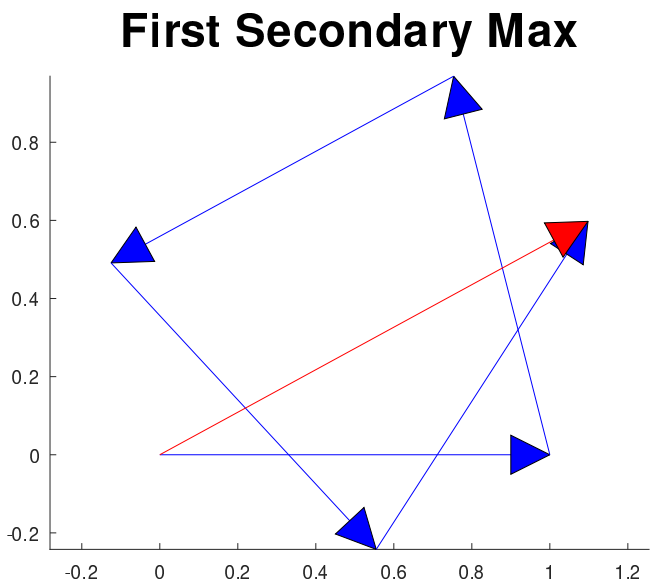


Figure 12.11: The phasor diagram for the first secondary maximum after the first minimum at $\delta_1 = 1.82$ radians. The red arrow represents the resultant from adding the five slit phasors and, you should note, is a bit *longer* than any single phasor but still order unity.

the next few sections, as we will have occasion to use these results in practical and important ways!

12.4: The Diffraction Grating – Rayleigh's Criterion for Resolution

You might wonder why we are spending so much time looking at interference through multiple slits, when we hardly ever run into problems involving interference through just *two* slits while shopping at the mall. There are two simple reasons. The first is that interference from many closely spaced slits is the basis for the *diffraction grating*, which in turn is the basis for modern spectrographs. Spectrographs are optical instruments used to identify e.g. atoms and molecules from their "signature" optical spectra, and are the basis for much of what we know of the Universe. For example, we know that the physical laws governing very distant stars very far away (and hence being observed today in their distant past due to the speed of light delay) are pretty much *identical to the laws we observe today* because their *spectra are the same* within a physically understandable red shift!

This may sound silly, but this is an *enormously* important result. If things like the gravitational constant G, the electric permittivity ϵ_0, the magnetic permeability μ_0, the speed of light c – constants of nature, as it were – weren't *constant* over time frames of billions of years, it would radically alter our perceptions and understanding of the Universe we find ourselves apparently living in. Instead we find that no matter how far away or how far back in time we look, the spectra of atoms in stars are pretty much the same, something that actually tests *many* of the constants of nature all at once. The physics governing those stars way back in the distant past and far away, then, seems to be pretty much the same as the physics we learn and use today.

Of course spectrographs are also useful throughout science and technology in a strictly mundane way. We have *many* occasions to wish to identify a material, and if we heat almost anything until it glows and then examine its light with a spectrograph, we can instantly identify at least all of the elements in the sample and their relative abundance, if not the molecules made up of those elements. Chemistry, engineering, and a variety of physical sciences use this capability every day, using machines that have more or less automated the process. It does seem wise for us to learn at least in general how this works, and what limits the resolution and accuracy of the process.

The second place understanding the interference of "many" slits will aid us is in bootstrapping our understanding of ***single slit*** diffraction, where light passing through a single *wide* slit interference "with itself". There, a mix of Huygens principle and our knowledge of N-slit interference will let us quickly come to understand how a single "wide" slit can produce a characteristic intensity pattern when cast on a distant screen. In the next two sections we will therefore apply the concepts we have just worked out for $2, 3, ..., N$ slits, beginning with the *diffraction grating*.

A diffraction grating is basically an opaque material with many transparent narrow slits inscribed through the opacity, each separated from its neighbor by a distance d – it's just the N-slit problem we just examined but for *large* N*!* We will imagine this grating to be normally illuminated by polychromatic light (with many frequencies/wavelengths) in such a way that N of them produce outgoing waves that recombine coherently at the screen, where in application the screen is indeed wrapped around in a cylinder at a distance D that is large compared to $d > \lambda$ (for any λ in the visible band).

As we saw in the previous section, the angles at which the primary maxima occur are determined only by the distanced d such that:

$$\theta_m^{\max} = \sin^{-1}\left(\frac{m\lambda}{d}\right) \tag{12.54}$$

independent of N – indeed, they are at the same angles for 2 slits as they are for 2000.

What changes as we increase the number of slits is the location of the *minima* and the secondary maxima in between. Consider the two minima that "bracket" each primary maximum. Again borrowing results from the previous section, we can see that they should occur at:

$$\theta_m^{\min} = \sin^{-1}\left(\frac{n\lambda}{Nd}\right) \tag{12.55}$$

for the particular values:

$$\begin{aligned} n_1 &= N \pm 1 \\ n_2 &= 2N \pm 1 \\ &... \\ n_m &= mN \pm 1 \\ &... \end{aligned} \tag{12.56}$$

where the index n_m can (as you can see) take on two values for each m, one for the minimum immediately before, the other for the minimum immediately after the mth principle maximum:

$$n_m = N * m - 1, N * m + 1 \qquad m = 1, 2, 3... \tag{12.57}$$

We now no longer need n_m. We can directly write these angles in terms of m alone as (factoring):

$$\theta_m^{\min} = \sin^{-1}\left(\frac{m\lambda}{d} \pm \frac{\lambda}{Nd}\right) \tag{12.58}$$

for each pair of values that bracket the mth maximum.

We now make the small angle approximation for both the maxima and the minima. This may well not be justified – many diffraction gratings will produce even the first principle maximum at a relatively large angle – but it suffices for us to understand what they do and the idea of "resolving power", and we can always take the actual inverse sines if needed for a particular actual grating. With this approximation, we get:

$$\theta_m^{\max} \approx \left(\frac{m\lambda}{d}\right) \tag{12.59}$$

and:

$$\theta_m^{\min} \approx \left(\frac{m\lambda}{d} \pm \frac{\lambda}{Nd}\right) = \theta_m^{\max} \pm \frac{\lambda}{Nd} \tag{12.60}$$

This is just what we need to understand what a diffraction grating does: it makes an absolutely perfect *spectrometer*, allowing us to cleanly resolve the spectral lines emitted by hot glowing atoms and molecules and thereby both identify them and make many inferences concerning their structure!

To see how this works, imagine that there are two "spectral lines" λ_1 and λ_2 being emitted by a given atom (such as the two emitted by the Sodium atom, with D1 at $\lambda_1 = 589.592$ nm and D2 at $\lambda_2 = 588.995$ nm, see homework). The *first* principle max for λ_1 occurs at the (presumed small) angle:

$$\theta_1(\lambda_1) = \frac{\lambda_1}{d} \tag{12.61}$$

while that for λ_2 occurs at:

$$\theta_1(\lambda_2) = \frac{\lambda_2}{d} \tag{12.62}$$

These two lines are *separated* in angle by:

$$\Delta\theta_{12} = |\theta_1 - \theta_2| = \frac{\lambda_1 - \lambda_2}{d} \tag{12.63}$$

The lines projected on the screen, however, are not infinitely sharp (even if the sodium wavelengths themselves are)! The *widths* of the first principle maxima at λ_1 or λ_2 are:

$$\Delta\theta \approx \frac{2\lambda_1}{Nd} \approx \frac{2\lambda_2}{Nd} \tag{12.64}$$

If the two maxima are too close together, their lines will *overlap* and we won't be able to tell that there are two lines there at all! On the other hand, if they are far enough apart, the lines won't overlap at all (except out in the irrelevant morass of secondary maxima and higher order minima) and we'll be able to easily see two lines. We need a *criterion* for the minimal resolution of two spectral lines (or anything else) cast as an "image" onto a screen, or a piece of film, or the retina. Enter *Rayleigh's Criterion for Resolution*.

12.4.1: Rayleigh's Criterion for Resolution

Lord Rayleigh was yet another eponymous physicist who studied the wave properties of "rays" and things such as the resolving power of spectral gratings or optical instruments. We have encountered him before in the context of "Rayleigh scattering", the original blue-sky theory. He established a very simple criterion for when two spectral lines from a diffraction grating or diffraction maxima from e.g. circular apertures are marginally resolved. It is this:

> Two lines are said to be *marginally resolved* if the principle *maximum* for one line is outside of the *first minimum* of the other.

That's it! Nothing to it. It is really slightly more general than this, however. We will also use it below to determine whether two point-like *images*, when focussed on a screen through a circular aperture, are marginally resolved, where instead of "lines" we simply talk about the diffraction maxima of the dots, but the idea is exactly the same. For us to be able to determine that there are two instead of one, they cannot overlap, and "overlap" is defined to be the maximum of each further away than the first minimum of the other.

12.4.2: Resolving Power

With that criterion in hand, we can talk about and derive the *resolving power* of a grating and see how we can determine whether or not any given grating will be able to resolve any given pair of closely spaced lines.

In order for our grating to resolve two lines the angular separation of their maxima has to be larger than the angle of the first minimum of each maximum. That is:

$$\theta_m(\lambda_2) = \frac{m\lambda_2}{d} > \frac{m\lambda_1}{d} + \frac{\lambda_1}{Nd} = \theta_m^{\min}(\lambda_1) \tag{12.65}$$

or

$$\Delta\lambda_{21} = \frac{m(\lambda_2 - \lambda_1)}{d} > \frac{\lambda_1}{Nd} \tag{12.66}$$

We can rearrange this, noting its symmetry under exchange of 1 and 2 and defining $\lambda \approx \lambda_1 \approx \lambda_2$ (the whole point is that they are very close together, right?) to define the *resolving power* of the grating:

$$R = mN = \frac{\lambda}{\Delta\lambda} \tag{12.67}$$

Note well that $R = \lambda/\Delta\lambda$ is a measure of the relative resolution of the grating at any wavelength λ. $R = mN$ tells you what this resolving power is, given the order of the maximum you are observing and the number of slits that are coherently illuminated by the beam which contribute to it. As N goes up, the first minima squeeze ever more tightly around the principle maxima and the resolving power improves. However, as m increases *all* of the angles increase, as well as all of the separations of the angles. Since the width of the principle maxima does *not* vary with m, higher order maxima have better resolution, all things being equal. If we want to know if we can resolve two lines with separation $\Delta\lambda$ (both very near λ), we can merely evaluate:

$$\Delta\lambda_{\min} = \frac{\lambda}{mN} \tag{12.68}$$

for the order considered and if the two lines are separated by more than this spread, they will be resolved.

There are other places in our daily lives where "diffraction gratings" can be observed. CD or DVD ROMs, for example, consist of many "tracks" carved into a shiny reflective plater and pitted by means of a laser to encode information. The reflective grooves behave just like multiple slits and split white light up into a veritable rainbow of colors when the reflective grooved surface is viewed at various angles. There is no *real* color to the shiny disk; all of the color arises from multiple slit interferences.

This same process works *backwards*, as well. A radio telescope is made out of a regular array of antennae spread out in a two dimensional lattice. If we imagine all of the antennae radiating coherently at the save frequency and wavelength, we expect the waves they emit to only constructively interfere and hence radiate most of their energy along certain directions. If we reverse this, however, by adjusting the phase of the signals *picked up* by the antennae and combining them into one phase delayed superposition signal, we can arrange it so that they only coherently *receive* from certain directions in the sky. In fact, by appropriately sweeping the phase delays, we can sweek the telescope across the sky and make a highly directional map of all of the radio signals emitted by the sun, by stars, even by remote galaxies. We even expect resolution to improve as we increase the number of antenna, in a way that should now be intuitively familiar.

Now, let us think about multiple slits and Huygens' Principle. Huygens' Principle states that all of the points on a wavefront behave like coherent radiators, which sounds a *lot* like what multiple slits that sample just some of those radiators do. The difference is that with a wavefront, the number of coherent radiators has to go to infinity at the same time that the distance between radiators has to go to zero at the same time the amplitude emitted by each radiator (which we've been treating as a given *constant* for the many slit problems) has to also go to zero, but in such a way that the total energy emerging from a piece of the wavefront is conserved!

Handling all of this correctly lets us understand *diffraction*, the interference of a wave that e.g. passes through a *single* slit with *itself*. Understanding diffraction is absolutely essential to the understanding of the diffraction/wave based limitations of optical instruments such as microscopes and telescopes. We begin by completely analyzing and solving for the diffraction intensity produced by light passing through a *single* slit of width $a > \lambda$, in the usual Fraunhofer approximation.

12.5: Diffraction

We have seen how coherent, monochromatic light passed through multiple slits, when it recombines after traversing different path lengths, interferes – sometimes creates a wave with an amplitude greater than that produced by a single slit, sometimes cancelling altogether – and that this creates a modulation of the intensity observed on a distant screen, basically transforming it into a pattern of light and dark bars (or something more complex if we have sources more complicated than "slits").

We have *also* seen that Huygens' Principle tells us that every point on a wavefront of an

advancing wave behaves like a "source" for the future time evolution of the wavefront. This suggest that we don't *need* multiple slits in order to see a wave interfere – all we need is *one* slit, but one that is wide enough that it contains "many" Huygens radiators in the wavefronts that are incident upon it!

Calling this interference would be very confusing – one slit? two? ten? – so we introduce a new term to describe "interference" of a wave with itself, or the interference patterns produced by *very* large numbers of slits/sources, so many that they form a near continuum. We call this kind of phenomena *diffraction*, and speak of the *diffraction* of a wave through a single slit, or the diffraction of a wave around an obstacle, or the diffraction patterns produced on a screen or piece of film by light that passes through one or more slits that are wide enough that the light that goes through them can interfere with itself.

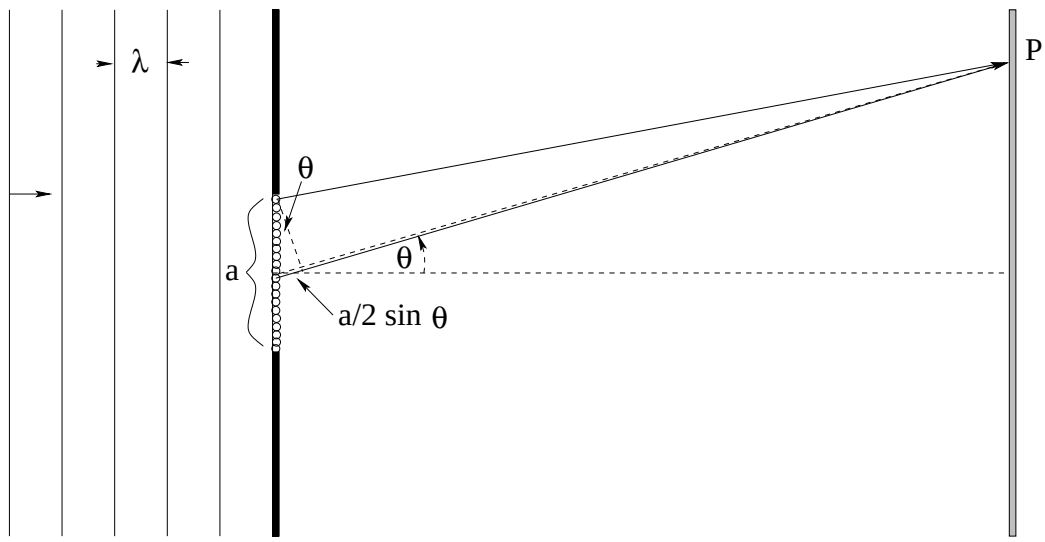


Figure 12.12: The geometry of single slit diffraction. Waves of some wavelength λ pass through a slit of width a, where a is typically somewhat larger than λ (to get an "interesting" diffraction pattern) and fall upon a screen under Fraunhofer conditions, where the screen is distant compared to a and λ and roughly equidistant from the center of the slit

The geometry of diffraction is straightforward and is represented in figure 12.12. Note its similarity to N slits – all of the N little round circles in the slit a represent Huygens radiators on the wavefront there.

As before, we'll assume that we have Fraunhofer conditions, so that the screen is far (compared to a and λ) from the slits, and we'll either ignore any radial variation in the field strength with distance or imagine that the screen bends in a half cylinder around the center of the slit. Note that we don't have to do this – we *could* work all of this out (and in later courses physics majors very likely will) but doing so doesn't help you understand the basic idea of diffraction itself so we won't bother[171].

Locating maxima and minima – especially maxima – will prove more difficult for a single slit (of width a) than it did for two or more very thin slits! Before we tackle actually solving for

[171] We'll also (as we've been doing) more or less ignore the vertical dimension of the slit (the one perpendicular to the paper) even though that is itself a "slit" and hardly seems to be as negligible as we've been making it out to be...

the intensity in a formally justifiable way, let's point out a couple of heuristic features that will – for the most part – suffice to help us understand at least the gross features of the diffraction pattern that results.

The first of these is the central maximum. At $\theta = 0$, all the radiators in the slit are basically equidistant from P and hence all of the coherent wavelets they spawn arrive in phase in the middle. We use this middle point of complete constructive interference of all of the Huygens radiators to define the *peak* amplitude and (time average) intensity of the light in the diffraction pattern, E_0 and $I_0 = 1/(2\mu_0 C)E_0^2$ respectively.

The second are the locations of the *diffraction minima* – angles at which the total amplitude and intensity are *zero*. We can find these using the following not-too-difficult mini-argument.

12.6: Diffraction Minima, Heuristic Rule

Consider the two waves emerging from the two Huygens radiators portrayed above in figure 12.12 and proceeding to the point P. As shown, the wave from the lower slit arrives having travelled a longer path, with a path difference of $\Delta r = \frac{a}{2}\sin(\theta)$.

We now apply the simple heuristic concept that served us well when we were trying to understand the two-slit minimum. If this path difference contains exactly $\lambda/2$ (one half of a wavelength) then the waves from these two particular radiators will *cancel* at P.

Now consider the second radiator down from the top. It also has a path difference of $\frac{a}{2}\sin(\theta)$ compared to the radiator second down from the middle and these two cancel. The third down from the top cancels the third down from the middle. In fact, *every* Huygens radiator in the top half of the slit cancels the corresponding radiator $a/2$ beneath it in the lower half of the slit. The field amplitude and intensity at P are *zero* (which is as low as one can get), making

$$\begin{aligned} \frac{a}{2}\sin(\theta) &= \frac{\lambda}{2}, \text{ or} \\ a\sin(\theta) &= \lambda \end{aligned} \tag{12.69}$$

a condition for a *diffraction minimum*.

Now imagine dividing the strip into fourths, as portrayed in figure 12.13. As you can see, if the path difference between the radiator at the top (0) and the radiator at $a/4$ contains $\lambda/2$ (a half a wavelength) they cancel, *and so does the wave from the radiator at* $a/2$ *cancel the wave from the radiator at* $3a/4$*!* Every point in the first quarter cancels a point from the second quarter and at the same time the corresponding points in the third and fourth quarter cancel. Again, no field amplitude arrives at P – this is a minimum with zero intensity. Multiplying out we get a second condition for a minimum:

$$a\sin(\theta) = 2\lambda \tag{12.70}$$

If we consider dividing the strip up into sixths, the condition $\frac{a}{6}\sin(\theta) = \lambda/2$ and the exact same argument shows that $a\sin(\theta) = 3\lambda$ is a minimum. If we divide it into eights we get $a\sin(\theta) = 4\lambda$. Clearly we can continue indefinitely; the general rule for a minimum is:

$$a\sin(\theta) = m\lambda \qquad m = \otimes, 1, 2, 3, ... \tag{12.71}$$

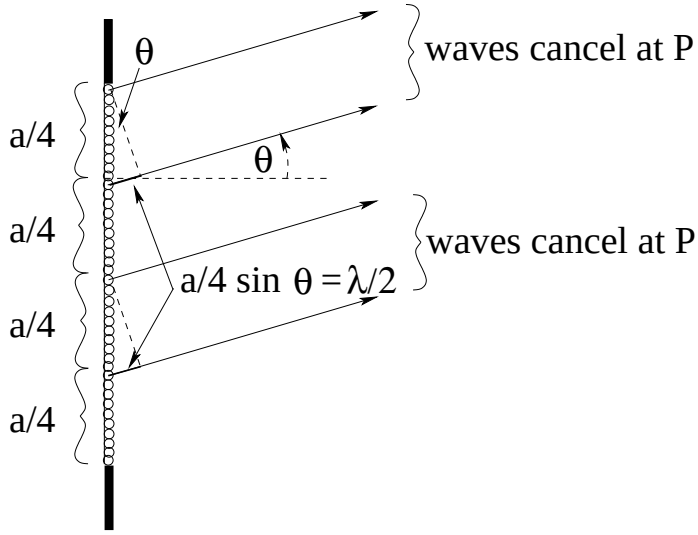


Figure 12.13: The slit, with the Huygens radiators divided into four equal segments. Light from the two pairs indicated cancels at P when the path difference $\frac{a}{4}\sin(\theta)$ contains a half of a wavelength, for all of the pairs that make up the slit.

where I've used $\otimes$ again to indicate that $m = 0$ is the principle *maximum* at the center, not a *minimum* and so must be skipped.

Finally, we know that diffraction will be symmetric, so that we have minima at all of the negative angles $a\sin(\theta) = -m\lambda$ but as before we'll manage this by hand to keep the equation simple.

Alas, no such simple argument can be made in order to find the angles of the diffraction *maxima* (except for the central principle maximum, already considered). We know there must *be* maxima in between each of the minima above but we expect from our discussion of N-slit interference that they won't occur at any "simple" values of the phase angle ϕ any more than they did at simple values of δ. We therefore abandon heuristics at this point and proceed to solve for the *exact* diffraction intensity as a function of phase angle ϕ (and hence θ, via the usual kind of inverse sines).

12.7: Exact Solution to Diffraction by a Single Slit

In figure 12.14 you can see a single slit with N radiators neatly drawn out. I chose $N = 7$ because it is enough to "cover" the slit without being so many that you can't see what is going on. In the end, of course, we will let $N \to \infty$ so that we *really* cover the slit with a continuum of radiators[172] so no particular choice for N much matters.

We have to be able to "scale" the field result itself. After all, the light we shine on the slit could be very intense or it could be weak. The slit could be large (letting a lot of light through) or it could be very small (not letting a lot of light through). We need a single parameter that indicates how strong the E-field is on the screen, or equivalently, how intense. We choose to

[172]... or, if this were a course in *optics* being given to majors or folks with mad math skills, we'd just write an *integral* for the field at an arbitrary P and not bother with all of this dividing up and summing...

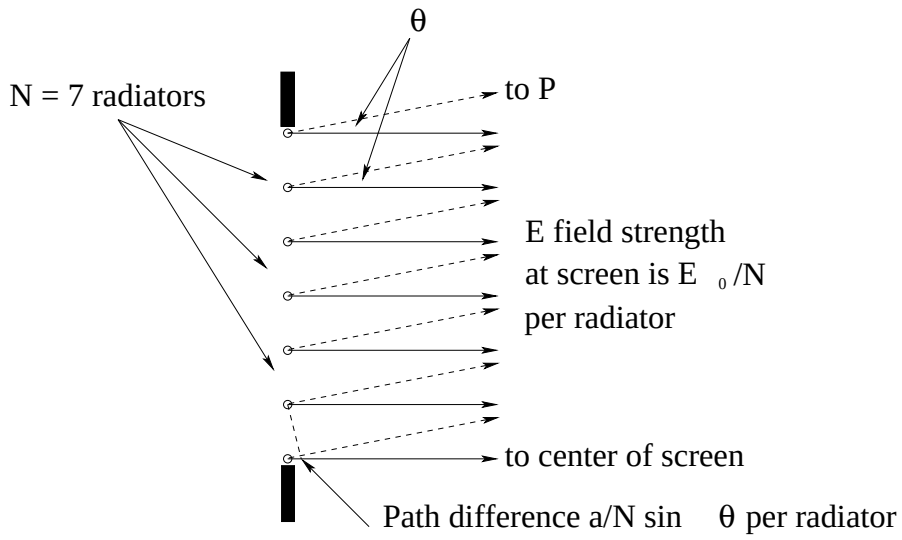


Figure 12.14: If we split the slit up into N radiators, the field amplitude at the maximum in the center of the screen from *each* radiator is E_0/N, where E_0 is the maximum amplitude from the entire slit there. When we consider the waves emerging at an angle θ directed towards point P, each radiator travels an additional distance of $\Delta r = \frac{a}{N}\sin(\theta)$ compared to the radiator immediately above it. Both of these relations scale with N, and hence will be useful when we try to let $N \to \infty$ and fill in the entire slit with radiators.

set E_0 to the value of the E-field that makes it through the slit to the screen in the *center of the principle maximum* at $\theta = 0$. With this interpretation, it is *exactly* like what we did for the interference of N "narrow" slits above. Indeed, at the end of this topic we can go back and *a posteriori* formally justify our narrow slit results, and define precisely just what "narrow" means!

If we split the slit up into N radiators, each with the same path length to the center of the screen (in the Fraunhofer limit, recall), then from symmetry and superposition run backwards each radiator must produce an individual E-field on the screen with strength E_0/N. That way, no matter what N is, the superposition of the fields at the center will remain equal to E_0, the measured/known/observed/assumed E-field there. As N gets large, this field amplitude (per radiator) will get very small (but nonzero) but the larger number of radiators will precisely compensate.

Next, let's think about path differences and phase differences. Recall that $a\sin(theta)$ is the total path difference to the point P between the wave from the (radiator at the) very top of the slit and the wave from the (radiator at the) very bottom of the slit. In the figure above, the top and bottom radiators aren't, of course, precisely "at" the top and bottom of the slits, but as we increase the number of radiators they will get closer and closer, and any error we make in assuming that they are there already for a finite N will go away.

We therefore can split $a\sin(\theta)$ up into N pieces, and make the path difference between adjacent radiators $\frac{a}{N}\sin(\theta)$. A very astute student might observe that for the 7 slits above, it really should be $\frac{a}{6}\sin(\theta)$ (or rather, that our general rule should be $\frac{a}{N-1}\sin(\theta)$ because the top radiator is at "zero") but in the limit $N \to \infty$ we will make an error of order $1/N$ using the first relation[173] so we'll just ignore it and use the first (easier) relation.

[173] As you can easily see by doing the binomial expansion of $a/(N-1) = (a/N)(1-1/N)^{-1}$, right...?

Let's turn this path difference between waves from adjacent radiators into a phase difference between adjacent radiators (by multiplying it by k, as always). Recall that we defined $\phi = ka\sin(\theta)$, so the phase difference between adjacent slits is just $\Delta\phi = \phi/N$. This phase difference *accumulates* as we count down the radiators from the top – the first slit down has a phase difference of ϕ/N, the second has a phase difference of $2\phi/N$, the third $3\phi/N$ and so on.

The wave we have to sum – using our ever-so-useful phasors, of course – is then (for $N = 7$):

$$\begin{aligned} E_{\rm tot} &= \frac{E_0}{N}\sin(kr-\omega t) + \frac{E_0}{N}\sin(kr-\omega t+\phi/N) \\ &+\frac{E_0}{N}\sin(kr-\omega t+2\phi/N) + \frac{E_0}{N}\sin(kr-\omega t+3\phi/N) \\ &+\frac{E_0}{N}\sin(kr-\omega t+4\phi/N) + \frac{E_0}{N}\sin(kr-\omega t+5\phi/N) \\ &+\frac{E_0}{N}\sin(kr-\omega t+6\phi/N) \end{aligned} \tag{12.72}$$

This is looking really tedious, and we're only at $N = 7$. However, if we *draw the phasor diagram for this sum*, it isn't so bad:

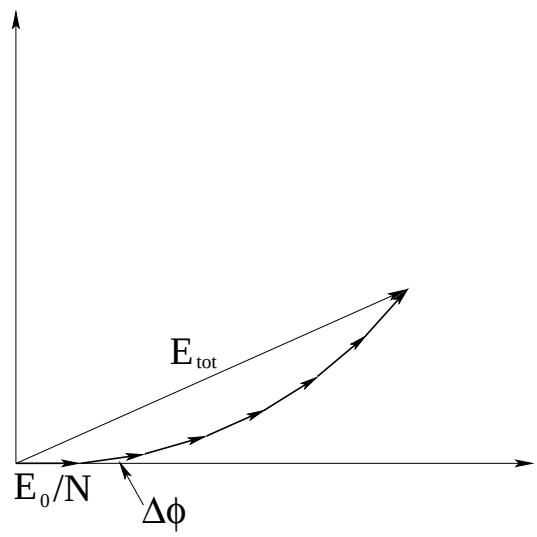


Figure 12.15: The phasor diagram for $N = 7$ Huygens radiators distributed across a. The amplitude of each radiator is E_0/N, and the phase $\Delta\phi = \phi/N$ accumulates.

The diagram in figure 12.15 (which we might have drawn for a 7-slit interference pattern!) shows us that as long as $\Delta\phi$ is *small*, the phasors gently arc up into what looks almost like a smooth curve even for only $N = 7$. In a seven *slit* problem however, as we increase θ then δ between two slits gets bigger and soon isn't small at all – we expect to get things like seven-pointed stars and so on that don't at all look like a smooth curve.

In this case of a *single* slit, however, as we make ϕ large, we can make $\Delta\phi$ *as small as we like* by increasing N! In fact, we can make it *infinitesimally* small, accumulating $d\phi$ as we go around a *smooth* curve. We won't actually do the following sums algebraically (so don't be intimidated by the notation) but we can in fact write the total field at the point P at the angle θ

in the Frauhofer approximation as[174]:

$$E_{\rm tot} = \lim_{N\to\infty} \sum_{i=0}^{N} \frac{E_0}{N} \sin(kr - \omega t + i\phi/N) \tag{12.73}$$

This sort of sum, accumulating infinitesimal chunks of E at infinitesimally different phase angles, is begging to be turned into an integral[175], but we will stop here and turn back to our user-friendly phasors. In this limit, the line of E_0/N-length phasors will form a *smooth arc* with a fixed length of E_0 . The total angle accumulated between the beginning of the arc and the end will be ϕ, the total phase difference between the top and bottom of the slits. Our "discrete" phasor diagram for 7 slits above will become the continuous phasor diagram illustrated in figure 12.16.

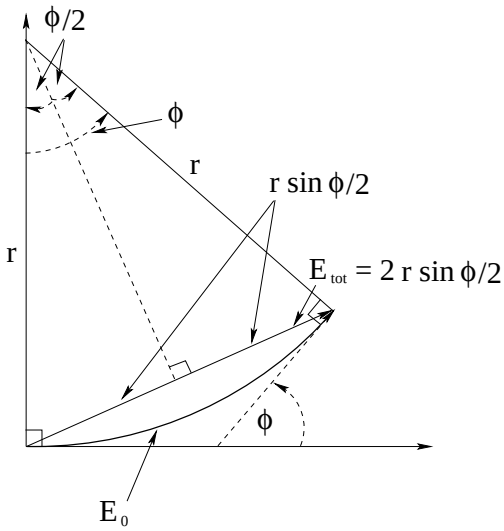


Figure 12.16: The phasor diagram for $N \to \infty$ Huygens radiators distributed across a. The "phasor snake" or "E_0-coil" bends smoothly around into a circular arc of length E_0, where we need to determine the length of the secant that cuts across, $E_{\rm tot}$.

I'll refer to the actual arc whose total length is E_0 as the "E_0-coil" as this is a convenient metaphor for the way the phasor diagram behaves as we increase ϕ as illustrated below, although "phasor snake" (one that eats its own tail like Ouroboros!) is almost as attractive and apropos.

Almost all of our work has been done for us in this diagram! Let's go over its features and results so that you understand them as we derive our final result. Note that the length of the arc is E_0 (we are just "bending it around", but all the superposition of all of the *amplitudes* of the infinitesimal phasor chunks still has to add up to E_0). The total phase difference between (a tangent to) the beginning of the arc and (a tangent to) the end of the arc is just ϕ, as illustrated with the lower ϕ angle. This *same* angle ϕ is the angle subtended by the circular arc as illustrated at the top – you can "see" by noting that the two r radii are perpendicular to the arc at both ends, so as we swing out the second r the angle accumulated by the tangent at

[174]Note that we are still ignoring that extra $\mathcal{O}(\mathcal{N})$ term on the end as there are $N+1$ terms in the sum.

[175]Ideally a complex exponential integral. Who actually *likes* to integrate sines and cosines and remember all of those silly sign change? $\int e^u du = e^u$, all we ever really need to know...

the bottom has to match the angle accumulated between the radii. From this we see that the arc length E_0 can be related to r by:

$$E_0 = r\phi \tag{12.74}$$

If we drop a perpendicular bisector (dashed line) from the center of the circular arc to the total field phasor $E_{\rm tot}$, we make two simple right triangles with vertex angle $\phi/2$. The opposite side of each of them has length $r\sin(\phi/2)$ so that:

$$E_{\rm tot} = 2r\sin(\phi/2) \tag{12.75}$$

We substitute $r = E_0/\phi$ into this (eliminating r in favor of E_0) to get:

$$E_{\rm tot} = \frac{2E_0\sin(\phi/2)}{\phi} = E_0\left(\frac{\sin(\phi/2)}{\phi/2}\right) \tag{12.76}$$

Finally, we go through the usual ritual to convert the field amplitudes to intensities:

$$I_0 = \frac{1}{2\mu_0 c}E_0^2 \tag{12.77}$$

so that:

$$I_{\rm tot} = \frac{1}{2\mu_0 c}E_{\rm tot}^2 = \frac{1}{2\mu_0 c}E_0^2\left(\frac{\sin(\phi/2)}{\phi/2}\right)^2 \tag{12.78}$$

or

$$I_{\rm tot}(\theta) = I_0\left(\frac{\sin(\phi/2)}{\phi/2}\right)^2. \tag{12.79}$$

This is what we have been trying to get – an exact formula for the intensity of the diffraction pattern as a function of θ (yes, it is actually given as a function of ϕ but recall that $\phi = ka\sin(\theta)$ so we also know it as a function of θ, at the expense of a little extra (and tedious, admittedly) arithmetic. But arithmetic isn't tedious to humans any more as long as an equation can be programmed into a computer, and this one is easy to code.

We'd like to find all of the maxima and minima of the intensity. From calculus, we know that we will get all maxima and minima in intensity at the values of $u = \phi/2$ for which:

$$\frac{dI(u)}{du} = \frac{d}{du}\frac{\sin^2(u)}{u^2}I_0 = 2\frac{\sin(u)}{u}\left(\frac{d}{du}\frac{\sin(u)}{u}\right)I_0 = 0 \tag{12.80}$$

We can now cancel out the 2 and I_0 and take the derivative in the parentheses:

$$\sin(u)\left(\frac{\cos(u)}{u^2} - \frac{\sin(u)}{u^3}\right) = 0 \tag{12.81}$$

At a glance, this equation has all of the right features. At $u = 0$, $\sin(u) = 0$. By inspection, we get an intensity of the *central maximum* I_0 here, where $u = \phi = \theta = 0$[176]. At all the other places where $\sin(u) = \sin(\phi/2) = 0$, we get a minimum. This occurs when:

$$\frac{\phi}{2} = \frac{\pi a}{\lambda}\sin(\theta) = \pi, 2\pi, 3\pi... \tag{12.82}$$

[176]We avoid the problem of "division by zero" calculus-fashion by taking the *limit*

$$\lim_{x\to 0}\frac{sin(x)}{x} = \frac{x - x^3/3! + x^5/5! - ...}{x} = 1 - x^2/3! + x^4/5! - ... = 1$$

or when:

$$a\sin(\theta) = m\lambda \qquad m = \emptyset, 1, 2, 3, ... \tag{12.83}$$

as before, so our heuristic rule is precisely derived.

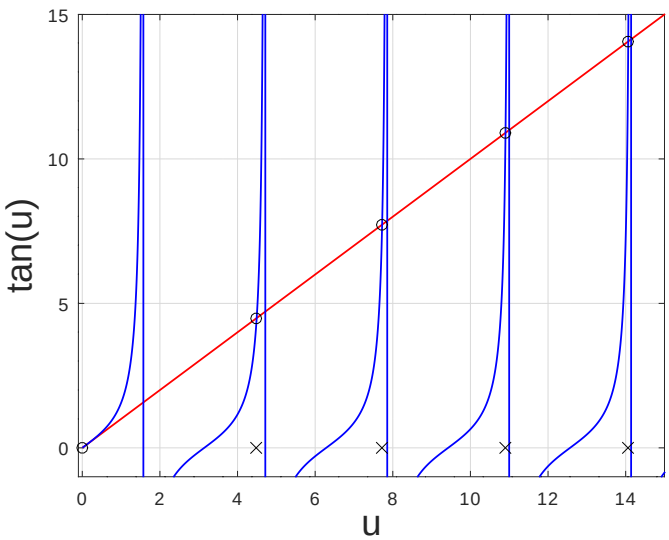


Figure 12.17: The graph from which the zeros corresponding to secondary maxima can be understood and extracted – approximately. One can obtain them more precisely using a computer algorithm to find the zeros.

The other zeros – the ***secondary maxima*** are obtained from:

$$\frac{\cos(u)}{u^2} - \frac{\sin(u)}{u^3} = 0 \quad \Rightarrow \quad u = \tan(u) \tag{12.84}$$

This is a transcendental equation[177] much like the one we graphed in the N-slit interference section. If one plots $u = \phi/2$ and $\tan(u) = \tan(\phi/2)$ simultaneously on a single set of axes, the intersections of the two lines are the relevant zeros. This is plotted in figure 12.17.

As one can see (once one does this) the maxima occur at angles *close to* but *just before* the condition(s):

$$\phi/2 = 0 \text{ (exact, and principle maximum)}, 3\pi/2, 5\pi/2, 7\pi/2... \tag{12.85}$$

That one would expect to be "halfway" between the minima, along the same lines as the results obtained for the N-slit interference problem. Note well the skipping of $\pi/2$, as the first solution is a *maximum*, not a minimum! The first four exact angles of secondary maxima (to three significant digits) are compared to these values numerically in the following table as extracted from the graph:

As always, one has to actually solve for θ from each of these values of u in order to plot the diffraction pattern with precisely located minima *and* secondary maxima as functions of θ.

It is somewhat useful to at least look at the phasor diagrams corresponding to the principle maximum and the first two minima and secondary maxima. In figure 12.18 the principle maximum (of length E_0 is illustrated for angle $\phi = 0$. The next two phasors show the (exact)

[177] Wikipedia: http://www.wikipedia.org/wiki/Transcendental Equation.

n	Exact	Approximate
3	4.48	4.71
5	7.72	7.85
7	10.9	11.0
9	14.1	14.1

Table 6: The exact value of $u = \phi/2$ in radians is in column 2. The approximate numerical value $u \approx n\pi/2$ is in column 3. Note that by the fourth root, the two are about the same at the precision of the table.

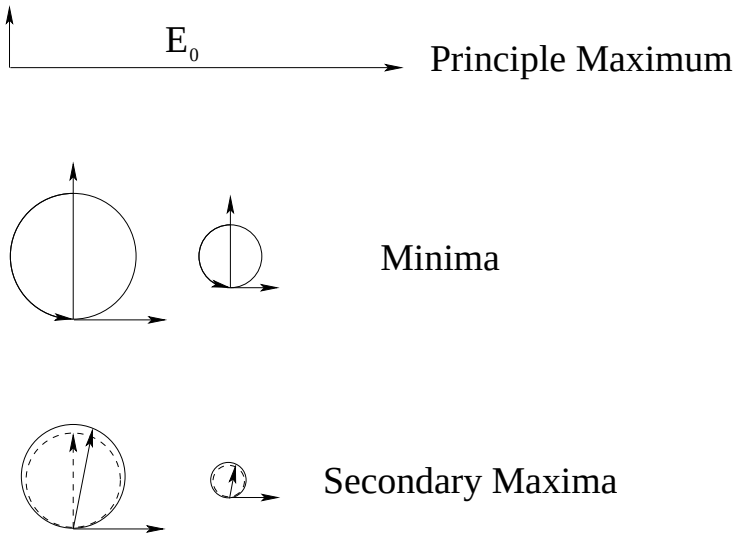


Figure 12.18: Phasor diagrams representing successive minima and maxima for single slit diffraction.

conditions for minima, where the "E_0-coil" is wrapped first one time around to reach its own tail at $\phi = 2\pi$ or *twice* around (the phasor snake having swallowed its own tail once and reached it a second time!) at $\phi = 4\pi$.

Note that the *diameter* of the E_0-coil has to get smaller as one wraps it around and around! The secondary maxima are now easy enough to understand. We don't get one at $\phi = \pi$ because we are still between the principle maximum and the first minimum, there is no maximum here. We do get one *near* $\phi = 3\pi/2$ (dashed circle and arrow), although we can gain a tiny bit of length by rolling the E_0-coil back to a slightly larger diameter as predicted by the transcendental solution above, ditto at/near $\phi = 5\pi/2, 7\pi/2...$ etc.

Before we move on, it is useful to note the approximate scaling of the secondary maxima field amplitudes and intensities. For the first secondary maximum, the phasor diagram suggests that the field amplitude should be the diameter of a circle where E_0 is wrapped around it 1.5 times. The second one is the diameter of a circle where E_0 is wrapped around it 2.5 times. We can extrapolate that if we let $n = 1, 2, 3...$ be the index of the secondary maximum:

$$E_n \approx \frac{2E_0}{(2n+1)\pi} = 0.21221, 0.12732, 0.09095...$$

From this we can also deduce that the secondary peak intensities should scale like this series

squared:

$$I_n \approx \left(\frac{2}{(2n+1)\pi}\right)^2 I_0 = 0.04503, 0.01621, 0.00827...$$

Note well that this series of secondary peak heights is ***independent of the relative size of*** a***!*** It doesn't matter if $a = 2\lambda$, or 4λ, or 7.334λ, the intensity peaks will scale down in *approximately* this way although as noted above, one will gain a *bit* of length unwinding the E_0 coil by a bit relative to the scheme pictured above. It's worth remembering this if you are asked to sketch a diffraction pattern (as you are in a homework problem) for some given value of a relative to λ. As always the *number* of minima or maxima attainable is limited by the fact that $|\phi| \leq 2\pi a/\lambda$, even though we can generate the "universal" phasor diagrams and scaling rules illustrated and estimated above past this limit.

It is now time to put it all together with a few examples.

Example 12.7.1: Diffraction Pattern of a Slit of Width $a = 4\lambda$

To draw the semiquantitatively correct $I(\theta)$ for a single slit, we must capture its *features* – both those we can compute or discover exactly as well as those that we can only guess at short of plotting the exact result. We'll find it a lot easier to plot not $I(\theta)$ but $I(\sin(\theta))$, so much so that I'm going to focus on this in the example. Note well that all we have to do to convert to or plot in terms of θ is take the inverse sines of the points we obtain.

We have seen above that we can exactly locate the principle maximum and the minima. We cannot *exactly* locate the secondary maxima, but we can guess their approximate location as roughly halfway between the minima in our drawing. Similarly, we can't exactly determine the intensity of the secondary maxima, but we do know that they have to get *smaller* as we increase their order, quite rapidly.

To facilitate drawing a graph with these features, we therefore begin by locating the minima:

$$\begin{aligned} a\sin(\theta_m) &= m\lambda \\ 4\lambda\sin(\theta_m) &= m\lambda \\ \sin(\theta_m) &= \frac{m}{4} \\ \theta_m &= \sin^{-1}\left(\frac{m}{4}\right) \end{aligned} \tag{12.86}$$

Let's arrange these for the values of m for which the inverse sine exists in a table. All angles are in radians.Don't forget to skip $m = 0$, the principle maximum!

m	$\sin(\theta_m)$	θ_m
1	$\frac{1}{4}$	$\sin^{-1}\left(\frac{1}{4}\right) = 0.25268$
2	$\frac{2}{4}$	$\sin^{-1}\left(\frac{1}{2}\right) = 0.52360$
3	$\frac{3}{4}$	$\sin^{-1}\left(\frac{3}{4}\right) = 0.84806$
4	$\frac{4}{4}$	$\sin^{-1}(1) = 1.00000$

Table 7: Diffraction minima for a single slit of width $a = 4\lambda$.

We see that it is a lot easier to draw the plot in terms of the *regular* $\sin(\theta_m)$ than it is in terms of θ_m. Of course, the latter is a lot more useful. Oh well, such is life. You should be *able* to do whichever one a problem requests on the homework or a quiz or exam. One reason I often accept results plotted in terms of $\sin(\theta_m)$ is that one doesn't usually need a calculator to do a decent job.

Figure 12.19 is a precise graph (generated with octave) of the diffraction pattern for $a = 4\lambda$ as a function of θ.

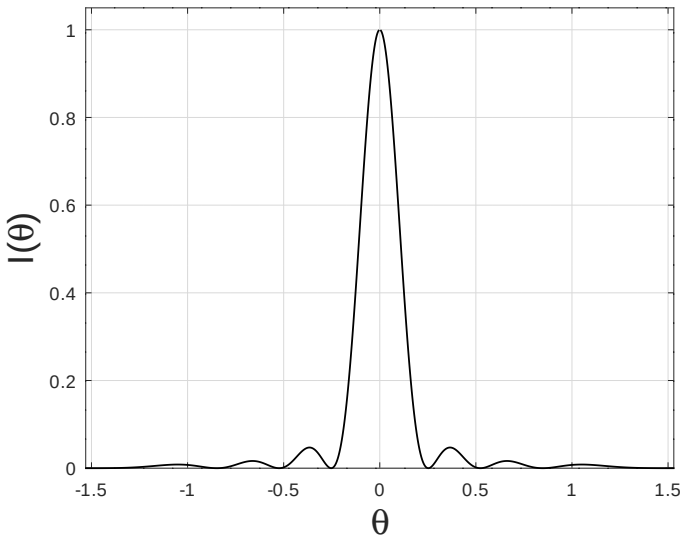


Figure 12.19: An exact graph of the diffraction pattern of a slit of width $a = 4\lambda$. Note the distortion of the horizontal scale by the inverse sine so that the minima are not evenly spaced. It's scaled so that $I_0 = 1$.

I redraw this, now as a function of $\sin(\theta)$:

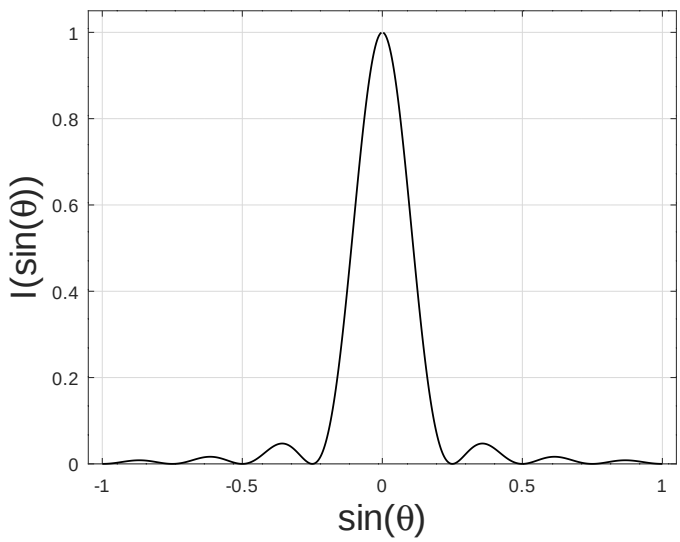


Figure 12.20: This graph, in terms of $\sin(\theta)$, is much easier to draw and requires no calculator to accurately locate the minima, which are evenly spaced.

Note well that the scaling of the three peaks visible in these figures appears to closely follow the series $0.05, 0.02, 0.01 \times I_0$ predicted to one significant digit by the scaling argument given above!

12.8: Two Slits of Finite Width

We are now ready to consider two slits of *finite* width. The result is very simple. We get interference maxima and minima at exactly the same angles we got them for very narrow slits. However, the field strength at those angles is *modulated* by the diffraction of the field through the individual slits. As a result, the field we observe as an angle of θ is the *product* of the field expressions for interference and diffraction:

$$E_{\rm tot}(\theta) = 2E_0 \cos(\delta/2) \left(\frac{\sin(\phi/2)}{\phi/2} \right) \tag{12.87}$$

Following the usual procedure (using the time average Poynting vector and relation between E_0 and B_0) we get the intensity

$$I_{\rm tot}(\theta) = 4I_0 \cos^2(\delta/2) \left(\frac{\sin(\phi/2)}{\phi/2} \right)^2 \tag{12.88}$$

Nothing to it. Note well that as always, $\delta = kd\sin(\theta)$ and $\phi = ka\sin(\theta)$, so this is an indirect function of θ linked by inverse sines.

Example 12.8.1: Two Slits of Separation $d = 8\lambda$ and width $a = 4\lambda$

We proceed exactly the same way we did for the previous example, except now we add two more tables: The angles of the *interference maxima* and the *interference minima*. We find these (as usual) from:

$$\sin(\theta_m) = \frac{m\lambda}{d} = \frac{m}{8} \tag{12.89}$$

for maxima and

$$\sin(\theta_m) = \frac{(2m+1)\lambda}{2d} = \frac{2m+1}{16} \tag{12.90}$$

for minima. The result is displayed in table 8. Using these numbers we can easily enough construct a combined interference/diffraction pattern, displayed in figure 12.21. For simplicity I only present the graph for $\sin(theta)$ – you can easily visualize or fill in a graph as a function of θ using the previous example as a guide to the distortion (or a piece of paper with an accurate graph scale on it). Note well the "squashed" interference that occur where there are diffraction *minima*. This illustrates a simple rule – when one of the two functions in the product above in $I_{\rm tot}$ are zero, zero wins!

Problems like this are graded on the basis of whether or not they contain the essential features illustrated herein. The various min's and max's should be correctly tablulated and located approximately correctly on the graph. The diffraction envelope should be qualitatively as shown, and the interference pattern should be drawn "under" it. If max's and min's occur at

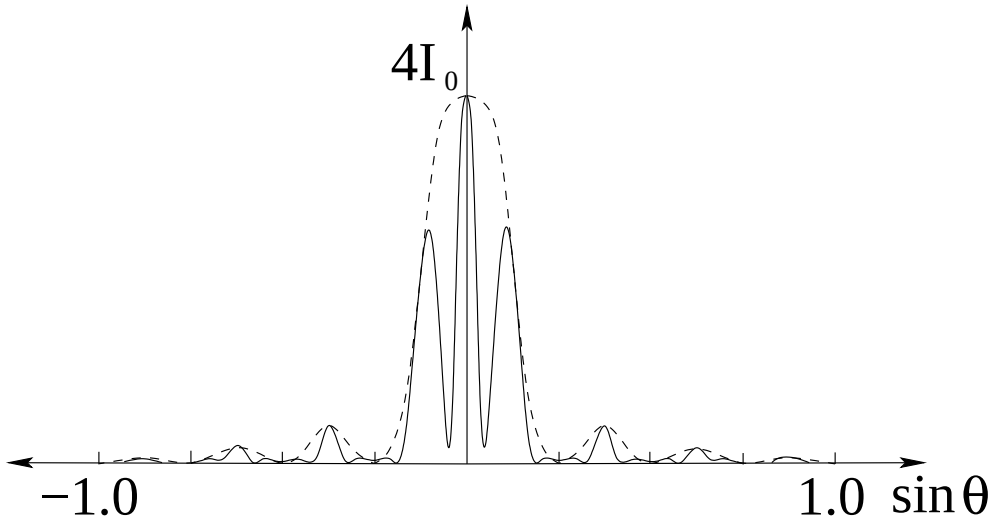


Figure 12.21: The graph of combined diffraction and interference, for $a = 4\lambda$ (same as before) and $d = 8\lambda$.

the same angle, the minimum wins. The maximum central intensity should be $4I_0$, where I_0 is the central intensity produced by a single slit.

Nothing to it!

12.9: Diffraction Through Circular Apertures – Limitations on Optical Instruments

Finally we are ready to understand how the use of waves with a finite (non-zero) wavelength affects things like vision and optical instrumentation. To start with, I have to give you a "true fact" concerning diffraction through a *circular aperture of radius* D – something that *can* be derived but that *I* won't derive just now in this work for you. It's not that the derivation is incredibly difficult or exotic – it proceeds more or less along the lines we've just used for single slit diffraction – it just is easiest to obtain using integration (which we avoided) and complex variables instead of phasors per se (which we have also mostly avoided).

In a nutshell, to obtain the result one has to do an integration in a sensible coordinate system (e.g. cylindrical coordinates) that sums up the differential electric field radiated from every point on the "disk" of Huygens radiators in the circular aperture, including their phase difference due to the path difference to an arbitrary point on the screen a distance Z away from the center of the aperture. To some people[178] this sounds like a really good time, but I'm guessing that for *most* students using this text it sounds like a still better time to *not* actually do it and hence you're inclined to forgive me for presenting something you actually have to just memorize/learn.

That true fact is this. The diffraction pattern produced on the screen by a circular aperture is itself a cylindrically symmetric "circle" of light, surrounded by alternating, ever fainter, rings

[178]Mostly physics or math majors or other mathochists, granted...

	Diffraction Minima	
m	$\sin(\theta_m)$	θ_m
1	$\frac{1}{4}$	$\sin^{-1}\left(\frac{1}{4}\right) = 0.25268$
2	$\frac{2}{4}$	$\sin^{-1}\left(\frac{1}{2}\right) = 0.52360$
3	$\frac{3}{4}$	$\sin^{-1}\left(\frac{3}{4}\right) = 0.84806$
4	$\frac{4}{4}$	$\sin^{-1}(1) = 1.57079$
	Interference Maxima	
m	$\sin(\theta_m)$	θ_m
0	0.0	$\sin^{-1}(0.0) = 0.00000$
1	$\frac{1}{8}$	$\sin^{-1}\left(\frac{1}{8}\right) = 0.12532$
2	$\frac{2}{8}$	$\sin^{-1}\left(\frac{1}{4}\right) = 0.25268$
3	$\frac{3}{8}$	$\sin^{-1}\left(\frac{3}{8}\right) = 0.38439$
4	$\frac{4}{8}$	$\sin^{-1}\left(\frac{1}{2}\right) = 0.52360$
5	$\frac{5}{8}$	$\sin^{-1}\left(\frac{5}{8}\right) = 0.67513$
6	$\frac{6}{8}$	$\sin^{-1}\left(\frac{3}{4}\right) = 0.84806$
7	$\frac{7}{8}$	$\sin^{-1}\left(\frac{7}{8}\right) = 0.94843$
8	$\frac{8}{8}$	$\sin^{-1}(1) = 1.57079$
	Interference Minima	
m	$\sin(\theta_m)$	θ_m
0	$\frac{1}{16}$	$\sin^{-1}\left(\frac{1}{16}\right) = 0.62540$
1	$\frac{3}{16}$	$\sin^{-1}\left(\frac{3}{16}\right) = 0.18862$
2	$\frac{5}{16}$	$\sin^{-1}\left(\frac{5}{16}\right) = 0.31782$
3	$\frac{7}{16}$	$\sin^{-1}\left(\frac{7}{16}\right) = 0.45282$
4	$\frac{9}{16}$	$\sin^{-1}\left(\frac{9}{16}\right) = 0.59741$
5	$\frac{11}{16}$	$\sin^{-1}\left(\frac{11}{16}\right) = 0.75804$
6	$\frac{13}{16}$	$\sin^{-1}\left(\frac{13}{16}\right) = 0.94843$
7	$\frac{15}{16}$	$\sin^{-1}\left(\frac{15}{16}\right) = 1.21538$

Table 8: Diffraction minima, interference maxima, and interference minima for a single slit of width $a = 4\lambda$.

of darkness (where destructive interference causes the total wave to cancel) and light (where partially constructive interference causes the total wave to peak, although never at the intensity seen in the central maximum). In fact, the *generic* shape of the diffraction pattern is much the same as that for a slit, only it is cylindrically symmetric instead of itself being a slit shaped bar with alternating bars of light and dark on the side. In this diffraction pattern the *first minimum* (the dark ring surrounding the bright(est) central maximum occurs at the angle given by:

$$D\sin(\theta_{\min}) = 1.22\lambda \tag{12.91}$$

Note that this is *almost* like the rule for the slit, $a\sin(\theta_{\min}) = \lambda$, except that we no longer get a pretty integer on the right and on the left we have the *diameter* of the aperture, not its short-direction width. It certainly makes dimensional sense.

Now consider viewing very distant, point-like objects through a circular aperture. I prefer to think of viewing stars, for example, as they are very distant indeed and appear to the eye as

mere points of light in the sky, through the aperture of your pupil, or the lens of a camera, or the lens of a telescope – it doesn't really matter what the aperture is as long as it is circular and symmetric.

The occurence of a lens in the aperture doesn't affect the diffraction – *every ray* gets bent by the lens to be focussed on the screen according to the angles in the diffraction patter, so the point-like object is focussed down not to a point, but to a circular dot. The *size* of the dot is basically determined by the angle of the first diffraction minimum, with smaller wavelengths being better resolved. Indeed, everything we learned in geometric optics, where source points on the object were mapped directly to image points by the lens, is what true physical optics predicts in the limit of *infinitely short wavelengths* (or more practically, wavelengths that are "infinitely" short compared to the aperture or length scales of the imaging apparatus)[179].

We can then ask: Suppose we are photographing a section of sky with our telescope and see a large, slightly asymmetric blob of "white" on our photograph corresponding to a light source in the sky. Is that blob the image of *one* object, or *two*? That is, is the source made up of the light from *two* objects (e.g. stars) or is it a slightly asymmetric single object (e.g. a lenticular galaxy)? Time to return to *Rayleigh's Criterion for Resolution!*

We can easily compute the capability of our telescope to resolve two objects that have a very small angle in between them using this criterion. Basically, if the peak produced by one object (center of the illuminated area on the film or charge-coupled device (CCD)[180] is separated from the other by at least the angle of the first diffraction minimum of the other, we can consider the two objects marginally resolved. This criterion depends on wavelength, and we intuitively expect our resolution to be better with e.g. blue or violet light than with red light[181]

The critical angle – which is certain to be a *very small angle* for any macroscopic aperture and optical frequency light – defining the diffraction resolution limit of an optical instrument is thus:

$$\theta_c \approx \sin(\theta_c) = \frac{1.22\lambda}{D} \tag{12.92}$$

Two stars with an angular separation greater than this critical angle will be clearly resolved on the film (assuming that the image is otherwise focussed on the film or CCD).

The same is true for two tiny features inside a bacteria or almost any two source objects imaged through a circular aperture. The central rays from object to image must be separated by more than $1.22\lambda/D$ or the two images will blur into one.

[179]This is actually a *very important result*, one worth reinforcing for possible math or physics majors. Geometric optics is the small wavelength limit of physical (wave) optics. Similarly, *classical mechanics* is the small wavelength limit of quantum (wave) mechanics! This answers one of the most important of questions from the Enlightenment – how light can behave like a particle (geometric) and wave (physical) at the same time, and extends it with the surprising result that microscopic objects like electrons and protons behave *exactly the same way*, with the same kind of schizophrenia producing particle-like behavior in one context or measurement apparatus, wave-like behavior in another.

[180]Wikipedia: http://www.wikipedia.org/wiki/Charge Coupled Device. A CCD is basically the "electronic film" used in digital cameras, consisting of a fine-mesh grid of photosensitive electrical units

[181]This same intuition has driven the invention of e.g. "blue ray" DVD formats that hold more information. Blue light has roughly half the wavelength of red light, so one can store roughly 4x as much information at the diffraction limit of resolution of blue light on disks compared to red. DVDs based on hard ultraviolet ($\lambda \sim 100 - 200$ nm) would hold a factor of 4 to 16 more data, and I'm quite certain that the minute I finish buying lots of blue-based movies UV DVD will be trotted out to replace it all yet again, this time on tiny DVDs...

Imaging nearly *anything* gets dicey when the objects themselves are the order of a wavelength in size or smaller. If you have ever seen water waves striking a pier support that is much smaller than a wavelength you know that they swirl right around it and recombine on the far side. A short distance away from the pier there is little sign in the shape of the wavefronts that there was a pier there at all. In order to reflect a wave or obstruct a wave, an object needs to be (ideally much) bigger than the wavelength of the wave.

Practically speaking, it is very difficult to create viewable images of objects much smaller than a half a micron using visible light. Bacteria are thus visible through a visible light microscope, but *structures* in or on the bacteria are not. Only the largest of viruses are visible with visible light.

To see objects smaller than the wavelength of visible light, one needs a wave with a smaller wavelength. Electron microscopes use electron "waves" to see objects as small as 5 nm – small enough to see most viruses in considerable (beautiful) detail[182]

We can see that physicians and physicists alike need to have a fairly clear idea of the role that waves play in the formation of the magnified images that permit us to see the very small or the very far away. It is quite easy to build microscopes and telescopes for which *diffraction*, *wave interference* and things like *chromatic distortion* are the limiting factors that prevent us from being able to see further, smaller, better. Even if you will never actively design a microscope or telescope, understanding their limitations will make you a better consumer of the information that they can provide.

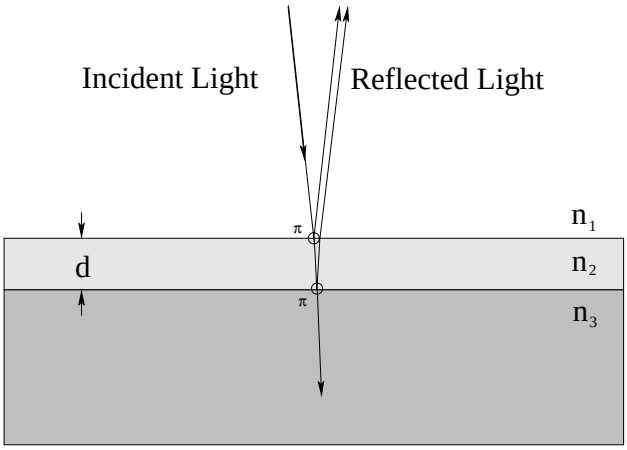
12.10: Thin Film Interference

Observing interference from slits thick or thin, at optical frequencies, is a bit of a rarity in everyday life. We just don't trip over visible light travelling through multiple pathways *within the coherence length of the light* to reach a common goal every day, given that the coherence length of light from hot/chaotic sources is the order of a few microns (tens to perhaps a hundred wavelengths). Exceptions do include – for a few people – diffraction limited viewing through visible light telescopes and microscopes, discussed above, or people who use spectrographs based on diffraction gratings. Well, I suppose I should include the rainbow of colors one can see on the bottom of CDs or DVDs, which are basically reflection-based diffraction gratings as light bounces off of the many tiny tracks scored in the reflective surfaces – *now* that is an everyday experience but it hasn't always been so.

Thin film interference, however, is something that we *might well* observe every day, or nearly so. Every time we blow a soap bubble, or see a slick of oil or gasoline on water, swirling around with many colors, we are observing thin film interference. Whenever we look at the lens of a camera and see a lack of reflections or those same "metallic" colors, we are seeing thin film interference. Thin film interference gives color and life to ornaments and has various other technological or social applications, even if those who observe it don't realize what it is.

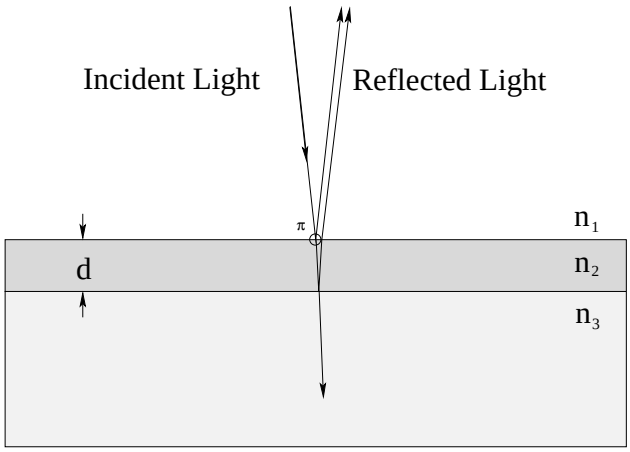
We'd like to understand it and learn to recognize it and see one or two of its applications.

[182]Wikipedia: http://www.wikipedia.org/wiki/Virus. This article has some lovely transmission electron micrographs of viruses, revealing detail that would be completely invisible to the eye even with the aid of a powerful visible light microscope.



$n_1 < n_2 < n_3$
(Two phase shifts of π)

Figure 12.22: One of the two basic diagrams for thin film interference. The total phase difference in the superposed reflected waves in the case $n_1 < n_2 < n_3$ or $n_3 < n_2 < n_1$ is just $\delta = k'(2d)$, as the phase shifts produced by reflecting off of the two surfaces are either both zero or both (as they are in this case) π, in which case they cancel.



$n_1 < n_2 > n_3$
(One phase shift of π)

Figure 12.23: The second of the two basic diagrams for thin film interference. The total phase difference in the superposed reflected waves in the case $n_1 < n_2 > n_3$ or $n_3 < n_2 > n_1$ is $\delta = k'(2d) + \pi$, as there is a phase shift of π produced by reflecting off of the surface of a material with a higher index of refraction only one one of the two surfaces..

Fortunately, it is (at this point) quite simple. Here's the idea.

In figures 12.22 and 12.23 a thin film of transparent material sits in between two other transparent materials. Each material has its own index of refraction, and we will for the moment use the convention that n_1 is the index of refraction of the material the light is coming *from*, n_2 is the index of the thin film itself, and n_3 is the index of the material the light is going *to*.

Incident light (often white light, a mixture of all the visible colors/wavelengths) is incident approximately "normally" onto (coming in perpendicular to) the surface between n_1 and n_2.

Some fraction of this light reflects off of the interface; the rest is transmitted into n_2. Of the light that makes it into n_2 and then is incident normally on the interface between n_2 and n_3. Again, some fraction is reflected and some is transmitted. Finally, the light that is reflected back up arrives at the interface between n_1 and n_2 a second time, this time coming from below, and a fraction of it is transmitted back into medium n_1, where the electromagnetic wave *combines* with the original reflected wave.

The interference we observe thus comes from adding two waves:

$$E_{\rm tot} = E_{12}\sin(kr - \omega t + \delta_{12}) + E_{23}\sin(kr - \omega t + \delta_{23}) \tag{12.93}$$

where (as we will see below) there is a chance of a phase shift occurring in *both* reflected waves compared to the phase of the incoming wave. Note also that it is almost certain that $E_{12} \neq E_{23}$, that is, the two reflected waves will very likely have somewhat different amplitudes as they recombine.

Presuming that these two waves have at least *approximately* equal field amplitudes and a consistent phase difference brought about at least partly by path difference (the wave that traverses the film twice travels a distance $2d$ farther than the wave that reflects of of the first surface), this superposition will partially cancel or partially add the waves for different wavelengths. Some wavelengths will be brightened, others diminished. The reflected white light will therefore take on those characteristic mauves and greens and poisonous shiny blues that are familiar to us all.

Of course, there are a few *details* we have to consider, and they are important; they are why we need *two* figures (and two phase shifts) to demonstrate two of the four possible patterns of sort order of the indices of refraction. In a nutshell, two things contribute to the overall phase shift between the recombined waves – the phase shift due to the path difference in the medium n_2 *and a phase shift caused by reflecting off of a medium with a higher index of refraction!* Let's begin by working out the former, as that is easiest, and then we'll talk extensively about the latter, as the phase shifts due to reflection off of the surfaces themselves will require us to go back to our intro physics 1 course and recall e.g. the *reflection of waves on strings* off of interfaces between a light string (where the speed of the wave is large) and a heavy string (where the speed of the wave is less).

12.10.1: Phase Shift Due to Path Difference *in the Thin Film!*

This one, as promised, is easy. The wave that traverses the thin film (twice!) goes an additional distance $\Delta r = 2d$ compared to the wave that reflects off of the upper surface. We are thus tempted to (after "reflection"[183] on what we have learned so far) to associate with this path difference an additional phase $\delta_{\rm path} = k(2d)$.

As it turns out, this heuristic guess is *almost* correct! But as the saying goes, "almost" only counts in horseshoes and hand grenades[184]. The problem is that the path difference accumulates *while the wave is in the thin film!* To get the phase difference right, then, we have

[183] Har, har...

[184] ...and possibly even other things that begin with 'h', such as hydrogen bombs. Being "almost" hit by a hydrogen bomb can ruin your whole day...

to use the wavelength (and hence wave number) *in the thin film medium* n_2, not the one we used in the originating medium n_1, or worse, the one that the light would have in a vacuum!

You should recall that:

$$\lambda_2 = \frac{\lambda}{n_2} \tag{12.94}$$

where λ is the wavelength of the light in a vacuum. This leads to a wavenumber of:

$$k_2 = \frac{2\pi n_2}{\lambda} \tag{12.95}$$

and a phase shift of:

$$\delta_{\rm path} = k_2(2d) \tag{12.96}$$

Basically, the wave that traverses the thin film accumulates phase at the spatial rate of k_2, not k, k_1, or $k3$.

Using k instead of k_2 is a very common mistake made by students of physics! *Don't let it be you!*

Next, let's examine the phase shifts due to the actual reflections themselves.

12.10.2: Phase Shifts Due to Reflections at the Surfaces

As you should remember from the treatment of waves in the first half of this course (see my [185] book online if all of this eludes you.), a wave pulse on a string that partially reflects off of the junction with a *heavier* string (slower speed) *flips over*, where a wave pulse on a heavier string that partially reflects of off the junction with a lighter one does not. The transmitted wave pulse in both cases does not flip.

Exactly the same thing happens for harmonic wave trains or wave pulses in the case of light. If a harmonic light wave reflects off of a denser medium (which usually has a higher index of refraction and a slower velocity of light) the reflected wave *inverts*. Inversion is basically multiplication by a minus sign, or equivalently (for harmonic waves) shifting the *phase* of the reflected wave by π or the heuristic equivalent half-wavelength. If a harmonic light wave reflects off of a lighter medium (lower index of refraction) the reflected wave does not flip, it retains it's original phase.

There are thus four permutations of sort order for the indices of refraction n_1, n_2, n_3. They are:

I *strongly recommend* that when you solve a problem involving thin film interference, you *circle the reflections* that have a phase shift $\delta_{ij} = \pi$ and write a little "π" next to each one, as I did in figures **??** and 12.23 above. Then you are less likely to forget to include it in your overall computation and understanding of the total relative phase shift. **Leaving out one or more of these phase shifts** (and getting the max's and min's backwards as a result) **is another common error. Don't do it!**

Now we are ready to put all of this together and and determine the heuristic conditions for maxima and minima. We'll do this twice, once for each of the two "opposite" rules one gets for max's and min's.

[185] http://www.phy.duke.edu/ rgb/Class/intro_physics_1.php Introductory Physics 1

Permutation	δ_{12}	δ_{23}	$\lvert\Delta\delta\rvert$
$n_1 < n_2 < n_3$	π	π	0
$n_1 > n_2 > n_3$	0	0	0
$n_1 < n_2 > n_3$	π	0	π
$n_1 > n_2 < n_3$	0	π	π

Table 9: Relative phase shift introduced between the wave reflected off of the $n_1 \to n_2$ interface and the transmitted wave reflected off of the $n_2 \to n_3$ interface. Note that in the first two cases (smoothly increasing or decreasing n) there is no net phase shift with n_2 "in the middle". In the second two cases, the index of refraction of the thin film medium is either higher than that of its neighbers or lower, but not in the middle.

12.10.3: No Relative Phase Shift from Surface Reflections

Consider the case where $\delta_{12} = \delta_{23} = 0$ or π. In both of these cases there is no *relative* phase shift due to the reflections. Either both waves flip (and hence accumulate phase difference only due to the path difference) or neither wave flips (ditto). Either way, the *total* relative phase shift δ is just due to the path difference:

$$\delta = k_2(2d) = \frac{2\pi n_2}{\lambda}(2d) = \frac{4\pi n_2 d}{\lambda} \tag{12.97}$$

We can now use our simple heuristic rules for max's and min's: If the path difference is an integer number of wavelengths λ_2 *in the thin film*, then we expect the two waves to recombine in phase and while the resultant amplitude may not be *twice* either of the two waves, it will certainly be larger than either one alone. Similarly, if it is an odd-half integer number of wavelengths in the film, we expect the waves to be exactly out of phase and to maximally cancel. We'll summarize this as:

$$2d = m\lambda_2 = m\frac{\lambda}{n_2} \qquad m = 0, 1, 2... \qquad \text{maxima} \tag{12.98}$$

$$2d = \frac{2m+1}{2}\lambda_2 = \frac{(2m+1)}{2}\frac{\lambda}{n_2} \qquad m = 0, 1, 2... \qquad \text{minima} \tag{12.99}$$

Of course, this is only heuristic. The "correct" way to arrive at the same place is to set δ to $0, 2\pi, 4\pi...$ for constructive interference and to $\pi, 3\pi, 5\pi...$ for destructive interference. It is left as a fairly simple (and hopefully by now, familiar) exercise for the student to show that if you do this, you arrive precisely at our heuristic rules.

12.10.4: A Relative Phase Shift of π from Surface Reflections

Consider the cases where *either* δ_{12} *or* δ_{23} is π and the other is 0. In both of these cases there *is* a relative phase shift due to the reflections. One of the two waves flips (and hence "suddenly" accumulate an additional phase of π and the other does not. No matter which wave flips the *total* relative phase shift δ must add or subtract this relative phase to the one from the path difference:

$$\delta = k_2(2d) = \frac{2\pi n_2}{\lambda}(2d) \pm \pi = \frac{4\pi n_2 d}{\lambda} \pm \pi \tag{12.100}$$

Note that the sign we get differ depending on which one flipped. However, we don't really care which sign we get. This is because $\sin(\theta + \pi) = \sin(\theta - \pi) = -\sin(theta)$, so we can simply move a π with either sign to whatever side of the equals sign that seems convenient to us. In order to get the best correspondance with our heuristic rules, we should probably use the minus sign no matter which one flipped (which I just proved that we can do):

$$\delta = k_2(2d) = \frac{2\pi n_2}{\lambda}(2d) - \pi = \frac{4\pi n_2 d}{\lambda} - \pi \tag{12.101}$$

That will let us move it over onto the same side as the other π's with a plus sign later.

The heuristic rules for max's and min's, are now *exactly the opposite* of the ones above:

$$2d = \frac{2m+1}{2}\lambda_2 = \frac{(2m+1)}{2}\frac{\lambda}{n_2} \qquad m = 0, 1, 2... \qquad \text{maxima} \tag{12.102}$$

$$2d = m\lambda_2 = m\frac{\lambda}{n_2} \qquad m = 0, 1, 2... \qquad \text{minima} \tag{12.103}$$

This is because the extra phase shift of π or minus sign in the wave corresponds to exactly *half of a wavelength path difference in the medium*, just enough to make the two rules swap places. In words, if the path difference contains an odd-half integer number of wavelengths in the medium, the phase shift of π at the surface contributes the equivalent of another half wavelength and the waves will recombine constructively *in* phase. Similarly, if the path difference in the medium contains an integer number of wavelengths, the extra phase shift puts them back exactly out of phase for (maximally) destructive interference and a minimum.

Again, the "correct" way to arrive at this heuristic is to set δ to $0, 2\pi, 4\pi...$ for constructive interference and to $\pi, 3\pi, 5\pi...$ for destructive interference. The extra factor of π is there, ready to be moved to the other side with whatever sign that pleases you. Again, a diligent student should verify that this leads straight to the heuristic rules.

12.10.5: The Limits of *Very* Thin Films

The occurrence of discrete phase shifts of π upon reflection from none, one, or both surfaces has one easily observable consequence. A *very* thin film, one that is much thinner than a wavelength ($d \ll \lambda$) will have *no* phase shift from path difference, as the film isn't thick enough. The only shifts that matter, then, are those that arise from the inversions reflecting off of a higher-n interface. There are as before only two combinations that matter – no *relative* reflection shift or a relative reflection shift of $\pm\pi$.

In the former case (two shifts or no shift's, no *relative* shift), light reflected from the upper and lower surface emerge in phase *for all wavelengths!* The surface becomes shiny white, even mirror-like.

In the latter case (one shift in either order), light comes off of the surfaces almost exactly out of phase for all wavelengths, and destructive interference results. Light is not reflected from the surface; it becomes extremely *transparent*.

Whether or not you know it, you have probably observed concrete examples of both of these limits. For example, a drop of oil or gasoline that falls onto a rain puddle over black pavement instantly spreads out and forms a thin film. We have all seen the initial rainbow

swirl of strange "metallic" colors, followed by the surface becoming shiny and grey. What one is seeing is the oil forming a layer on top of water with the order of indices of refraction $n_{\rm air} < n_{\rm oil} < n_{\rm water}$.

A second "experiment" – one that is greatly enjoyed by physics students the world over, including very young ones – is to blow soap bubbles[186]. All of us are familiar with the swirl of colors seen in the reflections from these spherical balls of thin soap film, and at this point you should understand that colors are the results of the enhancement of some wavelengths of light in the visible band and diminishment of others, constantly varying as the soap swirls around in the film (and the film thickness changes minutely) and as the angle of incidence and reflection of the light is varied by perspective.

If you blow a nice, big bubble that just hangs there for a time on a still day, supported by the slight buoyancy of the warm air of the breath with which you blew it, you will probably observe the following, although how successful you are may depend on the particular mix of soap you are using (some soap mixtures 'pop' more quickly than others).

As you watch, the color swirl will settle down and become colored not-quite rainbow like *rings* concentric around the vertical axis, and concentrated in the bottom half of the bubble. You may see several sets of rings at some point. What is happening is that the bubble soap is sinking under the influence of gravity and "bulging" the film at the bottom and thinning it out on top. At the same time, of course, the film is evaporating – getting thinner as the water molecules in the film thermally bounce free.

On the top, a curious thing happens. The film stops exhibiting color at all – it becomes *completely transparent!* In fact, as the water evaporates, the entire bubble may become almost completely invisible, revealed only by a hint of distortion at the outside edge of the sphere and an almost invisible tracing of lines where the soap is ever so slightly thicker and holding the bubble together.

This transparency is caused, as noted above, but light reflecting off of the *first* surface with a phase shift of π (functionally, a half of a wavelength) and reflecting off of the second surface with no phase shift. Once the film is much thinner than a wavelength, light in all wavelengths thus recombines *destructively*, largely cancelling the reflected wave. Light that isn't reflected is transmitted; hence the soap bubble becomes transparent.

This trick is used to advantage to make advanced optical coatings for e.g. binoculars, telescopes, microscopes, and other optical instruments. By covering the outer surface of the primary lens with a thin (< 100 nm) coating with a *higher* index of refraction than the glass, destructive interference in all visible wavelengths is assured, resulting in a lens that *maximizes light transmission*. High quality coated optics deliver 90+% of the light that is incident on them to the eye of the observer, which makes a big difference when compared to expected reflection/transmission intensities for the glass-air interface alone[187].

[186] That's right, this is an *assignment!* Go down to the store and get a bottle of bubble soap in any size that suits you. Blow bubbles, the bigger the better, ideally on a still, quiet, warm day where you get good 'hang time'...

[187] In my online book *Classical Electrodynamics II* I derive the *transmission coefficient*

$$T = \frac{4n_1 n_2}{(n_1 + n_2)^2}$$

for normal reflection. This is the fraction of intensity that is transmitted at an interface between two otherwise perfectly transparent media with differing indices of refraction. We omit discussing transmission and reflection

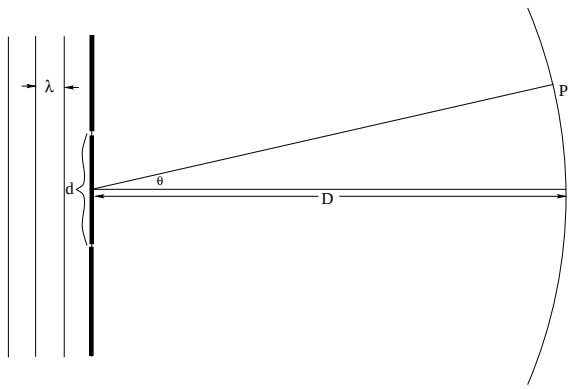
Homework for Week 12

Problem 1.

Physics Concepts

Make this week's physics concepts summary as you work all of the problems in this week's assignment. Be sure to cross-reference each concept in the summary to the problem(s) they were key to. Do the work carefully enough that you can (after it has been handed in and graded) punch it and add it to a three ring binder for review and study come finals!

Problem 2.



a) Use the phasor approach to derive the intensity:

$$I(\theta) = 4I_0 \cos^{\frac{\delta}{2}} \qquad \text{where} \qquad \delta = kd\sin\theta$$

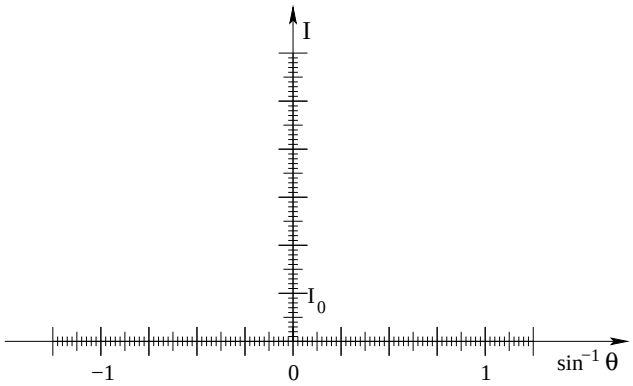
cast on a cylindrical screen a distance $D \gg \lambda$ away as a function of θ for two *narrow* ($a \ll \lambda$) slits separated by a distance d and illuminated by monochromatic light of wavelength λ as drawn above (D not to scale) where I_0 is the intensity due to a single slit alone.

b) For $d = 4\lambda$, find the (inverse sine of the) angles where the intensity is maximum and minimum.

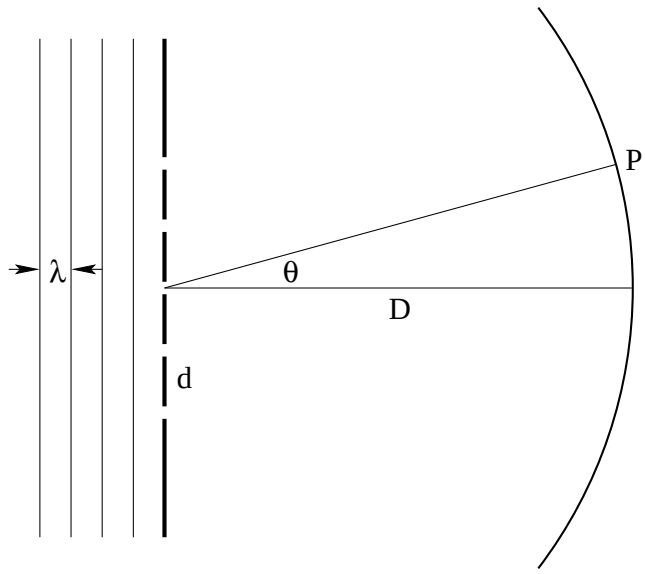
c) Hand-sketch the interference pattern for $\theta \in [-\pi/2, \pi/2]$ on (a copy of) the graph below.

coefficients in this book because they are too difficult to derive or handwave, arising from solving the boundary value problem on the surface between the two media.

However, for air ($n_a \approx 1$) and glass ($n_g \approx 3/2$) the expected transmitted fraction of the intensity from each air-glass surface (in either direction) is thus $T = 0.96$. For four surfaces (two lenses), this means that only 85% of the light makes it through to the eye, less if there are additional reflecting surfaces or lenses in the optical path, less still from filters or absorption by the glass (which is small but not zero). Coating can increase the transmitted fraction to 0.98-0.99 (per surface) and thus transmit an easy 10% more light.


I
I0
−1
0
1
sin−1 θ

Problem 3.



Five extremely narrow slits ($a \ll \lambda$) separated by a distance $d > \lambda$ are illuminated by light with wavelength λ which in turn project light onto a cylindrical screen at a distance $D \gg 5d$ as shown (D not to scale). Draw the phasor diagrams corresponding to the ***principle maxima***, the ***minima***, and (approximately only) to the ***secondary maxima***, assuming that the light reaching the screen from all slits is coherent. Graph what you expect the interference pattern to look like if $d = 4\lambda$ (in terms of I_0, the intensity of a single slit).

Problem 4.

a) Following the text and lecture, ***derive*** the intensity as a function of θ for the single slit problem for a cylindrical screen in the usual limit $D \gg a > \lambda$.

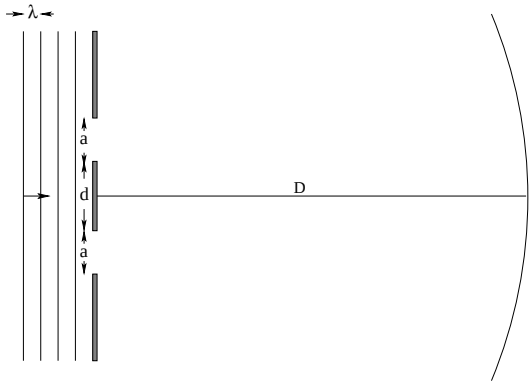
b) For $a = 3\lambda$, find the angles where the intensity is a ***minimum*** and put them in a table.

c) Sketch the diffraction pattern as a function of $\sin(\theta) \in [-1, 1]$

d) Sketch the diffraction pattern as a function of $\theta in[-\pi/2, \pi/2]$.

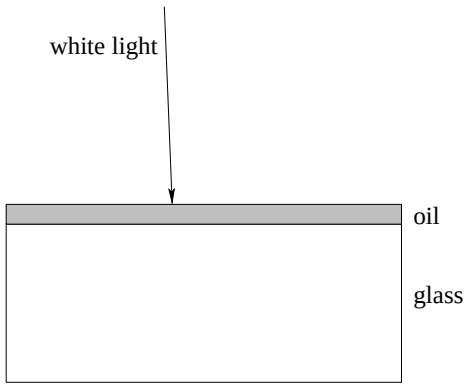
Compare/contrast these two sketches. Which one is easier to draw? Why?

Problem 5.



Suppose that two slits with *finite* width $a = 3\lambda$ and separation $d = 6\lambda$ are illuminated by plane-wave light with wavelength λ as shown above and cast the transmitted light onto a cylindrical screen a distance $D \gg d$ away. Determine all of the interference *and* diffraction minima and maxima (the latter can be approximate for diffraction and all can be given in terms of inverse sines) and sketch a *qualitatively* correct picture of the interference pattern underneath the diffraction envelope.

Problem 6.

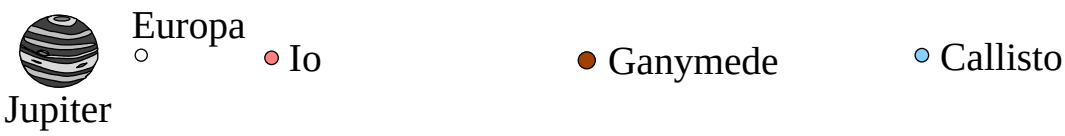


A thin film of oil with index of refraction $n_o = 5/4 = 1.25$ is smeared on a "thick" piece of glass ($n_g = 3/2 = 1.5$ as drawn above. The film is illuminated and viewed from directly above with white light.

a) What is the smallest (nontrivial) mean thickness t that the film must have such that reflected light to has a constructive interference *maximum* somewhere in the visible spectrum. At this thickness, does this maximum occur at the violet or red end of the spectrum?

b) Suppose that the film is only a few nanometers thick (much smaller than this minimum t). Does the film on the glass turn ***shiny*** (constructively reflecting all wavelengths) or ***transparent*** (destructively reflecting all wavelengths)? **Explain your answer.**

Problem 7.



Joe Braggart claims to have *really, really good vision*. "Why," he says. "My vision is *so* good I can make out the Galilean moons of Jupiter with my naked eyes on a really clear night. If I'd been around at the time of Galileo humanity wouldn't have had to invent the telescope in order to confirm the Copernican theory!"

Callisto is the moon with the largest orbit and has a maximum distance from Jupiter of just under 2×10^6 kilometers. At its closest point to the earth, it is around 600×10^6 kilometers away. Assuming that he is using visible light, is there a chance that he's telling the truth? Note well: This is a problem on ***resolution***, not lenses or the sensitivity of the retina, so please determine whether or not Jupiter and Callisto are *resolved* by the human eye at this distance.

Problem 8.

As shown in the text, the diffraction pattern cast by a single slit of width a (in the usual circumstance of a cylindrical screen a distance $D \gg a > \lambda$ away) is:

$$I(\theta) = I_0 \left(\frac{\sin \phi/2}{\phi/2} \right)^2$$

where $\phi = ka \sin \theta$

Use this formula and calculus to obtain an *expression* for the angles where **all the diffraction minima *and maxima*** occur. You might find it useful to recall that the differential of a function squared is given by:

$$df^2 = 2f\,df$$

Also recall (it is easy to show from the Taylor series, for example) that:

$$\lim_{x \to 0} \frac{\sin x}{x} = 1$$

and hence is not "undefined".

Advanced Problem 9.

Let's study diffraction gratings:

a) Following the textbook, derive for yourself the expression $R = mN = \frac{\lambda}{\Delta\lambda}$ for resolution for a diffraction grating with N slits of separation d.

b) Use it to determine the slit separation d such that the "famous" sodium spectral doublet with lines at $\lambda_1 = 589.0$ and $\lambda_2 = 589.5$ nanometers occurs at first order at the approximate angle of $\theta = \pi/6 = 30°$.

c) Then determine the miminum number of slits that have to be illuminated within the coherence length of the sodium light in order to resolve the doublet in the first order.

Part II

Electronics

Week 13: Alternating Current Circuits

Generation and Transmission

- **AC Generator:** If one spins a coil with N turns and cross-sectional area A at angular velocity ω in a uniform magnetic field B oriented so that it passes straight through the coil at one point in its rotation, one generates an *alternating voltage* according to:

$$\phi_m = \vec{\boldsymbol{B}} \cdot NA\hat{\boldsymbol{n}} = NBA\cos(\omega t + \theta) \tag{13.1}$$

$$V(t) = -\frac{d\phi_m}{dt} = NBA\omega\sin(\omega t + \theta) \tag{13.2}$$

  where θ is an arbitrary phase corresponding to the choice of when we start our clock. We will from now on choose the phase such that a standard harmonic alternating voltage source has its maximum at $t = 0$:

$$V(t) = V_0\cos(\omega t) \tag{13.3}$$

  Later, the choice of cosine makes the connection between complex and real approaches to driven AC circuits more consistent but we could just as easily have chosen sine.

- The most common models for household electrical distribution are represented in the following table (note well that $\omega = 2\pi f$ where f is the frequency of the source in Hertz): 208 is the potential difference between any two phases of a three-phase "Wye" main.

Volts	Hz	Purpose	Continent
120	60	lighting, small appliances, electronics	N. and S. America
208 or 240	60	heating, cooling, large appliances, 3 phase motors	N. and S. America
230	50	all household use	Everywhere else

Table 10: Common alternating voltages and frequencies in use around the world. There is a dazzling array of plug types in use around the world as well.

- **The Ideal Transformer:** The transformer is basically a pair of flux-coupled coils, one (the *primary*) with N_p turns connected to the *source* of alternating voltage, the other (the *secondary*) with N_s turns connected to the *load* that actually consumes the energy delivered from the source. If we let ϕ_m be the flux trapped in the core that passes through

a single turn, then:

$$V_s = N_s \frac{d\phi_m}{dt} \tag{13.4}$$

$$V_p = N_p \frac{d\phi_m}{dt} \tag{13.5}$$

or (taking the ratios of these two equations, in order)

$$\frac{V_s}{V_p} = \frac{N_s}{N_p} \tag{13.6}$$

Since an ideal transformer has no resistance, the power entering one side equals the power exiting the other, so:

$$V_s I_s = V_p I_p \quad \Rightarrow \quad \frac{I_s}{I_p} = \frac{N_p}{N_s}$$

Stepping *up* the voltage causes the current to step *down*. This underlies their function in the power grid that delivers energy to end users.

- The **Power Grid** step the voltage produced at the generator ***up*** to ***transmit at high voltage and low current***. Since high voltage can arc through air to ground and is enormously dangerous, it is then stepped ***down*** to *use at low voltage and high current*. There is *always* a step-down transformer at the very end of the line, that drops the power-line voltage from 12,000-14,000 volts to the ***much safer but still dangerous 120 volts*** (relative to ground) that we actually use.

Passive Circuits (with no driving AC voltage)

- **The Passive LC circuit:** In figure 13.1, the capacitor C on the left is initially charged up

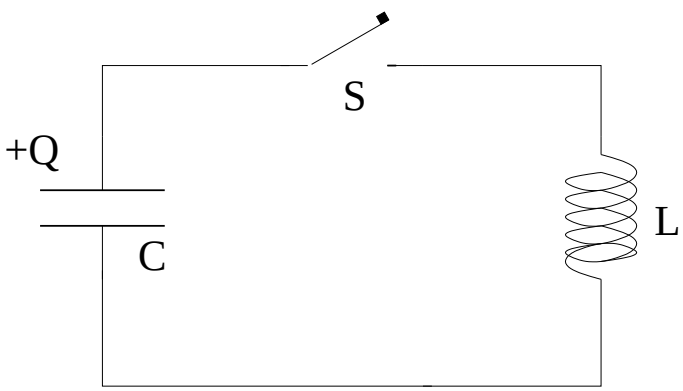


Figure 13.1: Undriven LC circuit

to charge Q_0. At time $t = 0$ the switch is closed and current begins to flow. If we apply Kirchhoff's voltage/loop rule to the circuit and solve the equation motion (as detailed in the text) we get:

$$Q(t) = Q_0 \cos(\omega_0 t) \quad \Rightarrow \quad I(t) = -\frac{dQ}{dt} = \omega_0 Q_0 \sin(\omega_0 t) \qquad \text{with} \quad \omega_0 = \frac{1}{\sqrt{LC}} \tag{13.7}$$

as the charge on the capacitor and the current in the circuit as functions of time.

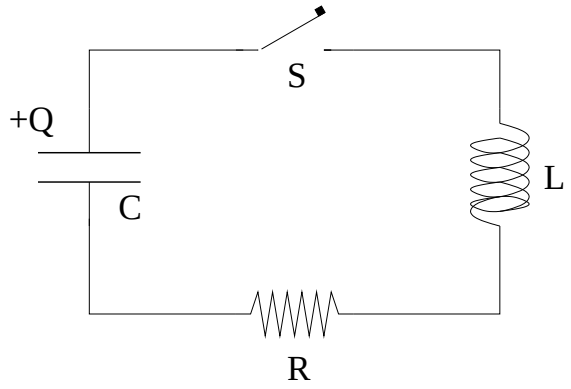


Figure 13.2: Undriven LRC circuit

- **The Passive LRC circuit:** In figure 13.2, the capacitor C on the left is initially charged up to charge Q_0. At time $t = 0$ the switch is closed and current begins to flow. If we apply Kirchhoff's voltage/loop rule to the circuit and solve the equation motion (as detailed in the text) we get:

$$Q(t) = Q_0 e^{\frac{-Rt}{2L}} \cos(\omega' t) \qquad \text{where} \quad \omega_0 = \omega_0 \sqrt{1 - \frac{R^2 C}{4L}} \tag{13.8}$$

  is the shifted frequency of the (**underdamped**) oscillator circuit.

- **Energy Loss in the Passive LRC circuit:** As we did for the strictly analogous damped simple harmonic oscillator, we define the $\mathcal{Q}$-factor of the circuit to be 2π times the fractional energy loss per cycle:

$$\mathcal{Q} = 2\pi \frac{E}{|\Delta E|} = \frac{\omega_0 L}{R} \tag{13.9}$$

  where $|\Delta E|$ is the energy loss over one cycle (period T) and E is the total energy in the circuit at the start of the cycle. The second form is valid in the weak damping limit, practically speaking when $\mathcal{Q} \gtrsim 3$ as shown in the text.

Active (Driven) AC Circuits

- **AC Voltage Across vs Current Through a Resistance R:** If the voltage across the

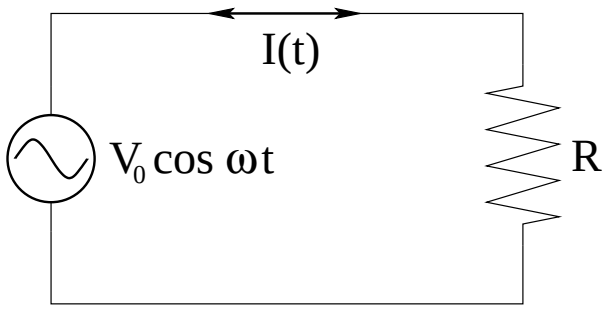


Figure 13.3: AC voltage across R

resistor is $V_0 \cos(\omega t)$ as shown, then (from KLR):

$$I_R(t) = \frac{V_0}{R} \cos(\omega t) \tag{13.10}$$

The current through a resistor is *in phase* with the voltage drop across it.

- **AC Voltage Across vs Current Through a Capacitor C:** If the voltage across the

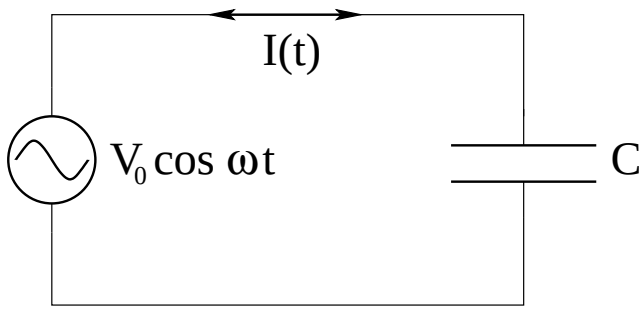


Figure 13.4: AC voltage across C

capacitor is $V_0 \cos(\omega t)$ as shown, then (from KLR plus differentiation):

$$I_C(t) = I_0 \cos(\omega t + \pi/2) \tag{13.11}$$

where

$$I_0 = (\omega C)V_0 = \frac{V_0}{\chi_C} \qquad \text{with} \quad \chi_C = \frac{1}{\omega C} \tag{13.12}$$

The current through a capacitor is $\pi/2$ *ahead of* the phase of the voltage drop across the capacitor. Note that χ_C (the "capacitative reactance") has units of *ohms*.

- **AC Voltage Across vs Current Through an Inductor L:** If the voltage across the in-

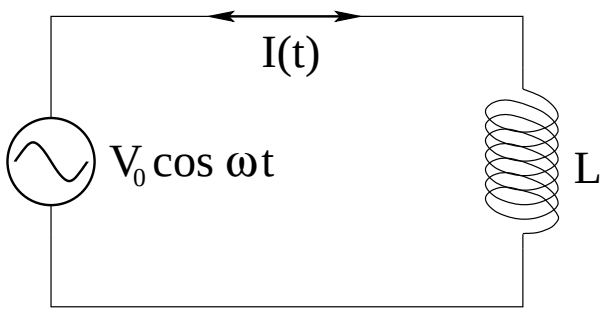


Figure 13.5: AC voltage across L

ductor is $V_0 \cos(\omega t)$ as shown, then (from KLR plus integration):

$$I(t) = I_0 \cos(\omega t - \pi/2) \tag{13.13}$$

where

$$I_0 = \frac{V_0}{\omega L} = \frac{V_0}{\chi_L} \qquad \text{with} \quad \chi_L = \omega L \tag{13.14}$$

The current through a capacitor is $\pi/2$ *behind* the phase of the voltage drop across the capacitor. Note that χ_L (the "inductive reactance") has units of *ohms*.

- **AC Voltage Across vs Current Through a Series LRC Circuit:** If the voltage across the LRC circuit is $V_0 \cos(\omega t)$ as shown, then (from either phasor analysis or the complex algebraic solution) the steady-state current in the circuit is given by:

$$I(t) = I_0 \cos(\omega t - \delta) \tag{13.15}$$

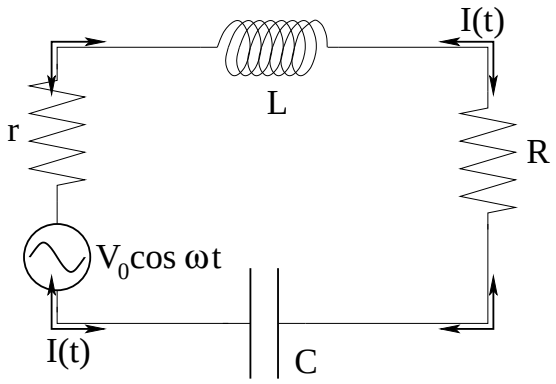


Figure 13.6: A series LRC circuit, for a non-ideal power supply with internal resistance r (and hence *total* resistance $R' = r + R$. R is considered the *load* resistance – where useful work is done by the circuit.

The current amplitude I_0 and phase δ (relative to the driving voltage) are ***not free parameters!*** They are entirely determined by the elements in the circuit!

- **Current Amplitude, Impedance, and Phase in a Series LRC Circuit:** The "effective total resistance" of the series LRC circuit is called the ***impedance*** and given the symbol Z (units of ohms). Z determines the current amplitude from the voltage according to:

$$I_0 = \frac{V_0}{Z} \quad \Leftrightarrow \quad V_0 = I_0 Z \tag{13.16}$$

Z itself can be determined from the following (phasor) triangle:

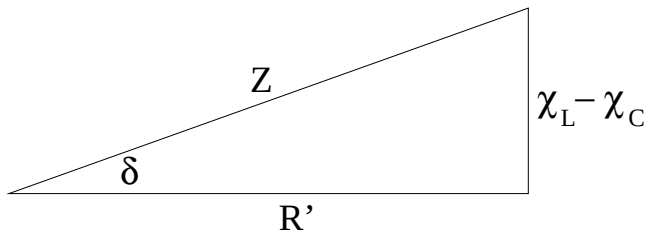


Figure 13.7: The impedance diagram for the *LRC* circuit, where $R' = r + R$ is the *total* resistance in the circuit.

From this triangle we can easily see that:

$$Z = \sqrt{R'^2 + (\chi_L - \chi_C)^2} \tag{13.17}$$

It also defines the phase angle:

$$\delta = \tan^{-1}\left(\frac{\chi_L - \chi_C}{R'}\right) \tag{13.18}$$

- **Power in the Series LRC Circuit:** The total ***average power*** delivered to the ***load resistance R*** (only) as a function of the driving frequency (and all of the other given parameters V_0, L, r, R, and C) can be written most compactly in terms of ***dimensionless parameters*** as:

$$P_{R,\text{avg}}(\omega) = P_{R,\max}\frac{1}{\left\{1 + \mathcal{Q}^2\left(\beta - \frac{1}{\beta}\right)^2\right\}} \tag{13.19}$$

where:

$$P_{R,\max} = \frac{1}{2}\frac{V_0^2}{(r+R)^2}R \qquad \beta = \frac{\omega}{\omega_0} \qquad \mathcal{Q} = \sqrt{\frac{R^2C}{L}} = \sqrt{\frac{\tau_{RC}}{\tau_{RL}}} \qquad \omega_0 = \frac{1}{\sqrt{LC}}$$

The power "wasted" inside the power supply has exactly the same form, using r instead of R in the maximum power:

$$P_{r,\max} = \frac{1}{2}\frac{V_0^2}{(r+R)^2}r$$

Because the average power provided by the power supply is proportionally split between the resistances, this circuit is sometimes referred to as a *voltage divider*.

- **Graphing the Resonant Power Curve:** Note that we just wrote the power delivered to

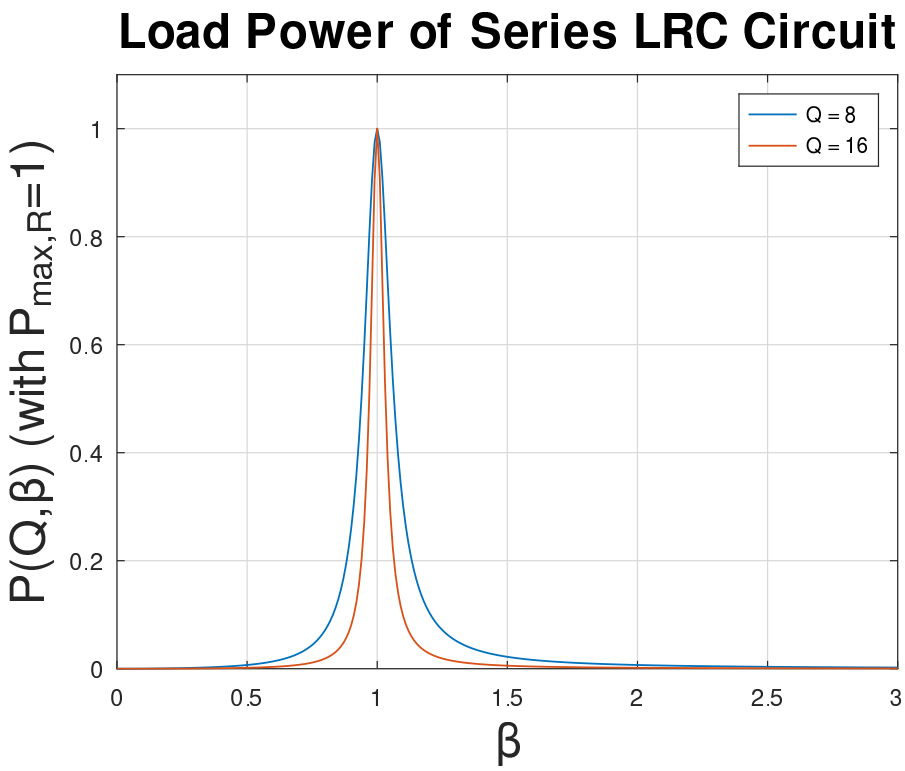


Figure 13.8: A graph of the dimensionless "filter function" for the two values $\mathcal{Q} = 8, 16$. This curve is exact (for $P_{\max,R} = 1$)! One can obviously rescale the abscissa to make the location of '1' any value ω_0 desired and rescale the ordinate to correspond to any $P_{R,\max}$ desired.

the load as a ***universal dimensionless filter function*** times the ***peak average power*** the circuit can deliver, which occurs when:

$$\beta = \frac{1}{\beta} \quad \Rightarrow \quad \omega = \omega_0$$

(a condition called ***resonance***, just as it was in the related discussion of drive harmonic oscillators in the first (mechanics) textbook in this series). It can be used to very simply graph *any* resonance curve just by rescaling the axes (by $PR, \max$ and ω_0, respectively and "squeezing" the graph according to:

$$\mathcal{Q} = \frac{\omega_0}{\Delta\omega} = \sqrt{\frac{L}{R^2C}} \qquad \Leftrightarrow \qquad \Delta\omega = \frac{\omega_0}{\mathcal{Q}} \tag{13.20}$$

$\Delta\omega$ is called ***the full width at half maximum***. Increasing $\mathcal{Q}$ (only) makes the graphed peak *sharper* as $\Delta\omega$ gets *smaller*.

- **Current Amplitude, Impedance, and Phase in a Parallel LRC Circuit:** The "solution"

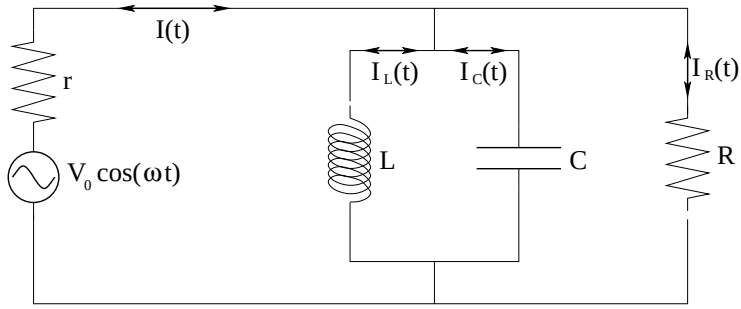


Figure 13.9: A parallel LRC circuit, for a non-ideal power supply with internal resistance r and loar resistance R. $R' = r||R = rR/(r+R)$ is the total *parallel* resistance of the circuit and is a critical parameter in its analysis.

to the parallel RLC circuit problem is a lot more difficult than it is for the series circuit, so much so that we only really properly solve it in the (advanced) section on solving circuit problems with complex algebra. Here we just summarize the results. The current is given by:

$$I(t) = I_0 \cos(\omega t - \delta) \tag{13.21}$$

where:

$$I_0 = \frac{V_0}{R}\frac{Z}{r} \qquad Z = \frac{R'}{\sqrt{1 + \mathcal{Q}^2\left(\frac{1}{\beta} - \beta\right)^2}} \qquad \beta = \frac{\omega}{\omega_0} \qquad \mathcal{Q} = \frac{L\omega_0}{R} \qquad \omega_0 = \frac{1}{\sqrt{LC}}$$

and:

$$\delta = \tan^{-1}\mathcal{Q}\left(\frac{1}{\beta} - \beta\right)$$

(note that this has the *opposite sign* of the δ defined for the series circuit above. Note well that:

$$R' = r \parallel R = \frac{rR}{r+R}$$

- **Power in the Parallel LRC Circuit:** The total ***average power*** delivered to the ***load resistance R*** (only) as a function of the driving frequency (and all of the other given parameters V_0, L, r, R, and C) can be written most compactly in terms of ***dimensionless parameters*** as:

$$P_{R,\text{avg}}(\omega) = P_{R,\max}\frac{1}{\left\{1 + \mathcal{Q}^2\left(\frac{1}{\beta} - \beta\right)^2\right\}} \tag{13.22}$$

where:

$$P_{R,\max} = \frac{1}{2}\frac{V_0^2 Z^2}{R^2 r^2}R = \frac{1}{2}\frac{V_0^2}{(r+R)^2}R$$

This is *exactly the same function of β and $\mathcal{Q}$ that was obtained for the series LRC circuit*, and at resonance the total average power provided by the power supply is still proportionally divided between r and R.

However, its behavior *away* from resonance is enormously different. The power delivered by the power supply is at a *minimum* in resonance (even through the power delivered ***to the load*** is ***maximum***). Same load power curve, but very different optimal applications, this is a popular circuit for e.g. crystal radio tuners but not so much for high power applications unless needed to match power supply impedances.

The graph of the power is obviously the same, but the role of capacitor and inductor as high/low pass filter elements (discussed next) is reversed (note $\beta \leftrightarrow \beta^{-1}$ in the power, but the term is squared so this has no effect) as they *short out the load* at high or low frequencies, rather than block the total current as they do for a series LRC circuit. For that reason it will not be redrawn here.

13.1: Introduction: Alternating Voltage

As we have seen in the previous chapter, if one spins a coil with N turns and cross-sectional area A at angular velocity ω in a uniform magnetic field B oriented so that it passes straight through the coil at one point in its rotation, one generates an *alternating voltage* according to:

$$\phi_m = \vec{\boldsymbol{B}} \cdot NA\hat{\boldsymbol{n}} = NBA\cos(\omega t) \tag{13.23}$$

$$V(t) = -\frac{d\phi_m}{dt} = NBA\omega\sin(\omega t) \tag{13.24}$$

This is, in fact, the functional form of the voltage that comes out of wall receptacles in your house, no matter what the voltage or frequency used by your particular country of residence. It is also the general functional form of electrical signals generated by many other means in (for example) radio transmitters.

In this chapter, then, we will learn to treat "arbitrary" harmonic alternating voltage sources as having the form:

$$V(t) = V_0\sin(\omega t) \tag{13.25}$$

where of course we can introduce an arbitrary phase (corresponding to the choice of when we start our clock). In this expression, remember that:

$$\omega = 2\pi f = \frac{2\pi}{T} \tag{13.26}$$

where f is the *frequency* of the harmonic oscillation in units of *Hertz* (cycles per second) and T is the corresponding *period*.

We will also look at slightly more general voltage sources that are *nearly* harmonic, in particular *amplitude modulated* harmonic sources such as:

$$V(t) = A(t)\sin(\omega t) \tag{13.27}$$

where $A(t)$ is a *slowly varying function of time* (making only small changes over many periods T of the harmonic part). More advanced students should note well that we will not *properly* treat this problem by means of e.g. a Fourier Transform, as knowledge of Fourier Transforms (however useful!) is not a requirement for this course. We will *barely* explore some of the benefits of treating voltages or currents given in a complex form:

$$V(t) = V_0 e^{i\omega t} \tag{13.28}$$

where V_0 may be a general complex number, $V_0 = |V_0|e^{i\delta}$ but again, advanced students should keep in mind the fact that this often makes things much *easier* once one has paid the price of learning how to use algebra over the field of complex numbers plus a few things such as Cauchy's theorem and Fourier Transforms. Some ideas, such as the importance of having *enough* bandwidth to encode an amplitude modulated (or otherwise encoded) signal on top of a given carrier frequency while nevertheless remaining well resolved from nearby carriers carrying information on other channels are very difficult to *prove* without using this more advance math, so students will have to content themselves with a few of this book's rare it-is-so-because-I-say-so without proper derivation or justification.

One very important thing all students should learn from this chapter is just how alternativing voltages and high-voltage transmission lines, together, are nothing less than the *basis for modern civilization* – a country's productive capacity and the comfort of its citizens is *directly linked* to its ability to generate electrical energy and distribute it widely in a cost-effective way.

Nothing convinces one more of this than the not-terribly-infrequent instances of *power outages* when hurricanes, ice storms, earthquakes, or solar storms interrupt the power grid for days or even weeks of time. During the downtime one immediately loses all refrigeration (so stored food spoils), heating and cooling (so one has to survive at the ambient temperature as best one can), the ability to turn light on and off with the touch of a finger (so one can stay up later and get up earlier than the sun), the ability to drive safely (no traffic lights), the ability to bank or shop indoors in shopping malls (no air conditioning, lights, electronic cash registers, check card readers), the ability to listen to music, compute, browse the internet (once local battery stores are exhausted). Over a single week life devolves to what it was like over a century ago before the advent of universally accessible, inexpensive electricity.

Life over a century ago, without electricity, *sucked!*

13.1.1: Electrical Distribution True Facts

The most common models for household electrical distribution are represented in the following table (note well that $\omega = 2\pi f$ where f is the frequency of the source in Hertz): 208 volts is the

Volts	Hz	Purpose	Continent
120	60	lighting, small appliances, electronics	N. and S. America
208 or 240	60	heating, cooling, large appliances, 3 phase motors	N. and S. America
230	50	all household use	Everywhere else

Table 11: Common alternating voltages and frequencies in use around the world. There is a dazzling array of plug types in use around the world as well.

potential difference between any two phases of a three-phase "Wye" main supply in the US

where the pole voltage are 120 volts relative to ground:

$$\begin{aligned} V &= 120\sin(\omega t) - 120\sin(\omega t \pm 2\pi/3) \\ &= 240\sin(\pi/3)\sin(\omega t \pm \pi/6) \\ &= 207.8\sin(\omega t \pm \pi/6) \end{aligned} \tag{13.29}$$

and 240 is similarly the difference between two 120 volt lines that are completely out of phase. Do *not* use this table as an authoritative guide to electrical main supplies around the world; there are many such authoritative guides and tables available on the internet[188] .

It is worth mentioning that (unfortunately) 60 Hz is a *particularly unfortunate* choice for distribution frequency because it is in "resonance" with certain cardiac frequencies and hence unusually likely to defibrillate the human heart. As little as 10 mA of 60 Hz AC across the heart can kill a person. It requires roughly five times as much DC (50 mA) to be equivalently dangerous!

As you can see, most power is distributed at only 50 or 60 Hz. This leads us to several important questions. Why distribute alternating voltage at all? Why use the particular frequencies that we use to alternate with, instead of (say) much higher frequencies or much lower ones (all the way down to DC voltage).

The reason we use alternating voltage is because it makes it easy to increase or decrease the voltage using *transformers*. In a moment we'll cover transformers and the reasons for using them in detail, but in a nutshell for now, we need to transmit the energy from the power station to where it is used at as high a *voltage* as possible. Transformers work "better" at higher frequencies than at lower frequencies, as they use induction; we need at least a *minimal* frequency in the tens of Hz to permit them to work at all well, but they'd work fine at 100's or 1000's of Hz too.

However, we cannot use these higher frequencies – in spite of the fact that they'd be much safer biologically because alternating *current* (AC) does not flow *uniformly* through a (cylindrical) conductor – most of the current flows near the *outer surface* of a conductor, and the current density drops of *exponentially* as one procedes further in with an exponential decay length δ_s called the *skin depth*. At 60 Hz this length is roughly 8.5 mm in copper; copper conductors "an inch in diameter" have at least some current density throughout their cross-section. At 10 kHz (an arguably safer frequency) it is 0.66 mm in copper, and an inch-thick cable carries *no* significant current over *most* of its cross-section.

If a wire is much thicker than the skin depth, its resistance is *significantly increased* because the effective cross-section in the

$$R = \frac{\rho L}{A} \tag{13.30}$$

expression isn't e.g. $A \approx \pi R^2$, it is roughly $A \approx 2\pi R\delta_s$ for $\delta_s \ll R$ (a much smaller number). 50 or 60 Hz are thus *compromises* between the need to use AC to transmit energy long distances and the need to minimize the resistance of the transmission wires along the way by making effective use of their entire cross-sectional areas, for cable cross-section diameter assumed to be an inch or less. Cables thicker than this are sometimes fabricated so that they are *hollow*, since there is little current carried by the central core anyway.

[188] Wikipedia: http://www.wikipedia.org/wiki/Mains_electricity. See also the many links in this article.

It is no exaggeration to state that alternating voltage generated using Faraday's Law and transmitted at high alternating voltages before being stepped down and used at lower voltages is the fundamental basis for modern civilization. Power distributed over long distances using step-up and step-down transformers has created the highest global standard of living in human history. Some 2/3 of the world's population uses nearly ubiquitous electricity to light, heat and cool their homes, to refrigerate and cook their food, to fuel devices that provide increasingly universal access to *information* in many of its sensory forms – musical, textual, visual, to provide transportation, to fuel industry and commerce and agriculture. If the electrical grid for any reason ceased to function we would regress to a medieval existence in a matter of weeks (as I have personally experienced as both hurricanes and ice storms have caused weeklong power outages in North Carolina on more than one occasion).

Let us understand the transformer and the role that it plays in the transmission of power.

13.1.2: The Transformer

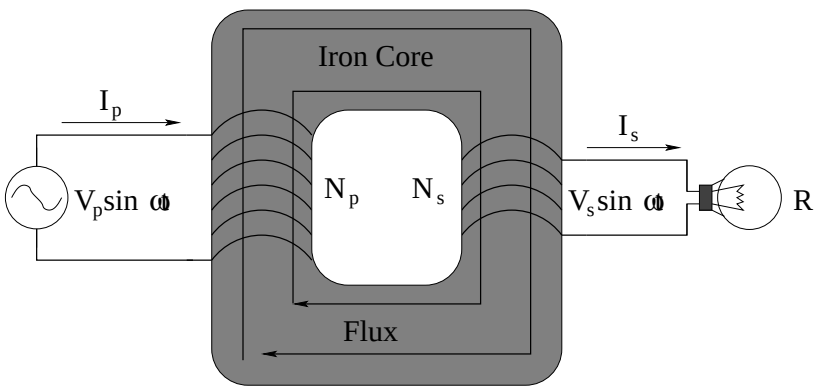


Figure 13.10: A transformer transforms voltage V_1 into a new voltage V_2, for time-varying (usually sinusoidal) voltages only.

The transformer is basically a pair of flux-coupled coils, one (the *primary*) with N_p turns connected to the *source* of alternating voltage, the other (the *secondary*) with N_s turns connected to the *load* that actually consumes the energy delivered from the source. All of the flux that passes through any turn in the primary or secondary coils passes (with as little loss as it is possible to arrange) through all of the turns in both coils. The flux is usually coupled by wrapping the coils around e.g. a torus of soft iron that traps flux, laminated to prevent *eddy currents* (called the transformer *core*).

If we let ϕ_m be the flux trapped in the core that passes through a single turn, then:

$$V_s = N_s \frac{d\phi_m}{dt} \tag{13.31}$$

$$V_p = N_p \frac{d\phi_m}{dt} \tag{13.32}$$

or (taking the ratios of these two equations, in order)

$$\frac{V_s}{V_p} = \frac{N_s}{N_p} \tag{13.33}$$

Note that we omit Lenz's law in this expression because we can wrap either coil either way around the core so that the voltages on primary or secondary side can be "in phase" or "exactly out of phase" as we wish.

A transformer can thus *step voltage up* to higher levels or *step it down* to lower ones, depending on whether $N_p < N_s$ or vice versa. This seems as though it would be obviously useful for many, many things, and of course it is. Sometimes we need a high voltage and a low current in a wire; other times we need a low voltage and a high current. Note well that we can't magically get a higher voltage and *more* current out of a transformer as this would violate energy conservation. In fact, if we compute the power delivered by the primary voltage to the transformer and equate it to the power consumed by the secondary circuit, then as long as the transformer itself doesn't get hot (removing energy from the circuit of its own accord):

$$P_p = V_p I_p = V_s I_s = P_s \tag{13.34}$$

or, if we use the fact that $V_s = V_p N_s / N_p$ and divide a couple of times, we find that:

$$I_s = \frac{N_p}{N_s} I_p \tag{13.35}$$

When the voltage goes up ($N_s > N_p$) the current goes down, and vice-versa.

Of course this *does* assume that the transformer itself and all of its wiring doesn't have any resistance and get hot, and the iron core of the transformer must *also* not get hot. However, the iron core is *itself* a conductor. When the magnetic flux through it is constantly changing it induces a voltage in *it* that causes a current to flow. That current, flowing in the resistance of the iron, generates heat! This kind of inductive heating is said to be caused by *eddy currents*, currents induced in any conductor by rapidly changing magnetic flux through the conductor.

It is also clearly undesirable, as the heat that appears in the iron core is *lost* and hence reduces the available power (voltage and current alike) on the secondary compared to what comes in through the primary. To minimize eddy currents, the iron core is usually made of *laminated* strips of iron separated by insulating resin or out of insulated *wires* of iron. The small cross-sectional area of the individual conductors thus minimizes flux, voltage and current, and thereby losses to heating through eddy currents.

Now, high voltage is dangerous. Dielectric breakdown can easily occur of the voltage is high enough – power can simply leap through the air in an electrical arc and fry whatever it passes through on its way to ground. Nevertheless, we find it very useful to use high voltage to *transmit electrical power long distances* by using the fact that current goes *down* as the voltage goes *up* for any given power being delivered.

13.1.3: Power Transmission

When electricity was first introduced into society on a grand scale (largely by Thomas Edison, to use in his recently invented light bulbs) Edison wished to power the world with direct current (DC) lines from his generating stations directly into your home, at a very low (and thereby safe) voltage. Edison had a number of patents on various aspects of DC power generation, storage, and metering, and had a vested interest in all of this technology. However, Edison was no mathematician, and did not *understand* electricity or Maxwell's equations (indeed, at

the time Maxwell's equations were only about 20 years old and there weren't a lot of people who weren't mathematicians or physicists who *did* understand them).

There is just one problem. At *low* voltages, delivering power across miles of wire to households can easily be shown to waste *almost all* of that energy heating the wires that carry it, and leave *almost none* for the households at the end! Edison's solution required there to be a DC generating plant within a mile, at most, of every household that received its energy, and required massive amounts of copper even then for its transmission lines.

At the same time, a George Westinghouse had hired a young man named William Stanley Jr[189]) to work on implementing an *alternating* current distribution system using AC transformers, which had just been invented. Stanley (working with a few others) in 1886 built a working AC distribution system that distributed the electricity over long distances with low currents at very high voltages in Great Barrington (basically, the neighborhood in which he lived). This system required much, much thinner wires and could distribute the energy over long distances but *also* required that the voltage be stepped down to a relatively "safe" voltage inside the homes and businesses where it was going to be used. Westinghouse almost immediately began to sell and build AC distribution systems based on Stanley's design for US cities that were eager to reap the benefits of electricity for its citizens.

Westinghouse also acquired at roughly the same time the patent(s) of a young man named Nikola Tesla[190] on polyphase generators and motors that would run on AC voltage, basically going "all in" on AC distribution. This led to the so-called *War of the Currents*[191] between the two companies – Edison's General Electric and Westinghouse's Westinghouse (both of which survive as supergiant corporations to this day).

Edison lost. We absolutely need to learn, and understand, *why* as it is of paramount importance *today*, some 130 years later, as we struggle to convert to renewable resource electrical generation, conversion of e.g. sunlight or the power of the wind in suitable locations into electrical current and its transmission across *thousands to as many as ten thousand miles* from those locations to where it will be consumed (say, from the Sahara desert to Finland, or India to Siberia).

So here's the trick of the power grid, Stanley's solution. The resistance of a wire is (recall) $R = \frac{\rho L}{A}$ (where A is the effective cross section at a given frequency). A copper wire just under a quarter inch thick has a resistance of roughly 1 Ohm/mile (rule of thumb). A wire a third of an inch thick has a resistance of roughly 0.1 Ohms/mile. Wires this thick are heavy and expensive and have to carry a *lot of energy*. Now, suppose we have a power station a mere ten miles from your home. The total resistance of all the wires between that power station and your

[189]Wikipedia: http://www.wikipedia.org/wiki/William Stanley Jr. William Stanley, incidentally, is also the inventor of the "Stanley stainless steel thermos". Interestingly, General Electric eventually bought out a controlling interest in Stanley's own electrical research and manufacturing company.

[190]Wikipedia: http://www.wikipedia.org/wiki/Nikola Tesla. Tesla was the original "mad scientist" – he is the original inventor of the radio (and was cheated of the patent), he invented his own ruinously inefficient and short range electrical distribution system based on the Tesla coil, invented the X-ray tube and photographed the bones of his own hand before Roentgen (but failed to publish or patent and lost the technical descriptions in a fatal fire that destroyed much of his work prematurely), he purportedly invented a "death ray", but destroyed it after a single apocryphal demonstration of its effects. He had a photographic memory and reportedly experienced direct insight into problems he was working on, bypassing all normal routes to invention or design. He is basically an enormously interesting person I a strongly recommend reading at least the wikipedia article on him.

[191]Wikipedia: http://www.wikipedia.org/wiki/War of Currents. Again, a worthwhile read.

home is easily order of an ohm. Now imagine that you turn on a single 100 Watt bulb (drawing roughly 1 A in current. The power station must provide 101 Watts for your bulb to burn – 100 Watts used by the bulb and $I^2R \approx 1$ Watt used in the *supply line*.

However, you then turn on the *rest* of your lights, your refrigerator kicks on, your AC starts up. Your house is now drawing more like 100 Amperes (delivered in parallel to the many appliances) and is using order of 10000 Watts. *So is the supply line!* Half of the energy being delivered to your home is wasted as heat along the way. A second consequence is that the *voltage* at your house is reduced to a fraction of the nominal voltage as you turn on more appliances and more of the voltage drop occurs across the supply resistance!

The solution is to *transmit at high voltage and low current* and *use at low voltage and high current*. If we step up the voltage by (say) 10,000 Volts (real long distance transmission is at much higher voltages than this) then in order to deliver the same *power* at the far end, instead of delivering 100 Amps at 100 volts one can deliver 1 Amp at 10,000 Volts! The resistive heating of the supply line is back to 1 Watt out of 10,000 delivered. Here the square in I^2R becomes your *friend* – delivering 10 kW at 100,000 V requires only 0.1 A and uses only 0.01 W heating the wire.

This is good for transmission, but bad for utilization. 100,000 volts can arc an appreciable distance through even *dry* air; that's why the insulators on high voltage transmission towers are so long! We'd hate to get electrocuted every time we changed a light bulb as power arced out of the socket through our bodies on the way to ground. With an entire power plant delivering the energy, even the (mere) 16,000 volt lines that run down the streets can literally make your body explode if you should stray within a few cm of a supply line.

In one of the few instances in my memory of a power outage at Duke, a squirrel was crispy-fried when it got inside the barbed wire fences at a major step-down transformer serving part of the campus. It strayed too near to the main power buses, which arced over (through the squirrel) blowing the transformer and shutting down power to the campus for a time. Imagine how exciting life would be if every time you went to plug in an electric light into your 16,000 volt household wiring or flicked a switch on a humid day, you risked being electrocuted by what amounts to a manmade lightning bolt!

"Exciting" isn't quite the right word for it. Consequently, there is *always* a step-down transformer at the very end of the line, that drops the voltage in our houses to the *much* safer but still dangerous 120 volts (relative to ground). Why such a high voltage? Wouldn't (say) 12 volts be even safer? Sure, but we still have to transmit the energy around *inside* the house as well, and there is a reason car jumper cables are made of much thicker wires than (say) a lamp cord!

Even inside the house we use devices that use anywhere from a watt or two up to almost 2000 watts for devices plugged into an ordinary receptacle. At 120 volts, these devices draw currents as high as 15 to 20 Amps (per circuit) within the house before circuit breakers or fuses interrupt the circuit. This is a low enough current that the resistive heating of the order of 10-50 meter long *household* supply lines remains "low", specifically low enough that the wires don't melt their insulation and/or set your house on fire if there is a short circuit momentarily drawing a much higher load!

Even this "low" resistance can waste a lot of heat! 14 gauge copper wire has a resistance

of around 0.25 Ohms per 100 feet of wire, wasting around 56 watts heating the wire all along its length when one draws the maximum National Electrical Code (NEC) permitted 15 amps of current. It also reduces the line voltage available to the appliance(s) at the end that are drawing all of that current by 3-4%) – there is a reason people refer to it as "110" volt household wiring when the voltage produced at the outdoor transformer is carefully regulated at 120 volts – the *actual* voltage at your appliance could be as much as 10lower or around 110 volts when a heavy load is turned on at the end of a long (50 meter) wiring run of 14 gauge wire!

Personally, while the NEC *does* permits one to use 14 gauge wire for wiring of normal "short" 15 amp household circuits, I prefer to do any primary runs in household wiring with the thicker 12 gauge wire (and not to use the thinner 14 gauge wire *at all* if I can help it) to minimize heat loss in the household wiring. The thicker wire is more expensive, but, as you can see, with thinner wiring you can easily waste anywhere from 1% to 5% of your energy bill simply heating the space inside your walls when you run appliances (and then paying again to air-condition that heat out of your house), month after month over decades! It's also safer! Even though the wire doesn't produce a *lot* of heat per foot, if it is running next to other wires in an insulated space, that heat can build up and significantly raise the temperature! Thicker supply lines run cooler at any given load.

All of this – and the National Electric Code itself – will make sense when you work out the algebra for yourself and apply the physics we have learned in this text so far, although there is a lot more to it than we have time to cover here – one can take entire courses in electronics. One of the homework problems has you do this very thing – explore the electrical distribution system quantitatively for an example mini-system. Be sure that you work through it, with the help of your instructor as necessary.

So much for the generation and efficient transmission of power, which we can see relies very much on AC currents and generators. Next we move on to the use of alternating voltages of *much higher frequency*, frequencies that we can associated with radio waves and information processing. The electrical circuits that allow us to generate, transmit, receive, encode and decode information in alternating flows of current are very nearly as important to modern society as the direct delivery of electrical power in the first place. They are also useful in the laboratory, and are key components of much medical apparatus, information technology apparatus, entertainment apparatus – they are ubiquitous, in other words. We begin by seeing how simple arrangements of resistances and inductances can *oscillate* in a way that is *mathematically identical* to the way a mass on a spring oscillates.

13.2: Passive AC Circuits

To make this section as simple as possible, we begin by noting that in the context of Kirchoff's rules and electrical circuits, a capacitor plays *precisely* the same role as a spring does in mechanics – it stores electrical charge and energy with a restoring "force" proportional to the charge. A resistance behaves *exactly* like a linear drag force does on the mechanical movement of the stored charge. An inductance behaves *exactly* like a mass does in a spring-driven harmonic oscillator, as a reservoir for the "kinetic" energy associated with flowing charge and the "momentum" that causes that charge to tend to continue flowing unless acted on by

opposing forces. Finally, a harmonically alternating voltage behaves *exactly* like a harmonically altering driving force in the damped, driven harmonic oscillator.

One can also build a circuit made entirely out of *water-filled pipes* that precisely mimics an electrical circuit. A section of the pipe containing a spring loaded piston that can store water on one side against the pressure difference maintained by the spring is a "capacitor". A sand-filled pipe that resists the flow of water is a "resistor". The water itself, which is massive and hence continues to flow in the (frictionless) pipe until slowed down by resistances or pressure differences is an "inductor". Finally, a pump that creates a harmonically oscillating pressure difference in the water, e.g. a harmonically driven pistor in a pipe, is just like an "alternating voltage".

Keep this in mind as we develop the following. Even though of course the algebra will be specific to the particular circuits being studied, the results will be *analogous* to identical results that arise from solving identical equations in other contexts you have already explored in mechanics. This conceptual repitition can help you learn the material more easily, and help you remember it for longer without additional reinforcement, provided (of course) that you properly studied harmonic oscillators the *first* time you encountered them.

13.2.1: Non-driven LC circuit

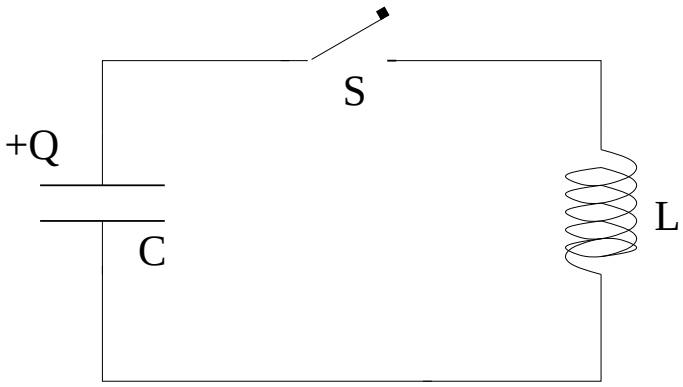


Figure 13.11: Undriven LC circuit

In figure 13.11, the capacitor C on the left is initially charged up to charge Q_0. At time $t = 0$ the switch is closed and current begins to flow. If we apply Kirchhoff's voltage/loop rule to the circuit, we get:

$$\frac{Q}{C} - L\frac{dI}{dt} = 0 \tag{13.36}$$

where

$$I = -\frac{dQ}{dt} \tag{13.37}$$

If we substitute this relation in for the I's and divide by L, we get the following second order, linear, homogeneous ordinary differential equation:

$$\frac{d^2Q}{dt^2} + \frac{Q}{LC} = 0 \tag{13.38}$$

We recognize this as the differential equation for a *harmonic oscillator!* To solve it, we "guess"[192]:

$$Q(t) = Q_0 e^{\alpha t} \tag{13.39}$$

and substitute this into the ODE to get the characteristic:

$$\alpha^2 + \frac{1}{LC} = 0 \tag{13.40}$$

We solve for:

$$\alpha = \pm i\sqrt{\frac{1}{LC}} = \pm i\omega_0 \tag{13.41}$$

and get:

$$Q(t) = Q_{0+}e^{+i\omega_0 t} + Q_{0-}e^{-i\omega_0 t} \tag{13.42}$$

or (taking the real part and using the initial conditions):

$$Q(t) = Q_0 \cos(\omega_0 t) \tag{13.43}$$

Note well that this overall solution methodology is *identical* to that used for the simple harmonic oscillator, with spring constant $k_{\text{eff}} = \frac{1}{C}$ and mass $m = L$.

One can, of course, analyze energy in this circuit. At any instant of time, the energy in the circuit is clearly all the energy stored in the capacitor:

$$U_C(t) = \frac{Q(t)^2}{2C} \tag{13.44}$$

This energy *over time* oscillates between the capacitor and the energy in the inductor:

$$U_L(t) = \frac{1}{2}LI(t)^2 \tag{13.45}$$

Show that the sum of these two energies is a constant, and that the constant equals the initial energy in the capacitor! This is precisely analogous to what happens to the conserved total energy as it oscillates between potential energy in a spring and kinetic energy of motion of the mass in a harmonic oscillator.

13.2.2: Non-driven LRC circuit

In figure 13.12, the capacitor C on the left is initially charged up to charge Q_0. At time $t = 0$ the switch is closed and current begins to flow. If we apply Kirchhoff's voltage/loop rule to the circuit, we get:

$$\frac{Q}{C} - L\frac{dI}{dt} - IR = 0 \tag{13.46}$$

where

$$I = -\frac{dQ}{dt} \tag{13.47}$$

If we substitute this relation in for the I's and divide by L, we get the following second order, linear, homogeneous ordinary differential equation:

$$\frac{d^2Q}{dt^2} + \frac{R}{L}\frac{dQ}{dt} + \frac{Q}{LC} = 0 \tag{13.48}$$

[192]Not really.

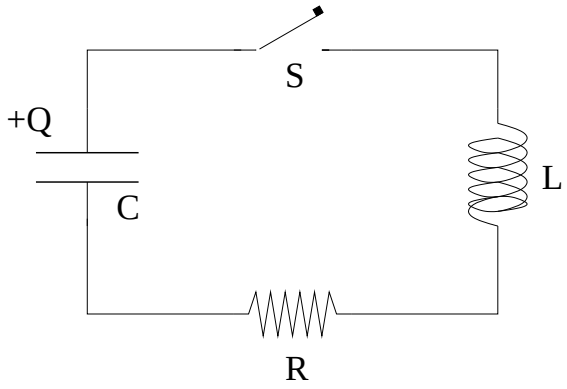


Figure 13.12: Undriven LRC circuit

We recognize this as the differential equation for a *damped harmonic oscillator*. To solve it, we "guess"[193]:

$$Q(t) = \hat{Q}e^{\alpha t} \tag{13.49}$$

where $\hat{Q}$ is some unknown – possibly complex – constant and substitute this into the ODE to get the ***characteristic***:

$$\alpha^2 + \frac{R}{L}\alpha + \frac{1}{LC} = 0 \tag{13.50}$$

We solve for:

$$\begin{aligned} \alpha &= -\frac{R}{2L} \pm \frac{\sqrt{\left(\frac{R}{L}\right)^2 - \frac{4}{LC}}}{2} \\ &= -\frac{R}{2L} \pm i\omega_0\sqrt{1 - \frac{R^2C}{4L}} \\ &= -\frac{R}{2L} \pm i\omega_0\sqrt{1 - \frac{\tau_L}{4\tau_R}} \\ &= -\frac{R}{2L} \pm i\omega' \end{aligned} \tag{13.51}$$

where $\tau_L = R/L$ $\tau_C = 1/RC$, $\omega' =_0 \sqrt{1 - \frac{\tau_L}{4\tau_R}}$. As before (in both the previous section and in mechanics) we have ***two complex exponential*** soutions so we might as well let the unknown constant (for each) be complex as well:

$$\hat{Q}_\pm = |\hat{Q}_p m|e^{i\phi_p m}$$

Then e.g.

$$Q_+(t) = |\hat{Q}_+|e^{i\phi_+}e^{-\frac{Rt}{2L}}e^{i(\omega' t)} = Q_+e^{-\frac{Rt}{2L}}e^{i(\omega' t + \phi_+)} \tag{13.52}$$

where Q_+ is the *real* amplitude of $\hat{Q}_+$ (and a very similar equation for the $-i\omega'$ solution).

Of course, we don't know what *complex* charge Q or *imaginary* charge Q could possibly be, so we must take the *real part of this* as our final solution:

$$Q(t) = \Re e\left\{Q_+e^{-\frac{Rt}{2L}}e^{i(\omega' t + \phi_+)}\right\} = Q_+e^{-\frac{Rt}{2L}}\cos(\omega' t + \phi_+) \tag{13.53}$$

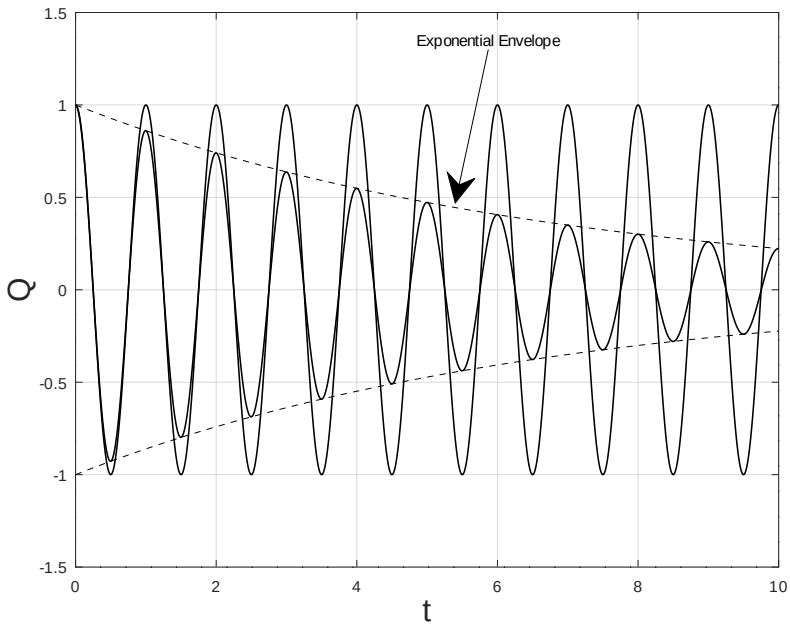


Figure 13.13: A *weakly* damped series LRC-circuit oscillates within an exponential damping envelope. $R/2L = 0.15$ and $T_0 \approx T' = 1$ in the time scale of this figure.

where Q_+ and ϕ are both real *constants of integration* that must be set from the initial conditions.

For the specific "standard" initial conditions given above, $Q_+ = Q_0$, $\phi_+ = 0$ and we get:

$$\boxed{Q(t) = Q_0 e^{-\frac{Rt}{2L}} \cos(\omega' t)} \tag{13.54}$$

This solution is plotted in figure 13.13 for the ***underdamped*** case (see below), where

$$\frac{RT_0}{2L} = 0.15 \approx \frac{1}{7} < 2\pi$$

so that the exponential damping time of the amplitude is roughly seven times the period $T_0 = 1$ on the scale displayed. Note that this is small *enough* that little deviation is observed between the damped and undamped solution over ten cycles of oscillation – $T' \approx T_0$.

Note that if we used $\hat{Q}_-$ and $-i\omega'$ to develop a real solution, its real part would have ***exactly the same form*** and hence would be identical to this once *its* constants of integration were set from the same initial conditions, so we don't even need to write it down, let alone keep it (or the general complex solution) around.

From this completely general solution for $Q(t)$, we can easily find the current through and voltage across all of the elements of the circuit. Given the current and these voltages it is easy to show that energy is conserved (it could hardly not be, given that we started with KLR) and that as we chase energy down we will find that the initial energy stored in the capacitor exactly balances the energy consumed in the resistor as $t \to \infty$. This is left as an exercise – the more of this that you work out on your own (rederiving things at least once as part of the process) the more you will learn.

[193]Not really.

Clearly the analogy with ordinary simple harmonic oscillators with linear damping is beyond strong – it is algebraically exact. We therefore must expect that series LRC circuits will *also* exhibit underdamped oscillation, critically damped exponential decay, and overdamped even slower expoential decay (just as the mass on the spring did), and that one can *drive* the circuit in *resonance* exactly the same way as well.

Let's look briefly at these limits:

Underdamped Oscillation – $\left(\boldsymbol{\frac{R^2C}{4L} < 1}\right)$**:** As always, when the system is underdamped it will oscillate within a decaying amplitude envelope as illustrated above.

It is worth looking at this condition a bit more closely:

$$\frac{R^2C}{4L} < 1 \quad \Rightarrow \quad \frac{R^2}{4L^2} \times LC < 1 \quad \Leftrightarrow \quad \frac{R}{2L} < \omega_0 = \frac{2\pi}{T_0} \quad \text{or} \quad \frac{RT_0}{2L} < 2\pi$$

We will consider ***weak damping*** to be the *specific* limit where:

$$\frac{R}{2L} \ll \frac{2\pi}{T_0} \quad \text{or} \quad \boxed{\boldsymbol{\frac{RT_0}{2L} \ll 2\pi}} \tag{13.55}$$

In this limit, $\frac{R^2C}{4L} \ll 1$ and:

$$\omega' \approx \omega_0, \quad T' \approx T_0 \tag{13.56}$$

(we can basically ignore the frequency shift to lowest order). We will now express the other two (less important) conditions in terms of ω_0 and the

Critically Damped Oscillation – $\left(\boldsymbol{\frac{RT_0}{2L} = 2\pi}\right)$ **:** As usual, when the LRC circuit is critically damped, the charge on the capacitor exponentially approaches zero at the *maximum* rate (but because it is an exponential, it never quite gets there).

Overdamped Oscillation – $\left(\boldsymbol{\frac{RT_0}{2L} > 2\pi}\right)$ **:** Also as usual, if the LRC circuit is overdamped, it approaches zero like a *mix* of exponential decays, but the fastest of them is still slower than the critically damped approach.

13.3: Energy Loss in Passive LRC Circuits – $\mathcal{Q}$-Factor

In a short while we will tackle the daunting task of damped *driven, active* AC circuits, including the series LRC circuit driven by an inline AC voltage such as $V_0 \cos(\omega t)$. There we will find that we can best understand the power delivered to the circuit and used to "drive a resistive load" – transmit average power from the voltage source to the resistor, basically, where it turns up as heat – in terms of the so-called ***quality factor*** or **$\mathcal{Q}$-factor**[194] we studied in the context of damped undriven simple harmonic oscillators in the first (mechanics) half of this introductory series. Let's work this out for the passive series LRC circuit we just worked out above, as even in this simple case, $\mathcal{Q}$ is a useful parameter for describing the damping of a circuit.

[194]Note well that I'm trying to use a different font for $\mathcal{Q}$ than I do for the capacitor charge Q because in this context they otherwise would both use *exactly the same standard symbol*, which could be quite confusing in the algebra!

In mechanics, the $\mathcal{Q}$-factor was defined to be 2π *times the reciprocal of the fractional energy loss in a single period of oscillation*:

$$\mathcal{Q} = 2\pi \frac{E(0)}{E(0) - E(T')} = 2\pi \frac{E}{\Delta E} \tag{13.57}$$

where $T' = \frac{2\pi}{\omega'}$ is the *shifted* period of the damped oscillator. We will evaluate this in terms of the known parameters of our circuit in exactly the same way we did for a linearly damped mass on a spring, only now the energy stored in the LRC circuit is the sum of the field energies in the capacitor and the inductor – resistors don't store energy – instead of in the total potential plus kinetic energy of an oscillating mass.

Evaluating $\mathcal{Q}$ from *arbitrary* initial conditions would be tedious because we'd have to sum the stored energy in both the capacitor and the inductor and the *current* in the series LRC circuit is a a bit messy[195]. ***If***, however, we start the LRC oscillator with *zero* current and charge Q_0 on the capacitor at $t = 0$ (the initial conditions we used in the previous section) the algebra is ***easy***. In this case (which, don't worry, gives the exactly correct *general* result averaged over many cycles due to the wonderful properties of the exponential function):

$$Q(t) = Q_0 e^{-Rt/2L} \cos(\omega' t) \tag{13.58}$$

$$E(0) = \frac{1}{2}\frac{Q_0^2}{C} \tag{13.59}$$

$$E(T') = \frac{1}{2} e^{-RT'/L} \frac{Q_0^2}{C} \tag{13.60}$$

We can now easily form $\mathcal{Q}$:

$$\mathcal{Q} = 2\pi \frac{\frac{1}{2}\frac{Q_0^2}{C}}{\frac{1}{2}\frac{Q_0^2}{C} - \frac{1}{2}\frac{Q_0^2}{C} e^{-RT'/L}}$$

$$\mathcal{Q} = 2\pi \frac{\cancel{\frac{1}{2}\frac{Q_0^2}{C}}}{\cancel{\frac{1}{2}\frac{Q_0^2}{C}} - \cancel{\frac{1}{2}\frac{Q_0^2}{C}} e^{-RT'/L}}$$

$$\mathcal{Q} = \frac{2\pi}{1 - e^{-RT'/L}} \tag{13.61}$$

We now assume *weak damping*. Practically speaking, this means both that:

$$\omega' = \omega_0 \sqrt{1 - \frac{R^2 C}{4L}} \approx \omega_0 \quad \Rightarrow \quad T_0 \approx T' \tag{13.62}$$

and (equivalently) that

$$\frac{RT'}{L} \approx \frac{RT_0}{L} \ll 1$$

In the weak damping case it is clear that we can use a ***Taylor series expansion of the exponential*** in $\mathcal{Q}$:

$$e^{-RT'/L} = 1 - \frac{RT'}{L} + \mathcal{O}\left(\frac{R^2 T'^2}{L^2}\right) + ...$$

[195] We would have to use the product rule to take the time derivative of $Q(t)$ to form $I(t)$, and then we have to *square* this two-term result to make $\frac{1}{2}LI^2(t)$, which would then have *three* time-dependent terms with mixed trig function cross-terms and two distinct exponentials in them – ***Yuk!***

Substituting in this last form, cancelling the 1's and keeping only the leading order surviving term, we get:

$$\mathcal{Q} = \frac{2\pi}{\cancel{1} - (\cancel{1} - \frac{RT'}{L})}$$

or (rearranging):

$$\boxed{\mathcal{Q} = \frac{2\pi}{T'}\frac{L}{R} = \frac{L\omega'}{R} \approx \frac{L\omega_0}{R} = \sqrt{\frac{L}{R^2C}}} \tag{13.63}$$

Note that this can be substituted back into:

$$\omega' = \omega_0\sqrt{1 - \frac{R^2C}{4L}} = \omega_0\sqrt{1 - \frac{1}{4\mathcal{Q}^2}} \approx \omega_0\left(1 - \frac{1}{8\mathcal{Q}^2} + ...\right) \tag{13.64}$$

where I used a binomial expansion on the last bit. From this we see that all of the approximations above are *consistent* as long as:

$$\mathcal{Q}^2 \gg \frac{1}{8} \quad \Rightarrow \quad \mathcal{Q} \gg 0.35$$

In practice, then, even $\mathcal{Q} = 2$ makes $\mathcal{Q}^2 = 4 \gg 1/8$.

Mind you, when $\mathcal{Q}$ is less than around 5 you will have errors and asymmetries in some of its applications – for example, the power curve for *active* AC circuit such as the resonant driven series LRC circuit is reasonably symmetric for $\mathcal{Q} > 5$ but has an easily visible asymmetry and perhaps a 10 or 20% error in its interpretation in that context, although it is still quite useful all the way down to perhaps 2 or 3.

13.4: Active AC Circuits

To go on, we need to introduce a classical harmonic oscillating voltage like that produced by an AC generator and use it to make an *active* AC circuit, one driven by a alternating voltage. Our first step is to determine what the relationship is between voltage (provided by the generator) across *each* circuit element, one at a time, and the current *through* that circuit element as a function of time. We begin with the resistor, as the easiest to understand and as a model for the other two.

Our goal initially will be to enable analyzing *simple* AC circuits using *phasor diagrams* to represent the sum of several harmonic functions of the same frequency but possibly different phases. The advantage of using phasors is that one solves the differential equation of motion for those circuits using geometry and trigonometry and nothing but real numbers. This method works well at first and is a great way of developing a semi-quantitative *conceptual* understanding of what certain very important circuits do and how they work, but then it rapidly loses steam as one makes the circuits just a tiny bit more complicated but at the same time much more realistic and useful.

In order to do better, it is necessary to use complex numbers throughout to express and solve the differential equation of motion. This actually makes the solution *easier*[196] once one masters the complex algebra, as *it keeps track of all of the phases and "trig identities" required*

[196]Where by "easier" I in fact mean "possible" in less than an insane amount of work.

to use phasors automatically and algebraically! Basically, one no longer needs anything beyond trivial trig definitions and the Euler relation $e^{i\theta} = \cos\theta + i\sin\theta$ to solve almost any problem where one could solve a *DC* circuit problem with "resistances" in place of reactances!

The complex approach is ***the*** way AC circuits of this sort are analyzed in electrical engineering courses for this reason, and is a very useful thing for physics majors and math majors taking this course to work through in detail as well. It is, however, beyond the needs of most students primarily taking physics to support their understanding of other sciences or achieve physics literacy at a level necessary to thrive in medical school.

For that reason, when I teach this course to physics or biophysics majors, engineering students, or math majors in the more "advanced" version(s) of this course, I usually *require* that they at least walk through the complex/advanced solutions (and work the corresponding advanced homework problems), while e.g. life science majors taking physics only study the relatively simple phasor solutions to the most important circuit diagrams so that they can understand their purpose and see how that purpose is realized in the way currents flow as the frequency of the applied voltage is varied.

The next few sections, then, are entirely devoted to the phasor approach and all students (including math and physics majors and engineering students) should work through them before contemplating (or *not*) the complex approach.

13.4.1: A Harmonic AC Voltage Across a Basic Circuit Element

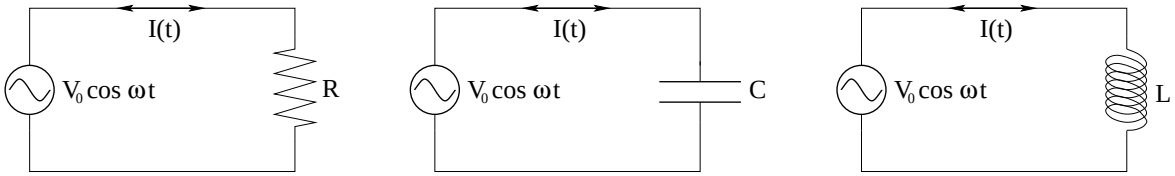


Figure 13.14: ***AC voltage drop across*** R***,*** L ***and*** C

Our first order of business will be to discover the surprisingly simple relationship(s) between the voltage drops across our basic circuit elements – R, L and C – and the current *through* these elements. For the phasor approach (which relies on applying some trig identities, be warned), we'll literally need to know these relations forward and backward. For the following complex approach, we can relax somewhat as the complex unit will encode them for us so we don't have to remember as much.

To do this, we need to imagine a circuit consisting of *only* a harmonic voltage $V_0 \cos \omega t$ applied directly across each circuit element in turn[197]. Note that this voltage need not be an actual e.g. AC generator – it might be an antenna, or a signal carried down a wire, or it could be the *resulting* voltage at that point of everything "upstream" of the circuit element in question. All of the resulting phases in the analysis below are considered *relative to* the phase of this incoming source of voltage, which we'll view as "the voltage drop across element X when the current through it is I".

In each case, we will apply Kirchhoff's Loop Rule and then use our known expression for the voltage drop across the element in question. Let's start with R, where the voltage across the resistor is given by our old friend, Ohm's Law. Then we can label:

$$V_R(t) = V_0 \cos(\omega t) \tag{13.65}$$

as the voltage across it and is illustrated in the first panel of figure 13.14 and write Kirchoff's Loop Rule as an "equation of motion" (so named because everything is now *dynamic* and varying in time) for the circuit:

$$V_R(t) - I_R(t)R = 0 \quad \Rightarrow \quad V_0 \cos(\omega t) - I_R R = 0 \tag{13.66}$$

Solving for the desired current, we get:

$$I_R(t) = \frac{V_0}{R} \cos(\omega t) = I_0 \cos(\omega t) \qquad \Leftrightarrow \qquad V_R(t) = I_0 R \cos(\omega t) \tag{13.67}$$

and we see that the harmonic current I_R through a resistor R is *in phase* with the voltage drop V_R across the resistor and has amplitude $I_0 = V_0/R$.

Next, we do the capacitor using the second panel of figure 13.14 to represent the harmonic no-phase voltage drop across C, Kirchhoff's Loop Rule, and the definition of capacitance to

[197] We could obviously use $V_R(t) = V_0 \sin(\omega t)$ or $V_0 \cos(\omega t + \phi)$ for an arbitrary ϕ as they simply represent a shifting of the zero of the clock used in the harmonic function – we choose this particular form to ***make*** $\boldsymbol{V_R(t)}$ ***the real part of the imaginary form:*** $\boldsymbol{V_0 e^{i\omega t}}$ and thereby enable the treatment of the complex solution in an "advanced" section below.

get the equation of motion:

$$V_C(t) - \frac{Q}{C} = 0 \quad \Rightarrow \quad V_0 \cos(\omega t) - \frac{Q}{C} = 0 \tag{13.68}$$

Again, we wish to find $I_C(t)$, the current through the capacitor, not $Q(t)$ per se. To get it, first we solve for $Q(t)$ and then differentiate to find $I_C(t)$:

$$Q(t) = CV_0 \cos(\omega t) \quad \Rightarrow \quad I_C(t) = +\frac{dQ(t)}{dt} = -(\omega C)V_0 \sin(\omega t) \tag{13.69}$$

We will use this equation to relate the harmonic voltage $V_C(t)$ across C (only) to the current through C (only) in more complicated circuits below, so it will be useful to express the current in terms of the given harmonic form of the voltage *plus a phase*. We therefore use the trigonometric identity[198] $-\sin(\theta) = \cos(\theta + \pi/2)$ to express the result in terms of the harmonic voltage :

$$I_C(t) = (\omega C)V_0 \cos(\omega t + \pi/2) \tag{13.70}$$

We can rewrite this equation in a familiar form by introducing a symbol for the "effective resistance" of the capacitor in the AC circuit, called the ***capacitive reactance*** χ_C:

$$\chi_C = \frac{1}{\omega C} \qquad \Leftrightarrow \qquad I_0 = (\omega C)V_0 = \frac{V_0}{\chi_C} \tag{13.71}$$

The units of χ_C are (hopefully obviously) *ohms*.

This definition makes conceptual sense! Capacitors act like open circuits with infinite resistance at $\omega = 0$ (DC circuits) and have zero "resistance" as $\omega \to \infty$. As long as the RC-circuit time constant $\tau_{RC} = RC$ is much larger than the period of the oscillating current, the capacitor itself has no *time* to build up any significant charge or potential difference before the current *reverses direction* and *discharges* any small charge/voltage that might have built up. We'll capitalize on this insight when we encounter AC "filter circuits" below.

Expressing the current in terms of the reactance we get a pair of equations that are the capacitor's version of the equivalent results for resistors derived in the previous section:

$$I_C(t) = \frac{V_0}{\chi_C} \cos(\omega t + \pi/2) = I_0 \cos(\omega t + \pi/2) \qquad \Leftrightarrow \qquad V_C(t) = I_0 \chi_C \cos(\omega t - \pi/2) \tag{13.72}$$

A key thing to remember from this is that:

> **The current is $\pi/2$ *ahead in phase* of the voltage drop across the capacitor. Equivalently, the voltage drop across the capacitor is $\pi/2$ *behind* the current through it.**

We will have occasion to use both of these statements in analyzing future circuits.

Finally, we repeat this process one more time for an inductance L. The methodology is basically the same. We start with the circuit portrayed in figure 13.5, write Kirchoff's Loop Rule using a zero-phase harmonic voltage $V_L(t)$ across the inductor, and then solve for $I_L(t)$:

$$V_L(t) - L\frac{dI_L}{dt} = 0 \quad \Rightarrow \quad V_0 \cos(\omega t) - L\frac{dI_L}{dt} = 0 \quad \Rightarrow \quad dI_L = \frac{V_0}{L} \cos(\omega t) dt \tag{13.73}$$

[198] These identities are difficult for most students – and myself! – to remember. One of the beauties of the complex approach is that you *don't have to remember* any trig identities to set things up. It's all built into the Euler relation!

Here we have to integrate both sides[199] to get:

$$\begin{aligned} I_L(t) &= \int \frac{V_0}{L}\cos(\omega t)dt \\ &= \int \frac{V_0}{\omega L}\cos(\omega t)\ \omega dt \\ &= \frac{V_0}{\omega L}\sin(\omega t) \end{aligned} \tag{13.74}$$

Again we write this as:

$$I_L(t) = \frac{V_0}{\omega L}\cos(\omega t - \pi/2) \tag{13.75}$$

and define the *inductive reactance* (in ohms):

$$\chi_L = \omega L \qquad \Leftrightarrow \qquad I_0 = \frac{V_0}{\omega L} = \frac{V_0}{\chi_L} \tag{13.76}$$

Again, this makes sense. At zero frequency, the "ideal" inductor has little or no resistance[200].

At very high frequency – where the period is much less than the LR circuit time constant R/L – the current is changing (and changing direction) so rapidly that very little current actually flows – the inductor de facto blocks the current with a very high effective "resistance". Again, this insight will help us to understand certain filter circuits below on conceptual grounds!

Putting this together, we get:

$$I_L(t) = \frac{V_0}{\chi_L}\cos(\omega t - \pi/2) = I_0\cos(\omega t - \pi/2) \qquad \Leftrightarrow \qquad V_L(t) = I_0\chi_L\cos(\omega t + \pi/2) \tag{13.77}$$

As before, we reduce this to a simple phase rule:

> **The current is $\pi/2$ *behind* of the voltage drop across the inductor. Equivalently, the voltage drop across the inductor is $\pi/2$ *ahead of* the current through it.**

These three matched pairs of results – one each for resistor R, capacitor C, and inductor L – can be assembled in a variety of ways in series, parallel, or mixed circuits. To make life maximally simple, I'll assemble all three into a single box that you can easily refer to when

[199] Doing an *indefinite* integral or setting the constant of integration to 0. Note that the constant of integration would correspond to the circuit having a constant "baseline" current in the infinite, zero resistance loop. That is clearly impossible as the voltage source and the wires themselves have *some* nonzero resistance, which would damp any such current to zero and leave one with only the harmonic part. The constant of integration is part of the *transient* response in the circuit discussed below in a broader context.

[200] In the real world, of course, inductors are generally made out of a coil of thin wire and many turns, have internal resistance associated with a paramagnetic or ferromagnetic core with its eddy currents (that generate heat!) and therefore have a *non-zero* resistance in their own right. This is one of many complications that the simple phasor+trig approach will ultimately have a hard time accommodating.

working through the following sections:

$$
\begin{aligned}
I_R(t) &= \frac{V_0}{R}\cos(\omega t) = I_0\cos(\omega t) &\Leftrightarrow\quad V_R(t) &= I_0 R\cos(\omega t)\\
I_C(t) &= \frac{V_0}{\chi_C}\cos(\omega t+\pi/2) = I_0\cos(\omega t+\pi/2) &\Leftrightarrow\quad V_C(t) &= I_0\chi_C\cos(\omega t-\pi/2)\\
I_L(t) &= \frac{V_0}{\chi_L}\cos(\omega t-\pi/2) = I_0\cos(\omega t-\pi/2) &\Leftrightarrow\quad V_L(t) &= I_0\chi_L\cos(\omega t+\pi/2)
\end{aligned}
$$

where in each case V_0 or I_0 should be interpreted as "harmonic voltage amplitude across the circuit element" or "harmonic current amplitude through the circuit element" in question, not as an overall voltage or current produced by an actual power supply somewhere in a complicated circuit.

13.4.2: The Series LRC Circuit – Phasor Approach

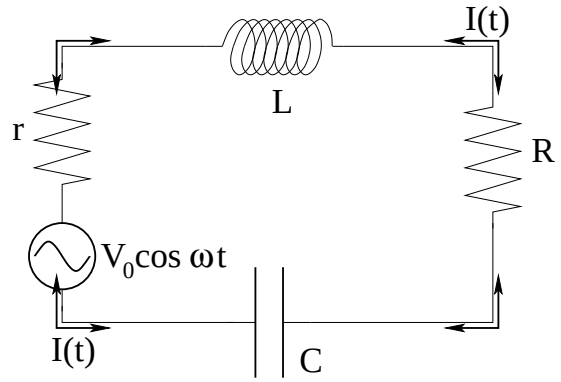


Figure 13.15: A series LRC circuit, for a non-ideal power supply with internal resistance r.

In figure 13.15 above we see a ***series*** LRC circuit, also known for obscure reasons as a "tank circuit"[201]. Note well that the order doesn't matter – different books might call this an RLC circuit or LCR circuit.

Let's analyze it by following – as closely as we can – the methodology developed in the previous section. In particular, we'll apply Kirchhoff's Loop Rule the (single loop) circuit to get a *differential* equation of motion.

First, however, note well that *I included an internal resistance for a non-ideal power supply in the diagram!* I did this deliberately. The second resistor, labelled R, is something called the *load resistance* in circuits of this sort. As we will see, the ***useful work*** done by circuits of this sort is almost invariably going to be done *delivering power to the load resistance* which might be, for example, an amplifier. Most of the ***useless, undesirable work*** done by the circuit will be *energy dissipated as heat inside of the power supply*. Also, in the case of antennas

[201] Apparently its mathematical description resembles in some fashion the way the fluids pulsed through a tank can resonate according to the tank dimensions. I include the term – once – just in case a student hears the term elsewhere and wonders what the hell it refers to.

specifically as voltage sources, they can only deliver a *tiny* amount of power and typically have a *large* internal "radiation" resistance.

While it is true that the specific circuit above might well be used in circuits where the internal resistance is *negligible* compared to the load resistance, $r \ll R$, even so we might want to know just how the total power coming in from the "power supply" on the left is being split up between r and R, which is going to be difficult if we don't include r from the beginning. In addition, the *parallel* LRC circuit we're going to analyze next will turn out to be *pointless* if $r = 0$, where actually it is quite useful and sometimes preferred when it is not!

This won't really complicate the first part of the solution. r and R are in *series*, so we can just define:

$$R' = r + R \tag{13.78}$$

and treat the circuit as if there is a single *effective* resistance R' in the circuit. Going around the loop above clockwise, then:

$$V_0 \cos(\omega t) - V_L(t) - V_{R'}(t) - V_C(t) = 0 \tag{13.79}$$

where:

$$V_L = L\frac{dI}{dt} \qquad V_{R'} = R'I \qquad V_C = \frac{Q}{C} \tag{13.80}$$

We wish to solve this equation for the ***common current in the circuit*** as we know that, given the current, we can solve for literally *everything* that there is to know about the circuit. In particular, we can easily find the voltage drop across each element, the power delivered to or provided by each element, and with a bit of thought, we can gain a conceptual understanding of the *purpose* of series arrangements of inductors, capacitors, and resistors in general circuit design (arguably one of the most important things in the chapter, which after all isn't about memorizing formulas without any regard for what they tell us at the cognitive level).

Before we properly begin, we need to note two things. The first is a mathematical observation. If we use the relations:

$$I = \frac{dQ}{dt} \qquad \frac{dI}{dt} = \frac{d^2Q}{dt^2}$$

we can convert the loop equation into a ***second-order, linear, homogeneous ordinary differential equation with a harmonic driving term***:

$$\frac{d^2Q}{dt^2} + \frac{R'}{L}\frac{dQ}{dt} + \frac{1}{LC}Q = \frac{V_0}{L}\cos(\omega t) \tag{13.81}$$

This equation is, note well, precisely equivalent to the damped, driven, simple harmonic oscillator equation studied in the Mechanics and Applications textbook associated with this one, with the mappings:

$$\left\{k \Leftrightarrow \frac{1}{C}\right\} \qquad \{m \Leftrightarrow L\} \qquad \{b \Leftrightarrow R'\} \qquad \{F_0 \Leftrightarrow V_0\}$$

One can literally copy every solution or conclusion from this chapter to that or vice versa by making these substitutions.

Next, another math-y item you may or may not know already. The solution to *inhomogeneous, linear* ordinary differential equations (in general, not just second order) can most generally be written as:

$$Q(t) = Q_h(t) + Q_i(t)$$

where $Q_h(t)$ is a solution to the associated *homogeneous* ODE:

$$\frac{d^2Q_h}{dt^2} + \frac{R'}{L}\frac{dQ_h}{dt} + \frac{1}{LC}Q_h = 0 \tag{13.82}$$

We literally *just solved this* in the first "passive" part of this chapter so you all should still recognize/remember that:

$$Q_h(t) = Q_{0h}e^{-R't/2L}\cos(\omega' t + \phi) \qquad \text{with} \qquad \omega' = \omega_0\sqrt{1 - \frac{R'^2C}{4L}}$$

is at least the *underdamped* solution to the homogeneous ODE.

This component of the solution – underdamped or not – is referred to as the ***transient***; it decays away so that after a few of the longest exponential decay times even the overdamped solution will be negligible. Its (somewhat complicated) time derivative contributes to the desired current $I(t)$, to be sure, but only initially is that contribution substantial!

The transient part of the solution *does* contain the two required *contants of integration* that can be varied to match any physically consistent set of initial conditions in the circuit, but that's generally – or at least *often, usually, mostly* – irrelevant to the purpose for building such a circuit in the first place. We will henceforth ignore it and concentrate on the second part. $Q_i(t)$ as the *unique* solution to the *inhomogeneous* ODE *after* the transient has died away. Note well that we effectively *used* this in the previous section without overthinking it to ignore any possible initial charge on the capacitor or current through the inductor as the hitherto-neglected internal resistance of the power supply and/or the non-ideal circuit elements themselves would be enough to eventually damp out the specific initial state(s).

The remaining part $Q_i(t)$ has *no free parameters* and *persists indefinitely*. For this reason it is more generally referred to as the ***steady state*** solution to the equation of motion. It alone suffices to reproduce the "inhomogeneous" function $V_0\cos(\omega t)$ when the "damped oscillator" differential form on the left is applied to $Q_i(t)$. Because the ODE is *linear*, it is trivial for you to verify that the sum of the transient part plus the steady state part always solves the inhomogeneous ODE for all possible nonzero constants of integration.

The last thing for us to note before we proceed is that – because we have taken the time to derive the relationships between harmonic voltage across and harmonic current through *each* of the circuit elements, we will find it ***much more convenient*** to solve for the common steady state *current* $I(t)$ through *all* of the circuit elements in series, not "just" the charge $Q(t)$ on the capacitor (steady state or not). This will allow us to express the voltage drop expected across each element in terms of the the common current and the reactances of each element as summarized in the box at the end of the previous section. We can summarize the relevant relationships between "voltage across" and "current through" from the box in *words* as:

> **The voltage drop across any circuit element with a harmonic current through it *is itself harmonic at the same frequency as the current* but with *no* phase shift (for R), a phase shift of $+\pi/2$ (for L) or $-\pi/2$ (for C) relative to that current.**

Since the sum of these voltages must *equal* the harmonic driving voltage with its particular given phase (of zero), we expect the real form of the steady state current that exactly solves

the single-loop series LRC problem to be:

$$I(t) = I_{ss}(t) = I_0 \cos(\omega t - \delta) \tag{13.83}$$

where the phase δ and amplitude I_0 are ***not free parameters***, but are explicit functions of V_0, ω, L, R, and C. Note that this expression *works* for each circuit element alone with the values of $\delta = 0, \pm\pi/2$.

From now on I'm dropping the "ss" subscript in this and related equations because we will *only* be interested in the steady state current after the transient has died away. For the remainder of the chapter we'll just call a currents in some specific part of the circuit $I_i(t)$ (with a subscript i labelling the particular element or branch the current passes through, if needed) with no additional "ss" marking; it is *assumed* to be the current in the steady state after the transient has decayed away.

This form 13.83 is not really an assumption or guess. It is, as I argued verbally above and as we'll see explicitly/algebraically below, the *necessary* form for the solution given a single harmonic frequency driving voltage[202].

We *will* assume that the current given above is positive when *clockwise*. This assumption is consistent with the one we made above for single circuit elements placed across an AC voltage, allowing us to just *substitute equations for the voltage given the current* from the box at the end of the previous section into Kirchoff's Loop Rule for the circuit equation 13.79! We get:

$$I_0 \chi_L \cos(\omega t - \delta + \pi/2) + I_0 R \cos(\omega t - \delta) + I_0 \chi_C \cos(\omega t - \delta - \pi/2) = V_0 \cos(\omega t) \tag{13.84}$$

Our goal, then, is to find the specific, unique values of I_0 and δ for which this equation is *true*. To accomplish this we'll use a *phasor diagram*[203] – we'll invent a bunch of "vectors" with appropriate lengths (called "phasors") at polar angles given by the arguments of the cosine functions in this equation, then add them up using the triangle rule until the sum of the x-components of the vectors in the figure represents our equation of motion 13.84.

There are two really cool aspects of doing this. One is that we've *already eliminated* all of the "second order differential equation" calculus associated with the equation of motion by

[202]**Warning:** It may not be a guess, but it is far from consistently expressed in textbooks and application. The minus sign of the phase in the equation is basically a convention. The particular symbol δ used for the phase is arbitrary – many books use e.g. ϕ instead. Some books use sine instead of cosine for the harmonic driving voltage, and hence use sine in the current as well. Electrical engineering texts might use ϕ instead of ω for the driving frequency, and they almost invariably use the complex approach discussed in a section following this one. To make matters worse, they typically use lower-case i for *current* forcing them to use j for the complex unit $j^2 = -1$ (physicists use j for e.g. current density). Some physics books use ν for angular frequency, although that's more common in the context of quantum theory. It's enough to drive you completely nuts – it's not easy for *me* and I'm far from being a beginner at all this stuff.

Sigh. I henceforth decree that *this* textbook contains *the one and only true notation*. Just kidding, I wish. Let me instead decree that – as has already been true in a number of places in the two-semester course – you need to learn to recognize symbols and meaning ***in context*** and not get *too* hung up memorizing particular symbols as having particular universal meanings, at least if you want to be able to use multiple reference works figuring something out. It's the best we can do.

[203]Wikipedia: http://www.wikipedia.org/wiki/Phasors. A portmanteau of "phas(e vect)or", not a space weapon from Star Trek... It's worth skimming this article, especially if you are an advanced student, although the presentation here is self-sufficient for our purposes.

doing it piecewise (finding the voltage in terms of the current and vice versa) in the box at the end of the previous section. Second, it turns out that this *one* equation is sufficient to obtain *two* unknowns – I_0 and δ – because it is really the real part of a 2D *complex* equation (or the x-component of a 2D vector relation that must also be consistent in the y-component) and hence in some sense is – two equations? Yes, I know, this is black magic, but are not physicists[204] the *wizards of the modern age?*

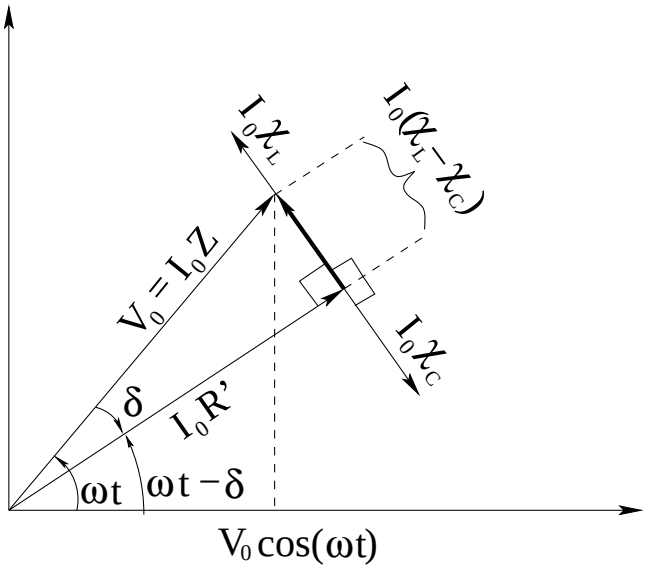


Figure 13.16: The phasor diagram for the series *LRC* circuit.

Figure 13.16 is the resulting phasor diagram. The voltage phasor of length $I_0 R'$ associated with R' is in phase with the current (at the angle $\omega t - \delta$). The voltages of the other two phasors have the right lengths and phases – $I_0 \chi_L$ for L, $\pi/2$ *ahead* of the common current and $I_0 \chi_C$ for C, $-\pi/2$ *behind* the common current. The three phasor themselves add up to the phasor of length V_0 oriented at the (time varying) angle ωt.

Because this is true, the sum of the x-components of the phasors on the diagram *exactly represents* the algebraic equation 13.84 as illustrated with the dashed vertical projector (for V_0 only). To solve the equation of motion for the current, then, we only need to express V_0 in terms of I_0, the common current amplitude and use the pythagorean theorem and some trig on the triangle in figure 13.16!

We accomplish this extending the methodology used in the previous section. That is, we'll write the equation for the applied voltage ***in terms of the amplitude I_0 of the current through it*** as:

$$V(t) = V_0 \cos(\omega t) = I_0 Z \cos(\omega t) \qquad \Leftrightarrow \qquad I_0 = \frac{V_0}{Z} \tag{13.85}$$

which might well be a *fourth* equation pair "in the box" at the end of the previous sections. In this equation we have defined a quantity Z called the ***impedance*** of the circuit, the equivalent of the *total* AC "resistance" to current flow, measured in ohms. Note well that there is no δ or other phase shift in this definition – this is the actual voltage maintained by the AC power supply so we use its phase to *define* the zero phase of *all* of the downstream harmonic phases used throughout the problem!

[204] Sigh. OK, OK. "...and mathematicians". There, now, are you happy?

Making this substitution lets us ***cancel the common factor*** $\boldsymbol{I_0}$ in the diagram and draw the triangle where the legs are all *effective resistances* – the impedance Z, the reactances $\chi_{L,C}$, or the resistance R'. Since the *relative* phases in the right trangle don't depend on ω or t, we can just draw the triangle in the most convenient orientation as in figure 13.17

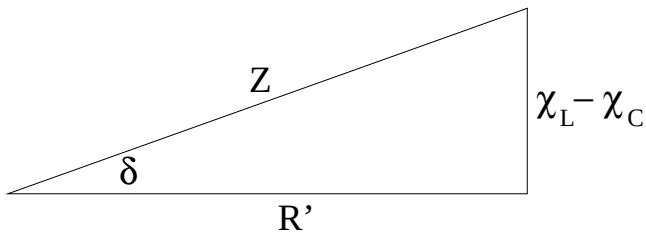


Figure 13.17: The impedance diagram for the *LRC* circuit.

From the pythagorean theorem, ordinary trigonometry and a bit of *algebra* (factoring L/ω out of the term with the reactances and using the definition of the resonant frequency of the *undamped* circuit $\omega_0 = 1/\sqrt{LC}$) we see from this figure that the impedance is given by:

$$Z = \sqrt{R'^2 + (\chi_L - \chi_C)^2} \tag{13.86}$$

(units/dimensions manifestly ohms) and the phase angle δ is given by:

$$\delta = \tan^{-1}\left(\frac{\chi_L - \chi_C}{R'}\right) \tag{13.87}$$

Substituting, we find the common current in the circuit loop!

$$\boxed{I(t) = \frac{V_0}{Z}\cos(\omega t - \delta) = \frac{V_0}{\sqrt{R'^2 + (\chi_L - \chi_C)^2}}\cos(\omega t - \delta)} \tag{13.88}$$

Note that there are ***several other useful ways of writing this result!*** I'm going to concentrate on the one way of writing it that is the most useful for the specific purpose of computing, and understanding the ***power delivered to the load resistance in the circuit*** in a dimensionless, easily scaled form. This is the topic of the next section.

13.4.3: A Universal description of the Current $I(\omega)$: Scaling into a Dimensionless Form

The expression just obtained for the current in equation 13.88 is the one you are most likely to encounter in most introductory physics books. On the good side, it let's you see that when the inductive reactance equals the capacitative reactance (that is, when:

$$\omega L = \frac{1}{\omega C} \quad \Rightarrow \quad \omega^2 = \omega_0^2 \quad \Rightarrow \quad \omega = \omega_0$$

or the circuit is in *resonance*) $Z = R$ and the current is at a maximum. On the other hand, we might also note that it is by no means easy to see how the current $I(\omega)$ will change as we change L, R; and C independently. This difficulty persists for most of the simple factorizations we might attempt. Surely there is an easier way to compress these parameters and hence visualize the behavior of the current (and later, the power delivered to the load resistance)!

Recall from earlier in the chapter the dimensionless ***Quality Factor***:

$$Q = \frac{L\omega_0}{R'} = \frac{1}{\omega_0 R' C} \quad \Rightarrow \quad \boxed{Q = \sqrt{\frac{L}{R'^2 C}} = \sqrt{\frac{\tau_{LR}}{\tau_{RC}}}}$$

that describes the energy damping of the circuit (where I used $R' = R + r$ in these expressions of Q, note well). We saw in the mechanics part of this course that the quality factor of a driven simple harmonic oscillator governed the "sharpness" of the resonant power curve, and since there is a *perfect* analogy between this problem and the series LRC circuit, we expect the same to be true here! Let's do some algebraic rearrangements of Z, and hence the current amplitude.

Our goal is to factor Z – which has units of resistance – into R' times a ***manifestly dimensionless, universal form*** that permits us to see at a glance 'how resonance works'. The result will let us obtain $I_0(\omega)$, the power delivered to the load resistance R $P_R(\omega)$ and so on in a form that scales with Q and ω_0 in easily understood ways. We'll start by factoring out the R' from Z as a function of ω:

$$Z(\omega) = R'\sqrt{1 + \left(\omega\frac{L}{R'} - \frac{1}{\omega R' C}\right)^2} \tag{13.89}$$

If we examine the definitions of Q above, parts of it are lurking in this expression. We multiply each of them factors involving L, R', C, and ω by $\omega_0/\omega_0 = 1$ and do a bit of rearrangement to put Z into the desired universal form:

$$\omega\frac{L}{R'} = \frac{\omega}{\omega_0} \times \frac{L\omega_0}{R'} = Q\frac{\omega}{\omega_0} = Q\beta \tag{13.90}$$

$$\frac{1}{\omega R' C} = \frac{\omega_0}{\omega} \times \frac{1}{\omega_0 R' C} = Q\frac{\omega_0}{\omega} = Q\frac{1}{\beta} \tag{13.91}$$

where we've defined the dimensionless/scaled frequency:

$$\beta = \frac{\omega}{\omega_0}$$

which is always equal to *one* at the resonance peak. Then:

$$Z(\omega) = Z(\beta) = R'\sqrt{1 + Q^2\left(\beta - \frac{1}{\beta}\right)^2} \tag{13.92}$$

which is our desired, manifestly dimensionless form.

There are a few interesting features of this remarkably symmetric form:

$$\lim_{\beta\to 0} Z(\beta) = \infty \qquad \lim_{\beta\to\infty} Z(\beta) = \infty \qquad Z(\beta = 1) = Z_{\min} = R' \text{ at resonance} \tag{13.93}$$

We can now use this form to evaluate the current amplitude and phase:

$$I_0(\omega) \Rightarrow I_0(\beta) = \frac{V_0}{R'}\frac{1}{\sqrt{1 + Q^2\left(\beta - \frac{1}{\beta}\right)^2}} \tag{13.94}$$

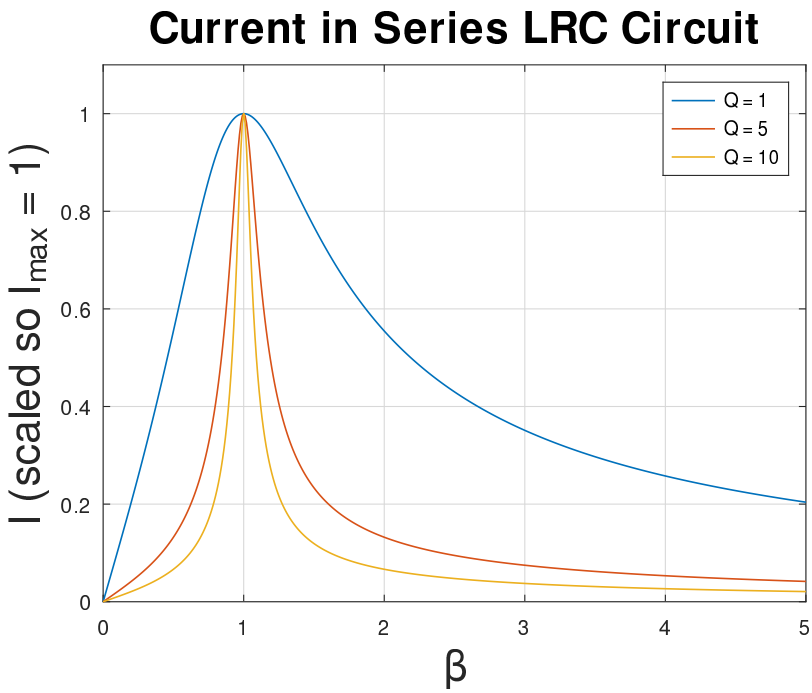


Figure 13.18: The amplitude $I_0(\omega, \omega_0, Q)$ for several values of Q, on axes that can easily be scaled by $I_{\max}$ and ω_0.

This is plotted below for several values of Q, with $I_{\max}$ set to 1. The curves shown can trivially be rescaled by simply *rescaling the abscissa and ordinate axes* so that the resonance occurs at ω_0 with a maximum of V_0/R', respectively!

We can apply *exactly the same transformations* to the expression we obtained for δ above:

$$\tan\delta = \left(\frac{\chi_L - \chi_C}{R'}\right) = \left(\frac{\omega L}{R'} - \frac{1}{\omega R' C}\right) = Q\left(\beta - \frac{1}{\beta}\right) \tag{13.95}$$

and use a trig identity (easily derived if you don't remember it) to get a *second* useful form for the current amplitude:

$$I_0(\omega) \Rightarrow \delta = I_{\max}\frac{1}{\sqrt{1+\tan^2\delta}} = I_{\max}\cos\delta \tag{13.96}$$

In words, this second result is:

> The amplitude of the current in the circuit is the maximum possible current times the dimensionless power factor $\boldsymbol{\cos\delta}$.

The power factor will be discussed in detail below – when we discuss power!

We have now succeeded in compressing pretty much all of the "interesting" behavior of any given resonant circuit into a single dimensionless number δ, that is a moderately complicated function but easily evaluated function of ω, L, C and R' (or just Q, ω and ω_0, the three key parameters that specify all of the resonance curves of interest since they all depend on the overall current. We can do no better than to write out two summary box with these results

encapsulated:

$$\boxed{I_0(\omega) = I_{\max}\sqrt{1 + Q^2\left(\beta - \frac{1}{\beta}\right)^2} = I_{\max}\sqrt{1 + \tan^2\delta} = I_{\max}\cos\delta}$$

where

$$\boxed{\beta = \frac{\omega}{\omega_0} \qquad \delta = \tan^{-1}\left\{Q\left(\beta - \frac{1}{\beta}\right)\right\}}$$

13.4.4: Power in a Series LRC Circuit

OK, we've just spent a lot of time and energy analyzing the series LRC circuit, but we haven't really thought about ***why*** it is worth all of this effort. What exactly does this circuit do? Clever students who remember the damped, driven simple harmonic oscillator problem might well guess that it has something to do with *resonance*, and they'd be right! Resonant electrical circuits like this one are the basis of ***almost all communication and processing of information using electricity!*** Radio, television, the internet, computers – *none* of this would be possible without RLC circuits! Even the light you are processing with your eyes in order to read is delivering its energy *and information content* to your rods and cones and brain using a quantum version of this same general process.

It is difficult to get *more* important than that. You literally would not be reading these words without it, because even if you are reading a paper version of this textbook, information processing electronics were used in *every single step* of its preparation and delivery. If you are (perhaps more likely) reading an electronic version – well, thank your lucky stars for RLC circuits!

In order to see why the circuit is useful, we have to examine the flow and delivery of *power* in the circuit. After all, physics (and engineering!) are all about "doing work" – making a desired thing *happen* by arranging things 'just so'. In particular, since only the resistance R "uses" the energy provided by the power supply – L and C simply absorb energy and then give it all back conservatively, while r dissipates it uselessly – being able to selectively deliver power ***to*** the load resistance R for only a certain range of frequencies will end up being at least *one* major point of this kind of circuit. Another is to filter useful signals out of a bunch of noise or mixtures of other useful signals obtained ***from*** an upstream source – an application of interest even to future healthcare professionals[205] .

In words, the series LRC circuit is fundamentally useful in electronics because it functions as a resonant ***band pass filter***. Given an input harmonic voltage of some frequency, it only allows a large current to flow (and hence deliver significant *power* from the energy source in the input power supply to the circuit *load* resistance R) when the frequency of the applied/input voltage is *nearly the same as the resonant, undamped frequency* $\omega_0 = \sqrt{1/LC}$ *for the circuit!*

[205]Wikipedia: http://www.wikipedia.org/wiki/Neural Oscillation. Future physicians or neuroscientists take note – the human brain does part of its thinking with wave patterns with particular frequency ranges. The heart is driven by periodic waves of neural activity. Whether or not the body or brain have RLC circuits *per se* in them, in order for us to *filter out these signals* to examine them without the "noise" of all of the *rest* of the electrical activity of the body or brain mixed in requires the use of filter circuits fed by pickups attached to the body (EKG) or head (EEG).

This is the electronic equivalent of pushing your little brother or sister on a swing *at* the natural frequency of the swing and thereby building up a large amplitude of oscillation! For frequencies *far* from this resonant frequency – in either direction, above it or below it – the current in the entire circuit and hence power input by the voltage and delivered to the load rapidly goes to zero. Let us understand this.

Consider the power delivered to or by each circuit element. The instantaneous power delivered *by* the input voltage *to* the circuit is just:

$$P_V(t) = V(t)I(t) = V_0\cos(\omega t)I_0\cos(\omega t - \delta) \tag{13.97}$$

If we use the trig identity[206] $\cos(\omega t - \delta) = \cos(\omega t)\cos(\delta) + \sin(\omega t)\sin\delta$ and $I_0 = V_0/Z$ we get:

$$P_V(t) = \frac{V_0^2}{Z(\omega)}\left(\cos^2\omega t\cos\delta + \cos\omega t\sin\omega t\sin\delta\right) \tag{13.98}$$

We don't usually care about the *instantaneous* power associated with elements in this circuit as some of that power is just energy moving in and out of the "conservative" devices L and C. We *do* care about the *time-averaged* power. Let's start by time averaging the power provided by the input voltage, as this is indeed the rate that power supply is doing net work on the system. We get:

$$\langle P_V\rangle = P_{\text{avg}} = \frac{1}{2}\frac{V_0^2}{Z(\omega)}\cos\delta = \frac{1}{2}\frac{V_0^2R'}{Z^2} = \frac{1}{2}I_0^2R' \tag{13.99}$$

where we used the fact that the time average of the square of *any* harmonic function of time is $1/2$. If you look back at the right triangle in the phasor diagram for the circuit from a few pages back, you can see how I identified the ***power factor*** mentioned in the last section:

$$\cos\delta = R'/Z$$

The power factor will be discussed in more detail shortly, which is why I keep putting down results in terms of it even when we have several other forms that also work.

Note that in evaluating this, the time average of $\cos\omega t\sin\omega t\sin\delta$ turns out to be *zero*. There are several ways of obtaining this result (e.g. u-substitution and direct integration) but likely the simplest one is that using the trig identities derived in the footnote we see that:

$$\cos(\omega t)\sin(\omega t) = \frac{1}{2}\sin(2\omega t)$$

and note that the time average of *any* harmonic function is zero. Either way, this term *does not contribute to the **average** power*, although it does contribute to the *peak* power delivered

[206] Which can be trivially proven with complex exponentials and the Euler equation in two lines:

$$e^{i(A-B)} = e^{iA}e^{-iB} = (\cos A + i\sin A)\times(\cos B - i\sin B)$$
$$\cos(A-B) + i\sin(A-B) = (\cos A\cos B + \sin A\sin B) + i(\sin A\cos B - \cos A\sin B)$$

or (equating real and imaginary parts:

$$\cos(A-B) = \cos A\cos B + \sin A\sin B \quad \text{and} \quad \sin(A-B) = \sin A\cos B - \cos A\sin B$$

by the input voltage[207] Physically, this second term describes instantaneous power going into the circuit (say, into L or C) and *coming right back out again* as L or C returns energy to the power supply/circuit unconsumed.

Now let's consider the power flowing *into* each of of the *other* circuit elements in the loop separately, ensuring that Kirchoff's Loop Rule still leads to strict energy conservation:

$$P_{R'}(t) = V_{R'}(t)I(t) = I_0^2 R' \cos^2(\omega t - \delta) \tag{13.100}$$
$$P_L(t) = V_L(t)I(t) = I_0^2 \chi_L \cos(\omega t - \delta + \pi/2)\cos(\omega t - \delta) \tag{13.101}$$
$$P_C(t) = V_C(t)I(t) = I_0^2 \chi_C \cos(\omega t - \delta - \pi/2)\cos(\omega t - \delta) \tag{13.102}$$

Again we don't much care about the peak values, but the averages are important. Obviously, the time average of $\cos^2(\omega t - \delta)$ is still $1/2$, so it is easy to find the average power delivered to the total series resistance R'. For the other two, note that:

$$\cos(\omega t - \delta \pm \pi/2)\cos(\omega t - \delta) = \mp \sin(\omega t - \delta)\cos(\omega t - \delta)$$

and the time average of either one is zero by the same argument as before! Then:

$$\langle P_R \rangle = \frac{1}{2} I_0^2 R' = \langle P_V \rangle = P_{\text{avg}} \tag{13.103}$$
$$\langle P_L \rangle = 0 \tag{13.104}$$
$$\langle P_C \rangle = 0 \tag{13.105}$$

In words, the average power provided to the circuit by the applied harmonic input voltage is delivered to the ***total resistance R' only!*** Some of this power appears in the load resistance R, some of it (as heat) in the internal resistance of the power supply, but none of it, on average, flows into the inductance or capacitance. This is precisely what we expect.

But what about L and C? Aren't they important too? They are indeed! They contribute to – sometimes even *dominate* – the limiting of the *peak* current provided to the system! When $\chi_L \gg R', \chi_C$ (the high frequency limit), it is primarily the inductor in the circuit that limits the peak current. Similarly, $\chi_C \gg R', \chi_L$ at low frequencies, it is the capacitor that limits current flow. The balancing of these three elements are therefore *crucial* design characteristics and are a major reason that we bothered above to separate out $\cos\delta$ and give it a name!

Now that this is established, and with our lovely dimensionless/scalable version of the current developed in the last section in hand, let's express the *power delivered to R and r independently* in similarly "maximally dimensionless, rescalable" terms that will depend only on V_0, ω, R', ω_0 and Q – the *most useful* descriptors of the resonant circuit. We will assume that:

$$I_{\max} = \frac{V_0}{R'} \quad \text{and} \quad Q = \sqrt{\frac{L}{R'^2 C}} \quad \text{and} \quad \beta = \frac{\omega}{\omega_0}$$

have already been computed or (in the case of β) will be used as our independent variable

[207] Note well that the term *itself* is zero if $\delta = 0$ *independent* of time simply because $\sin 0 = 0$, and the range of δ (as an inverse tangent) is $[-\pi/2, \pi/2]$ and doesn't contain any other zeros.

and express the results in terms of them:

$$P_{\mathrm{avg},R}(\beta, Q) = \frac{1}{2}\frac{I_{\max}^2}{1 + Q^2\left(\beta - \frac{1}{\beta}\right)^2}R \tag{13.106}$$

$$P_{\mathrm{avg},r}(\beta, Q) = \frac{1}{2}\frac{I_{\max}^2}{1 + Q^2\left(\beta - \frac{1}{\beta}\right)^2}r \tag{13.107}$$

where again I provide two slightly different forms that can be used to plot the result in easily evaluated and scaled ways.

At resonance ($\beta = 1$), the power reaches its maximum value for both internal resistance and load resistance:

$$P_{\max,R} = \frac{1}{2}I_{\max}^2 R = \frac{1}{2}V_0^2\frac{R}{(r+R)^2} = \left(\frac{1}{2}\frac{V_0^2}{r+R}\right)\frac{R}{r+R} \tag{13.108}$$

$$P_{\max,r} = \frac{1}{2}I_{\max}^2 r = \frac{1}{2}V_0^2\frac{r}{(r+R)^2} = \left(\frac{1}{2}\frac{V_0^2}{r+R}\right)\frac{r}{r+R} \tag{13.109}$$

$$\tag{13.110}$$

The term in parentheses is clearly ***the maximum possible average power delivered to the circuit in resonance*** given $R' = r + R$. The factor multiplying it is the ***fraction of that power delivered to each resistor***. This kind of circuit is therefore referred to as a *voltage divider*, as it divides the voltage drops and resulting power proportionally among the various resistive loads in the series circuit (there might be more than two, as I hope is obvious, and nothing but R' and the number of places the power is distributed would change).

Restoring the dimensionless variation of the power in terms of ω from above we get the *conveniently scalable form* for the average power delivered to the load resistor R specifically:

$$P_{\mathrm{avg},R}(Q, \beta) = \frac{P_{\max,R}}{1 + Q^2\left(\beta - \frac{1}{\beta}\right)^2} \quad \text{with} \quad P_{\max,R} = P_{\mathrm{avg},\max,V}\frac{R}{r+R} \tag{13.111}$$

(which obviously also describes the average power going into the source resistance with the simple substitution of $R \Rightarrow r$ in the divider fraction only). The total average power delivered by the power supply is split proportionally between the available series resistances, all of which share the *same* current! This makes sense!

Given a power supply with peak voltage V_0 and a purely resistive internal impedance r, the maximum *possible* power that can be delivered to the load resistance R for the given power available to the power supply occurs when $R = r$ – the Maximum Power Theorem you proved in the DC-circuits chapter. However, note well that the *real* Maximum Power Theorem in the context of AC-circuits involves more than just the load and internal resistances. Inductors often have nontrivial resistance (ignored in the "ideal" limit) that might need to be rolled into r as well. Capacitances, resistors, and even wires have a tiny bit of inductance. It turns out that the true maximum power occurs when the power supply impedance is *matched* in a particular way to the total load-side impedance.

Careful electrical circuit design has to at least be *able* to handle these nuances, even if the conclusion at the end is that one can ignore the non-ideal parts – sometimes one *can't!*

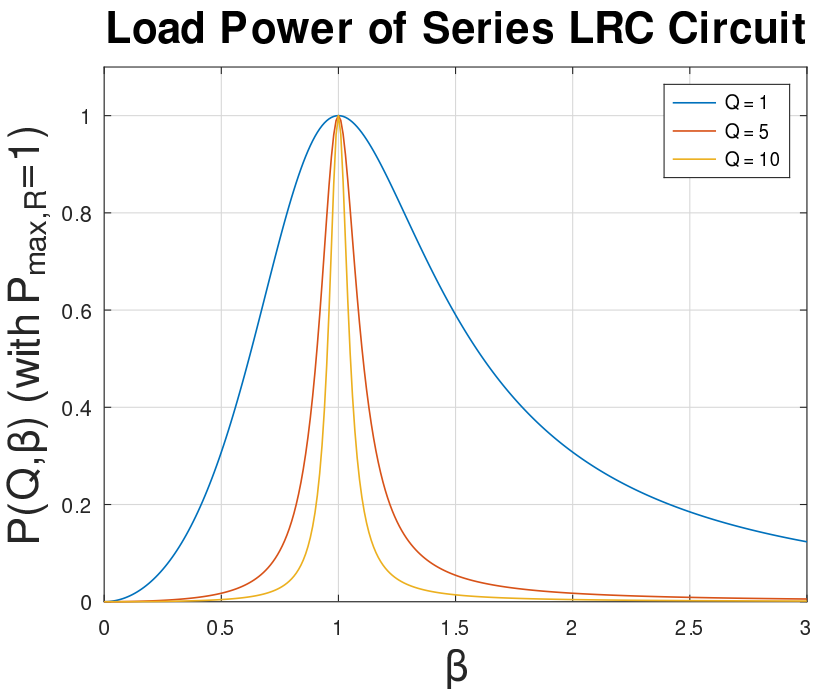


Figure 13.19: A typical series of resonance curves for $\mathcal{Q} = 1, 5, 10$, plotted on a scale such that $\omega_0 = 1$, $P_{\max} = 1$. This graph is *universal* and works for *any* values of $P_{\max,R}$, ω_0 and an appropriately given or evaluated $\mathcal{Q}$ once one scales the abscissa and ordinate axes!

However, exploring this (and proving the more general theorem) are beyond the scope of this course.

In any event, the final result equation 13.111 makes it easy to generate a *universal plot of the average power delivered to the load*. All we need to do is compute from the given quantities (V_0, R, L, C, and r:

$$\omega_0 = \frac{1}{\sqrt{LC}} \qquad P_{\max,R} = \frac{1}{2}\frac{V_0^2 R}{(r+R)^2} \qquad \mathcal{Q} = \frac{1}{(r+R)}\sqrt{\frac{L}{C}}$$

and scale the abscissa and ordinate axes accordingly before plotting the universal function for the resulting value of $\mathcal{Q}$. We display several of these plots in figure 13.19 for the *same* values of $\mathcal{Q}$ used to plot $I_0(\beta, \mathcal{Q})$ above.

For damped, driven LRC circuits, there is another meaning for $\mathcal{Q}$ in the weak damping limit. As you will (maybe) show in an "advanced" homework problem below:

$$\mathcal{Q} \approx \frac{\omega_0}{\Delta\omega}$$

where $\Delta\omega$ is called ***the full width of the power curve at half of its maximum value!*** It is also more casually used as an estimate of the ***bandwidth*** of the resonance, the range of frequencies *around* the resonant frequency that are significantly "passed" by the band-pass filter. These frequencies are essential for *encoding information on top of a harmonic carrier frequency* – one needs "enough bandwidth"[208] in order to be able to resolve the encoded information!

[208] A term that is nearly universally interpreted as meaning "enough mental capacity" in common parlance these days...

Note well that for $\mathcal{Q} = 1$, the approximation: is *pretty lousy* – as one might expect given our derivation of $\mathcal{Q}$ in the *weak damping limit* (which $Q = 1$ is not!) in the passive LRC section above. By $\mathcal{Q} = 3$, however, it is accurate to within a few percent and by $\mathcal{Q} = 5$, plotted above, it is visibly almost perfectly valid.

The one thing you should be certain to master in this section is the ability to sketch this *generic, dimensionless shape* for the resonance curve given any arbitrary value of $\mathcal{Q} \gtrsim 3$. The dimensionless $\mathcal{Q}$-factor is a measure of the *relative sharpness* of the resonance. Note that:

$$\Delta\omega = \frac{\omega_0}{\mathcal{Q}}$$

where $\Delta\omega$, again the ***full width at half-maximum power*** or ***bandwidth*** of the resonant circuit. This means that the larger $\mathcal{Q}$ is, the smaller the bandwidth and sharper the resonance in a way that scales in a simple way for specific values for $P_{\max}$ and ω_0.

The rules for drawing this are simple. Note that the power vanishes at $\omega = 0$ (where Z goes to ∞) and goes asymptotically to zero as $\omega \to \infty$ (where Z also goes to ∞ in a surprisingly *symmetric* way once we plot things in terms of Q and β). The steps you should follow are then:

- Identify and mark $P_{\max}$ on the ordinate (vertical axis).
- Draw a light dashed line across at that height.
- Identify and mark $P_{\max}/2$ on the ordinate (vertical axis).
- Draw a light dashed line across at *that* height.
- Identify and mark ω_0 on the abscissa (horizontal axis).
- Draw a light dashed line vertical line at that value.
- Mark the intersection of this line with the $P_{\max}$ line with a light dot.
- Determine $\Delta\omega = \omega_0/\mathcal{Q}$ and identify the interval of that width on centered on ω_0 on the abscissa, marking the points.
- Draw light dashed line vertical lines at those two values.
- Mark the intersections of these lines with the $P_{\max}/2$ line with light dots.
- Draw a smooth curve, starting at zero at $\omega = 0$, through the three dots and then down to die off as $\omega \gg \omega_0$.

These steps are graphically illustrated in figure 13.20 to generate a nicely peaked curve that is "hand drawn" with my usual mouse-driven plotting program (not graphed numerically with octave) for $\mathcal{Q} = 8$: Note that getting the shape "exactly right" is too much to expect when sketching a graph like this by hand, but it should have *all of the features* represented in the drawing "recipe" above!

In the homework "advanced" students will be asked to derive the relation:

$$\mathcal{Q} = \frac{\omega_0 L}{R'} = \frac{1}{R'}\sqrt{\frac{L}{C}} = \frac{\omega_0}{\Delta\omega} \tag{13.112}$$

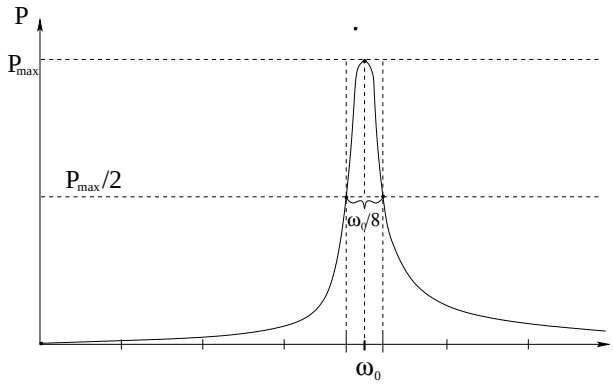


Figure 13.20: An illustration of how to "hand draw" a resonance curve for $\mathcal{Q} = 8$

(note well the use of R', not R) used above to establish the dimensionless forms. The best way to go about this is to determine $P_{\max}$ and ω_0, set $P_{\text{avg}}(\omega) = \frac{1}{2}P_{\max}$, and solve for the two values of ω *near* ω_0 for which this is true, following the hints given in the problem.

Before moving on, let's look one more time at the so-called power factor, as you're likely to encounter it as "important" in other textbooks. Recall that we can write average power delivered *to the total resistance* R' *above* as:

$$P_{\text{avg},R'} = \frac{1}{2}\frac{V_0^2}{Z^2}R' = \frac{1}{2}\frac{V_0^2}{Z}\cos\delta = \frac{V_{\text{rms}}^2}{Z}\cos\delta = V_{\text{rms}}I_{\text{rms}}\cos\delta \tag{13.113}$$

where:

$$\cos\delta = \frac{R'}{Z} \tag{13.114}$$

and where we have introduced the ***root-mean square voltage and current***:

$$V_{\text{rms}} = \frac{1}{\sqrt{2}}V_0 \approx 0.7V_0 \qquad I_{\text{rms}} = \frac{V_{\text{rms}}}{Z} = \frac{1}{\sqrt{2}}I_0 \approx 0.7I_0 \tag{13.115}$$

The "rms" forms of voltage and current let us drop the pesky factor of $1/2$ that frequently arises from averages in harmonic circuits (and in other harmonic contexts for that matter) and leaves us with quantities that are "pre-averaged" in a particular way so that they look more like their direct current counterparts, easier to remember[209].

Note well that the power factor is clearly defined in terms of the *total* resistance $R' = r + R$ used in Z' as well, not in terms of the load resistance R only. We already know how this power is split up between r and R from the voltage divider results above. We've also noted that the form containing $\cos\delta$ is *not convenient* for the purposes of computing or graphing the average load power[210]. What *does* $\cos\delta$ in this new form tell us?

Well, for one thing, it tells us that in a series LRC circuit, knowing the rms average (or peak) voltage applied *across* the circuit and the separate rms average (or peak) current per se are *not enough* to tell us what the average load power is. This is because the voltage peak

[209]These are not uncommonly encountered in the context of household wiring – when an "average current" delivered to an appliance is given so it can be used in estimating circuit capacity, it like as not is given as an rms current because the average current per se is zero.

[210]...because $\cos\delta = \dfrac{R'}{Z} = \cos\left\{\tan^{-1}\left(Q^2(\beta - \beta^{-1})^2\right)\right\}$ and we still have to evaluate either Z or Q and ω_0 or both!

may not occur at the same time as the current peak! This situation is *common* in LRC circuits – it is a part of what they do, why they exist! Let's examine three different possible limiting cases for the power factor and see how this works.

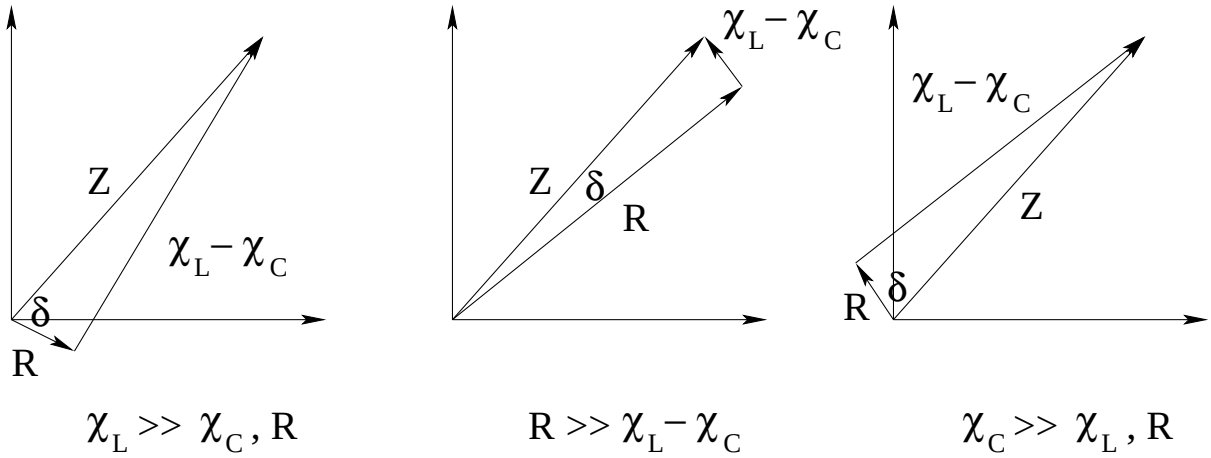


Figure 13.21: Three possible impedance diagrams corresponding to first, $\chi_L \gg \chi_C, R'$ so that $\delta \sim \frac{\pi}{2}$; second, $R' \gg \chi_L - \chi_C$ so that $\delta \sim 0$; and third, $\chi_C \gg \chi_L, R'$ so that $\delta \sim -\frac{\pi}{2}$.

Consider the three possible triangles for Z' drawn in figure 13.21. In the first one, we imagine that $\chi_L \gg \chi_C$ *and* $\chi_L \gg R'$. We are basically *far from resonance* on the *high frequency side* so that the impedance – the equivalent of the effective total "resistance" of this circuit at this frequency – is *large and dominated by induction*. That is in every cycle *most* of the energy provided to the circuit is stored in the inductor (and then given back!) and very little is used in the resistive load or adds energy to the capacitor. Note that:

$$\lim_{\chi_L \to \infty} \delta = \tan^{-1} \frac{\chi_L - \chi_C}{R'} \Rightarrow \delta \sim \tan^{-1} \infty = \frac{\pi}{2} \tag{13.116}$$

In this case the power factor is close to zero, so *independent of the individual magnitudes of* V_0 *or* I_0 little power is delivered to the circuit – the voltage and current are pretty much out of phase, with the voltage *leading* the current as is the case for inductances alone.

In the second figure, $\chi_L - \chi_C \ll R'$ (whatever their absolute magnitudes). In this case the energy provided by the voltage is *mostly* delivered as Joule heating of the total resistance of the circuit R', ideally (mostly) in the form of useful work delivered tothe load resistance; only a little goes into from the applied voltage into the capacitor or inductor (while they bounce whatever energy they have back and forth between them, not taking or giving much back to the applied voltage or resistor).

$$\lim_{R \to \infty} \delta = \tan^{-1} \frac{\chi_L - \chi_C}{R'} \Rightarrow \delta \sim \tan^{-1} 0 = 0 \tag{13.117}$$

and the power factor is close to ***one***. We are in, or "close enough to" resonance, and the current and voltage are nearly *in phase*. Again there is no guarantee that the power delivered in resonance will be *large* – that depends on how large R' is as in this limit $Z \approx R'$ – it just means that the impedance is mostly resistive (another way of saying the same thing) so that the load is getting all the power it reasonably can.

Finally, the third figure is "like" the first, only now it is $\chi_C \gg \chi_L$ and $\chi_C \gg R'$. This is likely to be true in the *low* frequency limit as $\omega \to 0$, far from resonance. Most of the power now

goes into and out of the *capacitor*, with little being dissipated by the resistor or going into the inductor. The phase angle is now:

$$\lim_{\chi_L \to \infty} \delta = \tan^{-1} \frac{\chi_L - \chi_C}{R'} \Rightarrow \delta \sim \tan^{-1} -\infty = -\frac{\pi}{2} \tag{13.118}$$

The power factor is again close to zero so that little steady state power is delivered to the circuit – the voltage and current are out of phase, with the voltage *lagging* the current as is the case for capacitances alone. The total circuit impedance is dominated by the *capacitive reactance*.

From these figures, we can deduce the importance of the power factor in circuit design. In all three cases, the *useful* work in the circuit is likely to be associated with the average power delivered to the ***total resistive load*** $\boldsymbol{R'}$. The maximum possible average power will always be given by:

$$P_{\max} = \frac{1}{2} V_0 \times I_{\max} = \frac{1}{2} \frac{V_0^2}{R'}$$

– we can't deliver *more* total power to the resistors than they would get if there were no L's or C's in the circuit (so $Z' = R'$) or in perfect resonance (ditto). However, as we've seen, V and I can be far enough out of phase that the product of their peak magnitudes is no longer close to the true average power associated with useful work – power is indeed flowing into and out of the circuit, it just isn't making it, on average, into the resistive load. In this case one has to provide a significantly higher voltage amplitude to reach some target power delivery than one would if it were applied directly to the load resistance.

Why then, one might ask, do we not just leave out capacitors and inductors and apply our voltage directly to the resistive load? Well, we *do* this when we can – for example, running an electric resistive space heater, or using an old-fashioned incandescent light bulb – but some of the things that do useful work or perform useful tasks – notably *electric motors* and *transformers* – are based on magnetic coils with many turns. They have, in other words, a *large inductance* L, a nonzero actual resistance from the many turns of wire in their coils, and a kind of "resistance" associated with the work done in association with their purpose. To discuss this in any detail requires a look at impedance matching for nontrivial source impedance, which as we noted even before starting this topic, is beyond the scope of this course.

Next, let's study the *parallel* LRC circuit.

13.4.5: The Parallel LRC Circuit

The parallel LRC is circuit drawn in figure 13.22 for the "realistic" case where the power supply has a non-zero internal resistance that may well not be small relative to the load resistance R similar to what we did for the series LRC circuit above.

In the case of the series circuit, the internal resistance just became part of the total resistance, allowing us to fairly easily solve the equation of motion for the current in the single loop in terms of the overall impedance of the circuit including r. However, including r in our real/trig based phasor analysis in *this* case makes the parallel problem, um – let's just say "difficult"[211]. On the other hand, if we analyze it with the complex methods worked through in the next section or two, it is entirely tractable (which is an excellent reason to use them instead of

[211]...and leave out modifiers like "insanely".

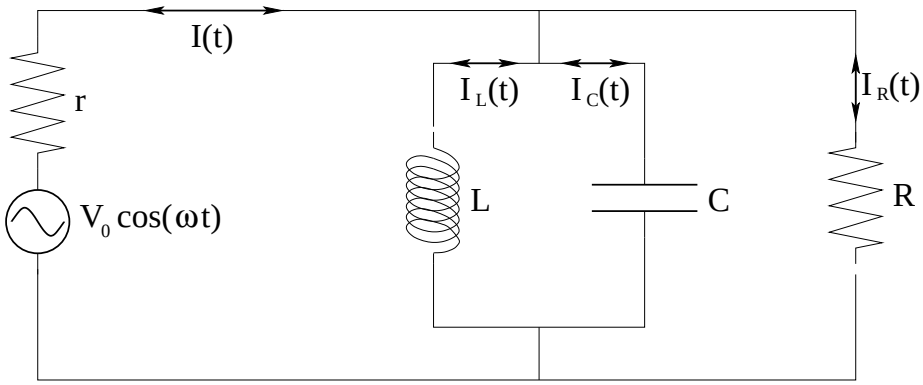


Figure 13.22: A parallel LRC circuit, for a non-ideal power supply with internal resistance r.

the phasor+trig approach for *all* circuit problems, as is pretty much universally done by people professionally working with electrical circuit design)!

To make it tractable, I'll give the results of that analysis in a useful form as that can be understood and even remembered without recourse to actually *working it out* with complex algebra. The net result is that if there is an internal resistance r in the power supply, one can compute an effective impedance Z' based once again on an equivalent resistance R'. This is good news! The bad news is that the resulting impedance can only be used *directly* in an expression for the voltage drop V_p across the three parallel elements in the circuit in terms of the applied voltage amplitude V_0. Still, this is sufficient for us to get the currents in all four elements, the relations between voltage and current for each element, and the power delivered to each element as a function of time or on average.

A ***very important difference*** between the series and parallel result is that the correct effective resistance to use in computing Z' is the effective resistance of $\boldsymbol{r}$ ***and*** $\boldsymbol{R}$ ***in parallel, not in series***, often written in electronics textbooks as:

$$r \parallel R = R' = \frac{rR}{r+R}$$

or the like.

It is difficult to come up with a simple heuristic explanation of why this should be so[212], but it is the unambiguous result of the complex analysis. So – for once without any attempt to derive them *here* – are the results of the complex algebra worked through in detail in the next optional/advanced section of the chapter. I wish I could do better, but I can't.

Let:

$$V_p = V_0 - Ir$$

be the common voltage drop across the three active circuit elements in parallel. Note well that I am ignoring whatever is happening to the phase because that's exactly the part that complex algebra handled automatically and that is so difficult to represent with real equations plus trig[213]. The amplitude of V_p in terms of the amplitude of V_0 and known parameters ***turns***

[212]For me, anyway. Anyone reading this who wants to send me one, feel free!

[213]Not *that* surprising – remember, we had a similar difficulty with the *passive* LRC circuits earlier in this chapter.

out to be:

$$V_p = \frac{V_0}{r}\left(\left\{\frac{1}{R}+\frac{1}{r}\right\}^2 + \left\{\frac{1}{\chi_C}-\frac{1}{\chi_L}\right\}^2\right)^{-1/2} = \frac{Z}{r}V_0 \tag{13.119}$$

Note the occurence of the inverse of:

$$\frac{1}{R'} = \left(\frac{1}{R}+\frac{1}{r}\right) = \left(\frac{r+R}{rR}\right) \quad \Rightarrow \quad R \parallel r = R' = \frac{rR}{r+R} \tag{13.120}$$

(squared) in this equation.

Inverting and filling in the phase (also obtained from the complex algebra we omit here) we get:

$$Z = \frac{1}{\sqrt{\frac{1}{R'}^2 + \left(\omega C - \frac{1}{\omega L}\right)^2}} = \frac{R'}{\sqrt{1+\left(\omega R'C - \frac{R'}{\omega L}\right)^2}} = \frac{R'}{\sqrt{1+\mathcal{Q}^2\left(\frac{1}{\beta}-\beta\right)^2}} \tag{13.121}$$

and:

$$\delta = \tan^{-1}\left(\omega R'C - \frac{R'}{\omega L}\right) = \tan^{-1}\left\{\mathcal{Q}\left(\frac{1}{\beta}-\beta\right)\right\} \tag{13.122}$$

and (putting in the correct time dependence including the phase of the voltage across R and the current through R, which have to be *in phase*):

$$V_p(t) = \frac{Z'}{r}V_0\cos(\omega t - \delta') \quad \Rightarrow \quad I_R(t) = \frac{V_p(t)}{R} = \frac{V_0}{R}\frac{Z'}{r}\cos(\omega t - \delta') \tag{13.123}$$

where for the moment I'm *only* giving the current through the load resistance. The instantaneous power is then

$$P_R(t) = V_p(t)I_R(t) = \frac{Z'^2V_0^2}{r^2R}\cos^2(\omega t - \delta') \tag{13.124}$$

and we can do the time average with the stroke of a pen. Continuing to work to put the result in a suitably dimensionless form, the average load power is:

$$\begin{aligned} P_{\text{avg},R} &= \frac{V_0^2}{r^2R} \times \frac{R'^2}{1+\left(\omega R'C - \frac{R'}{\omega L}\right)^2} \\ &= \frac{1}{2}\frac{V_0^2}{r^2R} \times \frac{\frac{r^2R^2}{(r+R)^2}}{1+\left(\omega R'C - \frac{R'}{\omega L}\right)^2} \\ &= \frac{1}{2}\frac{V_0^2R}{(r+R)^2} \times \frac{1}{1+\left(\omega R'C - \frac{R'}{\omega L}\right)^2} \end{aligned} \tag{13.125}$$

We now use the *same* conversions of this form to one involving:

$$P_{\max,R} = \frac{1}{2}\frac{V_0^2R}{(r+R)^2} \qquad \omega_0 = \sqrt{\frac{1}{LC}}$$

and

$$\mathcal{Q} = \frac{L\omega_0}{R'} = \omega_0\tau_{LR'} = \frac{1}{\omega_0R'C} = \frac{1}{\omega_0\tau_{R'C}}$$

used in the discussion of current and power in the series LRC circuit. ***Surprisingly enough***[214] we get:

$$P_{\text{avg},R} = P_{\text{max},R}\frac{1}{1+\left(\omega R'C - \frac{R'}{\omega L}\right)^2} = \frac{P_{\text{max},R}}{1+\mathcal{Q}^2\left(\frac{1}{\beta}-\beta\right)^2} \tag{13.126}$$

Amaingly, the average power delivered to the load resistance R has the *exact same dimensionless form* and *scales identically with* $\mathcal{Q}$ as we obtained for a series LRC circuit with an internal resistance r. We don't even have to regraph it! The curves we drew before *still work*, except for one tiny change – the R' in the expressions for Z, $\mathcal{Q}$, and δ is now the effective *parallel* resistance:

$$R \parallel r = R' = \frac{Rr}{R+r}$$

and even the *total average power* has the same form and is clearly divided in *exactly the same way* as it was for a series LRC circuit.

It is left as a fairly easy exercise for the interested student to prove that this result makes sense in the $r = 0$ limit, where there is no "resonance" per se as the power delivered to the load resistance is:

$$P_{\text{avg},R} = \frac{1}{2}\frac{V_0^2}{R} = \frac{V_{\text{rms}}^2}{R}$$

independent of $\boldsymbol{\omega}$***,*** $\boldsymbol{L}$ ***and*** $\boldsymbol{C}$ as one expects when the voltage across R (and the other two active circuit elements) is *exactly* $V_0\cos(\omega t)$ (because there is no voltage drop across an internal resistance r). The second case – $r \neq 0$ and $R = 0$ – is a bit trickier and omitted here.

There is one very important step in this that I'll note now. Recall that we found:

$$Z = \frac{R'}{\sqrt{1+\mathcal{Q}^2\left(\frac{1}{\beta}-\beta\right)^2}}$$

For a series circuit, recall, we had $R' = r + R$ and found that $\lim_{\beta\to 0,\infty} Z(\beta) = \infty$, the impedance *diverged* for very large or small frequencies, cutting off the power or current, and was always *greater than or equal to* R'.

Here the exact opposite holds, as one might *expect* from our studies of resistors in parallel, where the total effective resistance was *smaller than the smallest* of those resistors. Now:

$$\lim_{\beta\to 0} Z = 0 \qquad \lim_{\beta\to\infty} Z = 0 \quad (!) \tag{13.127}$$

What gives? Consider: If one is (say) at a very low frequency, the inductive reactance is low, say $\chi_L = \omega L \ll R', \chi_C = 1/\omega C$. The inductance then *shorts out the load resistor*, shunting all of the current from the power supply through the inductance and causing all of the voltage drop in the circuit to occur *across the internal resistance* r*!* At high frequencies $\chi_C = 1/\omega C \ll R', \chi_L = \omega L$ the same thing happens but now it is the capacitor that shorts out the load resistor!

The two circuits thus end up having an identical power curve as far as power delivered to the load resistance is concerned, but ***very, very different ones as far as power delivered to the internal resistance*** $\boldsymbol{r}$***!***

[214] To me, at least, the first time I worked it out...

In a series circuit, the power supply delivers little or no power to the circuit far from resonance. Basically, the current delivery to the whole circuit is blocked by the reactances except near resonance. If one is paying for fuel used in that power supply, well, it doesn't use much except when it matters. The power supply also remains *cool* when far from resonance with little energy wasted as heat internally.

In a parallel circuit, the power supply delivers its ***maximum possible power*** when it is far from resonance! However, it is all burned in the power supply's internal resistance! If it were fuel driven, or was prone to get hot and catch fire, this would be bad! However, if it is an *antenna*, the maximum current/power it can deliver is *tiny*, making a parallel LRC circuit one of the favorite ones to use when building an e.g. crystal radio receiver, or certain high/low pass filters!

In other words, the two kinds of circuits have very different optimal application circumstances. Keep this in mind when we look at high pass and low pass filters later!

Before moving on, I'll go ahead and plot this for a few more (somewhat different) values of $\mathcal{Q}$ on our usual, scalable coordinate frame. The rules for drawing these resonance curves are clearly the same as they were before, save for the use of a different R' in $\mathcal{Q}$.

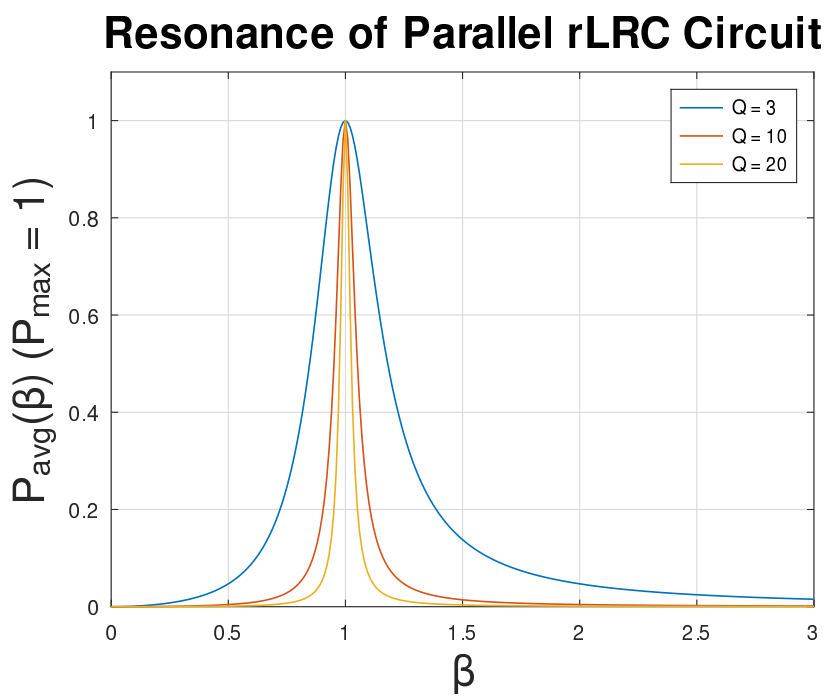


Figure 13.23: A typical series of parallel $rLRC$ resonance curves for $Q' = 3, 10, 20$, plotted on a scale such that $\omega_0 = 1$ and the peak load power $P_{\max} = 1$.

13.5: Filter Circuits

In the previous section, we examined (and in one case, explicitly derived) expressions for the power for both parallel and series AC circuits for the useful case where the power supply has a nonzero internal resistance r as well as a "load resistance" R to which power is intended to

be delivered.

As noted several times in the text, circuits of these general forms behave like *filters* in AC circuit design – ***high pass filters*** that deliver significant power to the load only for frequencies above some cutoff frequency, ***low pass filters*** that deliver significant power to the load only for frequencies below some cutoff frequency, and ***band-pass filters*** that deliver significant power to a load only in a (usually fairly narrow) band of frequencies around a central *resonant* frequency. The previous section basically examined two different ways of building band-pass filters!

In this section we'll do a simple tabulation of the "archetypical" circuits that do each of these and discuss a few of their advantages or disadvantages in circuit design, followed by an actual application to the case of a "classic" AM radio receiver (a so-called "crystal radio").

To do this we'll start with the expressions for the power we obtained in the previous section that is delivered to the *load resistance* R *only* in a form that leaves both L and C explicitly in the result so that we can most easily graph them to obtain their filter behavior in circuits that contain only (say) r, R and C but no L or vice versa. Here they are:

- For a series "$rLRC$" circuit we have:

$$P_{R,\text{avg}} = \frac{1}{2}\frac{V_0^2 R}{R'^2 + (\omega L - \frac{1}{\omega C})^2} \tag{13.128}$$

  where:

$$R' = r + R$$

  (the total resistance *in series* in the circuit) and:

- For a parallel "$rLRC$" circuit we have:

$$P_{R,avg} = \frac{1}{2}\frac{V_0^2 R}{\frac{1}{R'^2} + \left\{\omega C - \frac{1}{\omega L}\right\}^2} \tag{13.129}$$

  where:

$$R' = \frac{rR}{r + R}$$

  (the total resistance *in parallel* in the circuit).

Let's proceed.

13.5.1: Low-Pass Filter $rLRC$ Circuits

There are two simple circuit designs that are suitable for a low pass filter. The choice of which one to use can be quite involved (and beyond the scope of this course) but at least *one* of the major factors in the choice (already introduced in the previous section) will be discussed in context as we go.

As it turns out, we can either use a series circuit with *no capacitor*, or a parallel circuit with *no inductor*. Let's see how each of these works.

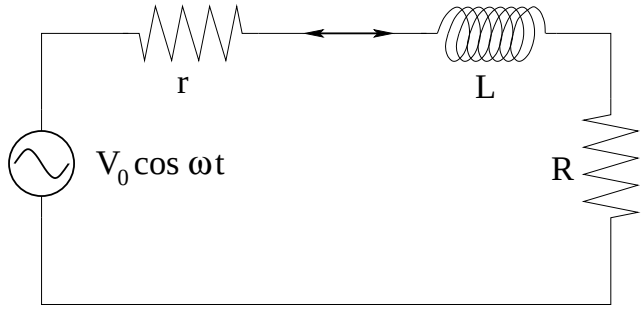


Figure 13.24: A series rLR circuit used as a low-pass filter.

We'll start with the power supply (including r, its internal resistance), an ***inductor*** $\boldsymbol{L}$, and the load resistance R all in ***series***. This circuit is pictured above. Recall that the inductive reactance is $\chi_L = \omega L$ which *increases as* ω *increases!* This in turn increases the impedance of the circuit and suppresses the current, reducing the power delivered to the load! Easy enough to conceptually understand...

The average power delivered to R only in this "series low pass" (SLP) circuit is (from equation 13.128 above, with "no C" – which algebraically corresponds to $C \to \infty$ and hence $\chi_C = 0$) given by:

$$P_{\rm SLP} = \frac{1}{2}\frac{V_0^2 R}{R'^2 + (\omega L)^2} = \left\{\frac{1}{2}\frac{V_0^2}{R'^2}R\right\} \times \left(\frac{1}{1 + \frac{(\omega L)^2}{R'^2}}\right) = \frac{P_{R,\max}}{1 + \frac{(\omega L)^2}{R'^2}} \tag{13.130}$$

where:

$$P_{R,\max} = \frac{1}{2}\frac{V_0^2}{R'^2}R$$

is the maximum average power delivered to the load resistance R! and where:

$$\frac{1}{1 + \frac{(\omega L)^2}{R'^2}}$$

can thought of as a ***dimensionless filter function***, one that (as you can see below) goes to zero for large ω, to one as $\omega \to 0$, and in between smoothly connects the two behaviors.

Obviously, this circuit *has no resonant frequency* ω_0 – or if you prefer, is "resonant" at $\omega_0 = 0$ where the delivered power *does* peak – – so our previous definition of a dimensionless frequency $\beta = \omega/\omega_0$ is impossible. At the same time, using a rescalable dimensionless form for the power curve is really, really *useful*, so let's give it a try.

Note that the equivalent of our "full width at half maximum" occurs when $P_{\rm SLP}(\omega) = P_{\max}/2$. This, in turn, occurs when:

$$\frac{\omega L}{R'} = 1$$

We should recognize L/R' as the $\boldsymbol{LR}$ ***exponential time constant*** $\boldsymbol{\tau}$ we obtained in chapter/week 8! This suggests that we should define the dimensionless parameterization of the angular frequency to be:

$$\psi = \omega\tau = \frac{\omega L}{R'} = \frac{\omega L}{r + R} \quad \Rightarrow \quad \frac{1}{1 + \frac{(\omega L)^2}{R'^2}} = \frac{1}{1 + \psi^2}$$

and write:

$$P_{SLP}(\psi) = P_{R,\max}\frac{1}{1+\psi^2} \tag{13.131}$$

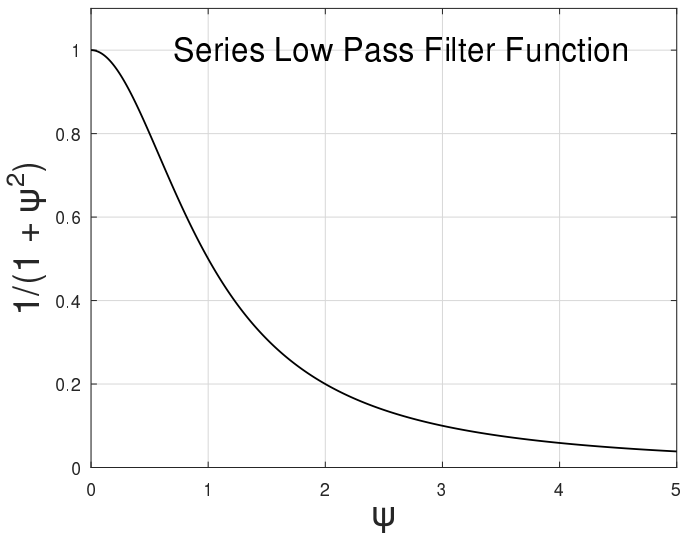


Figure 13.25: A "generic" power graph of series rLR circuit used as a low-pass filter. Note well that this figure is *identical* to the one corresponding to a low pass rRC circuit in *parallel*!

We can now display the "universal" filter function this circuit represents in dimensionless form. It is plotted in figure 13.25. This graph is now *completely generic!* All we have to do to make it work for *any* SLP circuit like this one is to a) scale the vertical axis so that the peak at 0 frequency has the right value; and b) scale ψ back into $omega = \psi/\tau$ with $\tau = L/R'$! Nothing to it!

There are two things to note about this kind of low pass filter. First, as is clear from the graph, significant power is delivered to the load resistance only when:

$$\psi < 1 \quad \Rightarrow \quad \chi_L < R'$$

When the inductive reactance is larger than the total resistance, it quickly cuts off the current as we anticipate earlier. When it is less than the total resistance, power flow to the load is not significantly impeded.

Second, this is a sensible filter circuit to use if it is undesireable for the power supply to heat up when it isn't actually delivering power to the load. In the cut-off high frequency region, the total current goes down, so resistive heating in the power supply goes down. Basically *no* power is used in either r or R for very high frequencies! If $r = R$ (impedance matching!) we expect to get the maximum possible power into the load at $\omega = 0$.

There is a second way to build a low pass filter. One can make a ***Parallel Low Pass*** (PLP) filter by putting a capacitor in ***parallel*** with the load resistance as shown in figure 13.26. When ω is small, the capacitative reactance $\chi_C = 1/(\omega C)$ is *large*, more or less blocking current so that it must flow through the load resistor. When ω is large, however, the capacitor *shorts out the load resistor* as its effective "resistance" becomes much smaller than R! A large current then flows out of the power supply (through its internal resistance r) but it *bypasses* the load resistor, delivering little or no power to it as the voltage across it drops.

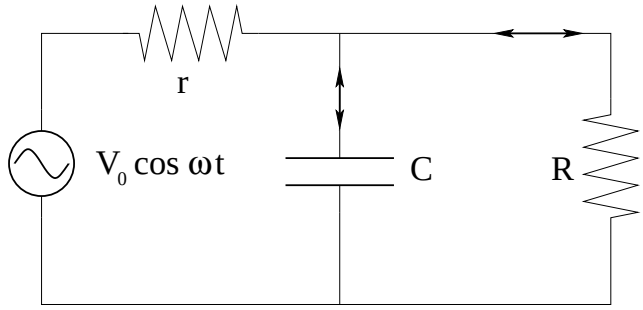


Figure 13.26: A parallel rRC circuit used as a low pass filter.

The power delivered to the load resistor (only) is given by (following equation 13.129 above, with $L = 0 \quad \Rightarrow \quad \chi_L = 0$):

$$P_{\rm PLP} = \left\{ \frac{1}{2} \frac{V_0^2 R}{(r+R)^2} \right\} \times \left(\frac{1}{1+(\omega R' C)^2} \right) = \frac{P_{R,\max}}{1+(\omega R' C)^2} \tag{13.132}$$

where:

$$R' = \frac{rR}{r+R}$$

is the effective resistance of r and R in *parallel* (even though it is far from obvious that this should be the case from looking at the circuit itself).

Examining this, we see that if we define the dimensionless:

$$\psi = \omega\tau$$

but now with $\tau = R'C$ the exponential time constant of the *capacitative* RC circuit we get:

$$P_{PLP} = P_{R,\max} \frac{1}{1+\psi^2} \tag{13.133}$$

which is *identical* to the low-pass filter function result we obtained with the SLP circuit above! I don't even need to re-graph it! The only real change is that the crossover occurs when the *capacitative* reactance equals the load resistance, not the *inductive* reactance.

There is, however, a *critically* important difference, one we've spoken of before! When $\chi_C \to 0$ as $\omega \to \infty$, the current amplitude provided by the power supply tends towards its maximum possible value:

$$I_{\max} = \frac{V_0}{r}$$

so that all of the voltage drop occurs across its own internal resistance! The average power that appears there is also a global maximum at:

$$P_{\rm avg,max} = \frac{V_0^2}{2r}$$

If the power supply is being driven by e.g. a heat engine of some sort, this basically means that it is running at its maximum capacity and generating lots of heat for no useful work! However, there are a number of circumstances in circuit design where parallel filter circuits are useful. As we'll see below, *low* voltage *low* power sources may benefit from circuits of this general type.

A second application of circuits of this sort is to "clean up" a DC power supply. A large capacitor-based low pass filter is built into most DC power supplies in order to "short out" high frequency voltage surges that arise when e.g. a switch is closed and arcing causes sharp spikes in the voltage, or when lightning hits a power line not too far away sending a similar very short duration pulse down the line. They waste less energy in normal operation by completely blocking DC voltage and it's also somewhat easier to get good, accurate capacitors than it is good inductors. For that reason this is probably the most common form of low-pass filter.

13.5.2: High-Pass $rLRC$ Circuits

Well, if LR in series is a low pass and RC in parallel is low pass, a pretty good (and correct!) guess is that LR in parallel is a high pass and RC in series is a high pass. Indeed, this would be even more obvious if we wrote out the power in terms of the reactances of these elements – one simple swaps the inductive reactance for capacitive and vice versa.

Nevertheless, we'll work through it. The RC version of the high pass filter is drawn in figure 13.27:

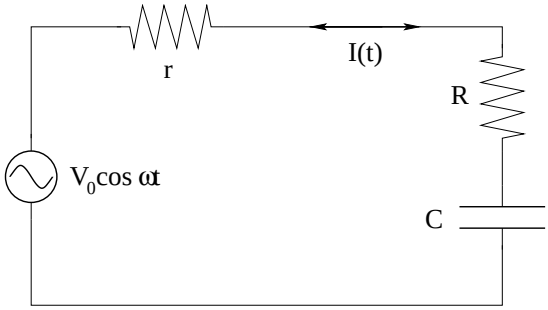


Figure 13.27: High pass rRC circuit with all elements in *series*.

The ritual we must follow should now be pretty obvious. We write out expression for the average power delivered to the load from our series $rLRC$ circuit result equation 13.128 above (with $L = 0 \quad \Rightarrow \quad \chi_L = 0$), express it in terms of the divided r vs R power, make the argument dimensionless, and get:

$$P_{R,\text{avg}} = \frac{1}{2}\frac{V_0^2 R}{R'^2 + \frac{1}{(\omega C)^2}} = \frac{1}{2}\frac{V_0^2 R}{(r+R)^2} \times \frac{1}{1 + \frac{1}{(\omega R' C)^2}} = \frac{P_{R,\max}}{1+\psi^{-2}} \tag{13.134}$$

where $R' = r + R$ is the total circuit resistance in *series*. If we let:

$$\tau = R'C \quad \Rightarrow \quad \psi = \omega\tau$$

with τ the exponential time constant of the RC part of the circuit as we did in the previous section this becomes the maximum (divided) power delivered to the load resistor times the ***Series High Pass*** (SHP) dimensionless filter function:

$$P_{rRC} = P_{R,\max}\frac{1}{1+\psi^{-2}} \tag{13.135}$$

If we multiply this on the top and bottom by ψ^2 this becomes slightly easier to evaluate and graph without worrying about dividing by zero:

$$P_{rRC} = P_{R,\max}\frac{\psi^2}{1+\psi^2} \tag{13.136}$$

The filter function itself is as before "generically" plotted in figure 13.28:

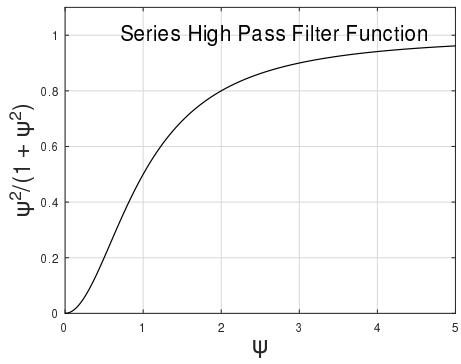


Figure 13.28: A "generic" graph of the dimensionless series rRC circuit used as a high pass filter (SHP). Note well that this graph is *identical* to the one corresponding to a high pass rLR circuit in *parallel* (PHP)!

Note that the figure is literally the low pass power function upside down (or subtracted from 1 if you prefer). As before it is scaled so that the maximum possible load power is 1 and so that the cut-off frequency is 1 – the single graph will work for any similar circuit if one simply scales the axes units appropriately to match.

Finally, we'll repeat one last time for the parallel rLR circuit portrayed in figure 13.29. The average load power in this case is:

$$P_{rLR} = \frac{1}{2}\frac{V_0^2 R}{(r+R)^2}\frac{1}{1+\frac{R'^2}{(\omega L)^2}} \tag{13.137}$$

where

$$R' = \frac{rR}{r+R}$$

is the equivalent resistance of r and R in *parallel*. Again we define:

$$\tau_{LR'} = \frac{L}{R'} = \frac{L(r+R)}{rR} \quad \Rightarrow \quad \beta = \omega\tau_{LR'}$$

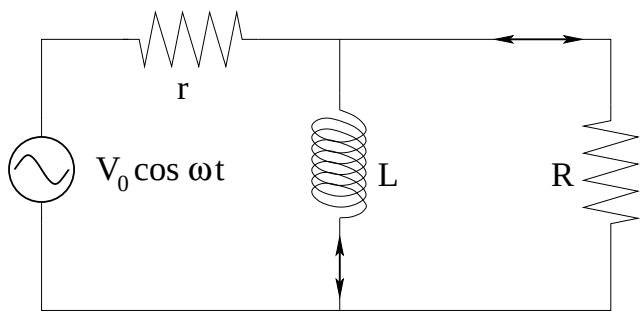


Figure 13.29: High pass rLR circuit with all elements in *parallel*.

so that:

$$P_{rRC} = \frac{1}{2}\frac{V_0^2 R}{(r+R)^2}\frac{1}{1+\frac{1}{\beta^2}} = \frac{1}{2}\frac{V_0^2 R}{(r+R)^2}\frac{\beta^2}{1+\beta^2} \tag{13.138}$$

which is identical to the result already plotted above.

As before, there are probably more advantages to using the capacitive form of the high pass filter instead of the inductive version simply because one doesn't have to mess with the resistance and heating of the inductance. However, specific circuits where the power supply or load itself has some inductive or capacitive character (which can easily occur if the power supply is e.g. a transformer of some sort or when the load is a motor) can only be perfectly balanced by introducing the corresponding correction to the circuit impedance as indicated in the Maximum Power Theorem, in which case the inductive version might be called for. Stuff like this is why EE's often make big bucks – it's not for the faint of heart.

13.5.3: Band Pass Filters

Well if series rLR circuits are low pass filters, and series rRC circuits are high pass filters, what's a series $rLRC$ circuit? A *band-pass filter!* It prevents the power supply from delivering power to the load resistance when $\omega \ll \omega_0$ *or* when $\omega \gg \omega_0$. Only when $\omega = \omega_0 = 1/\sqrt{LC}$ does the circuit allow the maximum current to pass (in phase with the power supply) and hence deliver maximum power to the load resistor.

Generic/scaled plots of series band-pass filters for various values of Q are already plotted above and will not be repeated here. I'll only note that in:

$$P_{\text{series}} = \frac{V_{\text{rms}}^2 R\omega^2}{R'^2\omega^2 + L^2(\omega^2 - \omega_0^2)^2} \tag{13.139}$$

don't forget that the impedance Z^2 in the denominator contains $R'^2 = (r + R)^2$, not R^2. This will only be important if r is on the same order of magnitude as R – more often $R \gg r$ and one can safely ignore r.

That's not so clearly the case for parallel $rLRC$ circuits!

$$P_{\text{parallel}} = \frac{1}{2}\frac{V_0^2 R}{(r+R)^2}\frac{1}{1 + R'^2C^2\omega^2\left\{1 - \frac{\omega_0^2}{\omega^2}\right\}^2} \tag{13.140}$$

where $R' = rR/(r+R)$ is the equivalent *parallel* resistance. In this case the entire rationale for the functioning of the circuit *as* a band pass filter *depends* on $r \neq 0$! When $r = 0$, the average power delivered to the load resistance is a *constant independent of* ω*!*

We need to think of this circuit as one where at very low frequencies the (presumed zero resistance!) inductance *shorts out the load resistor* and causes all of the voltage drop to occur within the *source* resistance r! At very high frequencies, it is the capacitor that shorts out the load resistor. In resonance, however, the inductance and capacitance have *equal and opposite* reactances so that the contribute nothing to the circuit reactance and it again becomes a de-facto simple voltage divider with the power split up proportionally between r and R.

This sounds a lot more complicated than the behavior of a series $rLRC$ circuit, but it turns out to be nearly ideal for radio circuits where the "resistance" of the antenna is not generally negligible compared to the load resistance. In this case the voltage received is "relative to ground", and one can indeed think of of all of the radio frequencies one does *not* want to

receive as being shorted out to ground by either the capacitor or inductor unless the two are in resonance and effectively cancel out of the circuit altogether.

Let's take a look at a classic crystal radio circuit based on this idea and see how it all works in the case of amplitude modulated (AM) radio transmission/reception.

13.6: The AM Radio and Bandwidth

We now know enough to understand how one of the ***foundations of modern civilization*** – the *radio* – works. The circuits of certain modern, common versions of the radio (such as FM radios or microwave-based radios) are beyond our scope here[215]. However, the *simplest* (and original) way to transmit things like voice and music via electromagnetic (radio) waves is to use *Amplitude Modulation* (AM) to encode the signal onto a *carrier* wave. Here's how it works.

First one builds an *oscillator* at the fixed frequency of the carrier (which is generally a much higher frequency than any frequency in the signal). Without going into any details, the LC circuits studied above (combined with an amplifier) can be used to drive *themselves* to a stable, single frequency output (especially when stabilized with and tuned to a "natural" electrical oscillator such as a piezoelectric crystal). For our purposes this frequency doesn't have to be *too* precise – a bit of slow drift in phase or frequency is OK, for example – but we'll pretend that it is a single, pure harmonic wave at a carrier frequency ω_c.

Next, we need to collect the signal being encoded in electronic form. This is easily done with e.g. a microphone, which creates a voltage proportional to the air pressure variations that it experiences when we speak into it or play music into it. This sort of signal is called an *analog* signal (as opposed to a digital signal) that can take *any value* and that varies over time.

Third, we combine the two. We use the varying voltage from the microphone as the relatively slowly varying *amplitude* of carrier. The three signals (unmodulated carrier, modulating signal, encoded/modulated carrier) are shown in figure 13.30. The final AM encoded voltage is used as input to an *amplifier* that drives the voltage supplied to the *transmission antenna*, typically a tall radio tower being driven at a power of tens to hundreds of kilowatts. The resulting *radio signal* – electromagnetic *radiation* of the sort we will study in the next chapter – propagates for long distances at the speed of light and falls upon the *receiving antenna* of your AM radio.

There it creates an alternating voltage with the same shape as the voltage applied to the transmitting tower. However, this voltage is now *very weak* – the intensity of the radio wave diminishes with roughly the square of the distance from the radio tower – and is mixed in with many other *equally strong or even stronger* signals from other radio sources (other radio stations, the sun, electrical motors, many things create radio waves) at various frequencies.

To tune in *just* the carrier (plus enough bandwidth to allow its amplitude modulation to make it through the receiver circuit) we build a circuit that effectively *shorts out* all of the signals but the desired carrier at ω_0 by providing them with an easy path to ground through either an inductor (for lower frequencies) or a capacitor (for higher frequencies). The *simplest* circuit

[215] In part because I don't fully understand them myself well enough to teach them – no human alive can understand *everything* that there is to understand...

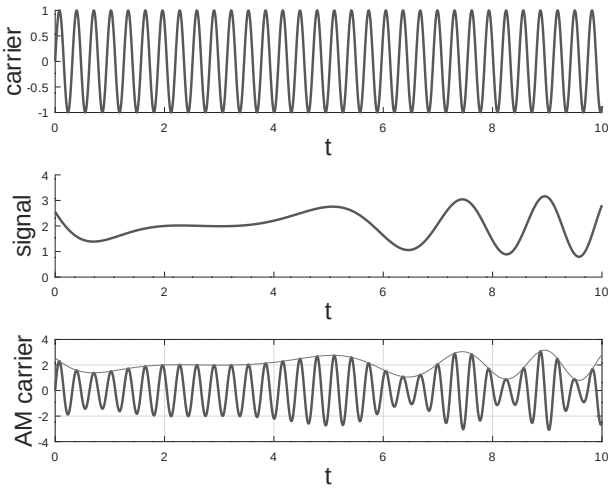


Figure 13.30: (a) The unencoded carrier with an arbitrary normalization voltage $V_c = 1$ volt and angular frequency ω_0. (b) The signal to be encoded. A *DC bias* has been added to the AM signal so that the voltage is always positive. This DC bias can be removed at the far end with a simple high-pass filter; (c) The AM encoded carrier used as (for example) the power supply to the antenna of a radio station. Note that for real AM signals the carrier frequency is much higher compared to the highest frequencies in the signal, which improves the averaging that takes place in the decoding rectifier.

that accomplishes this is our parallel LRC circuit above.

However, we have to add two features in order to make it a tunable AM radio. First is a way to tune it! We note that we do the best possible job of filtering out unwanted frequencies when the condition $\omega_0^2 = 1/LC$ and when $R = r$, so our receiver resistance/impedance matches the internal resistance of the voltage source. We therefore have to be able to adjust L, C, or both in order to tune in our AM encoded carrier.

It is beyond our scope in this work to discuss all the various aspects of this decision. The antenna, diode (crystal), headphones or amplifier input all have some impedance – characteristics of resistance, inductance and capacitance – and have to be corrected for. Also, we need to be able to tune the Q of the circuit so that the receiver bandwidth is adequate to pick up all of the encoded signal while still being narrow enough to reject nearby AM encoded stations. Many simple crystal radio designs that use wire wrapped around e.g. a simple tube of some sort allow one to vary L across a range (which adjusts ω_0 and Q simultaneously) – this is especially wise if one's headphones and/or antenna have enough capacitance already to make it difficult to add a tuning capacitor "in range" to permit tuning. Others use fixed L (and hence fixed Q) and a variable capacitor to tune. Still others may do both – allow one to vary L (possibly to one of a small set of discrete values) and then use a continuously tunable C to find the signal.

In an idealized circuit for the simplest of crystal radios in figure 13.31, I arbitrarily show a variable C (that's the arrow symbol) and also introduce the symbol for an antenna and ground. The resistance r is a *mix* of the physical resistance of the antenna wire and its "radiation resistance" and is the quantity that needs to be impedance matched (more or less) by the load

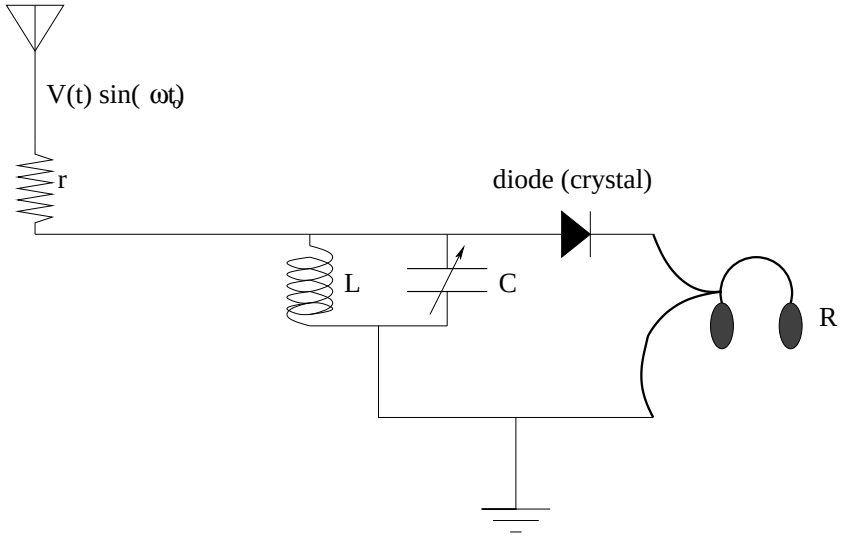


Figure 13.31: A very simple, idealized crystal radio circuit using a variable capacitor instead of variable inductance (or variable both). Note also the presence of a *diode decoder* – a one-way gate for current (which flows only in the direction of the "arrow").

R for maximum power delivery at resonance. Recall that providing an easy (low impedance) path to ground through either L or C for a given frequency will effectively short out the antenna so that all its power at that frequency will be dissipated in the antenna, not in R. Only when LC has *infinite* collective impedance at resonance will the power delivery be balanced in r and (matched) R.

This simple parallel signal *alone* would suffice to tune in the AM carrier, but if we listened to the headphones *without* the diode decoder visible in the circuit, we'd hear – *nothing!* That's because the carrier is at a very high frequency (typically over 500 kHz) that is well above the range of human hearing. We have to *remove* the carrier, leaving the signal.

Diodes act as a one-way gate for the voltage, allowing current to flow only in the direction of the "arrow" in the diode. This process is called "rectification" (literally right-sidification), and a single diode is a *half-wave rectifier*, cutting off of the *negative* parts of the current and passing only the positive "right side up" voltage/current variation. Placing a small capacitor in the line containing the headphones (usually not necessary, as the diode and the headphones together have some capacitance) removes the DC bias and "smears" out the top-half carrier waves to fill in a good approximation to the original signal.

The original diodes were *crystals* of e.g. lead galena in a mount with an adjustable wire whisker in contact with the crystal – hence "crystal radio". The wire whisker created a semi-conducting interface with the crystal that in turn only passed current in one direction (with a very high back resistance that effectively prevented it in the other). However, lots of other conductor interfaces will provide the same effect, including a graphite pencil (basis of so-called "foxhole radios" used by GIs in World War II, usually built out of surplus junk scavenged on a battlefield).

Of course using a single diode in a circuit wastes *half of the power* picked up by the incoming antenna! It is much better to use *four* diodes turned into a *full-wave rectifier*. Look over the following circuit in figure 13.32 (intended to replace the entire diode/headphone arrangement

in the circuit above) and understand how as the *voltage* oscillates positive to negative, the *current* through the headphones only passes in just one direction.

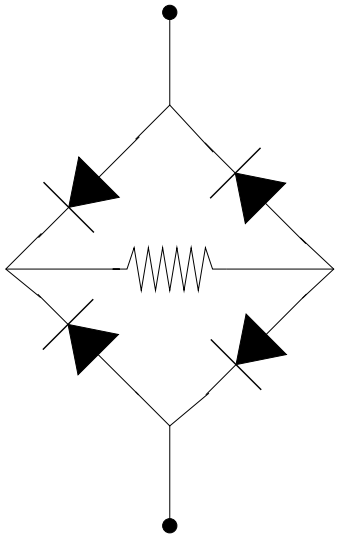


Figure 13.32: A full-wave rectifier made out of four diodes. The "headphones" are the resistance in the center of the diamond of diodes. Verify that the current always passes through this resistor from left to right, regardless of whether the voltage difference top to bottom is positive or negative.

This arrangement basically *flips* the negative half-waves and fills them into the "holes" between the positive ones, recovering the full energy. Again, when smeared out a bit by an RC time constant by the capacitance of the headphones, this accurately reconstructs the decoded AM signal, without any bias, with a bit of high frequency "ripple" that the human ear cannot hear. A schematic of the flipped (but not smeared) signal is shown below in figure 13.33. Compare it to the original signal and you can see that as long as the headphones are massive enough to be unable to respond to the very high frequency ripple *anyway*, you'll be able to hear the music, voices, or whatever that was encoded on the carrier to a high degree of accuracy.

This section should provide you with more than enough information to understand and even build a crystal radio of your own. Note well: this *general process* of encoding and decoding information on to/off of carrier signals is one of the *fundamental bases* of modern civilization. High pass, low pass, and band pass/reject circuits are ubiquitous. Even if you yourself never actually *build* an electronic circuit, knowing a bit about how they work and in particular knowing what things such as "impedance matching" are and why they matter can really improve your understanding and ability to work with electronic devices in many laboratory environments.

In this chapter we have already remarked on the content of the next one. We have learned all of Maxwell's Equations already, but one of them is *broken*; in particular, it doesn't take into account the fact that *charge is conserved* and that there is a certain *ambiguity* in the particular open surface S one can choose that is bounded by any given (specified) closed curve C. We need to fix this, adding the *Maxwell Displacement Current* to Ampere's (broken) Law.

When we do, we will discover an amazing thing: time varying electromagnetic fields satisfy the *wave equation* and hence *propagate like a wave*. Under some circumstances those waves

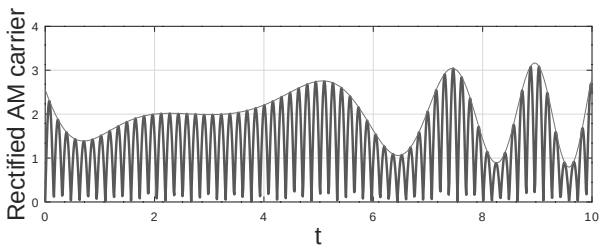


Figure 13.33: The AM encoded signal after it has been received by a tuned, band-pass filter and full-wave rectified. Note that the *average* output voltage will very closely track the original signal.

form *radio* waves, like the AM encoded carrier wave we have just studied. In others, however, those waves are what we know as *light!*

13.7: Complex Representations of LRC Circuits (Advanced)

In the sections above, I've *tried* to give at least some explanation of the simplest realistic series and parallel LRC circuits. In the case of series circuits, we could do a halfway decent job for the case where the power supply has an internal resistance, but our method of phasor diagram plus some trigonometry broke down in the case of a realistic parallel LRC circuit and I was forced to import results from *this* section without deriving them.

This problem only gets worse as we add additional circuit branches or loops, especially ones where the circuit elements theselves are not ideal. Inductors, for example, usually have a non-zero resistance as well, which means we should really put that resistance R_L in series with our "ideal" L in a real world circuit. Similarly, resistances have some capacitance, and even the *wires* used in a circuit have some inductance.

How can we handle these inconvenient realities in less-than ideal circuits? Indeed, how can we handle at least simple true multi-loop circuits, where perhaps a *mix* of AC voltages with different frequencies is applied across circuit elements in series with loops of elements in parallel?

These are not idle questions! In real circuit design, different arrangements of *precisely these forms* serve important purposes such as making high-pass filters (filters that remove all of the low-frequency voltages from the power delivered to a given "load" resistance), low-pass filters (that obviously do the opposite), band-pass filters (that select a possibly narrow range of frequencies and otherwise eliminate *both* low(er) *and* high(er) frequencies from the power delivered to the load.

Band pass filters, in particular, are at the heart of pretty much *all signal processing* in our information-rich civilization, where information (such as voice or music or digital data) is literally encoded in electromagnetic voltages or waves and transmitted in a *mix* to receivers that only want to select *one* specific "band" that contains the information – for example, a radio station – they wish to access while excluding all of the *other* radio stations broadcasting to the same antenna at the same time! Low pass filters are built into nearly all DC power supplies to ensure that they don't transmit transient (but "high frequency") pulses generated by arching in a switch or nearby lightning into delicate and expensive electronics that they would destroy. High pass filters can similarly remove "DC bias" from a circuit that might exceed the design parameters for something intended to do work with only the high-frequency AC component.

Let's begin an exploration of this more complicated world by using complex algebra and *no trigonometry at all* to obtain our results. Treatment of this topic cannot possibly be truly exhaustive in a general introductory *physics* textbook, but we can at least span the most common forms of high, low and band-pass filters. This should be "enough" to leave future EE's well-prepared for their first real electronics course (with a bit of remapping of notation as mentioned above, sorry) and at the same time leave future *physics* students well prepared for a real electronics course (yes, we have them just for our majors as we certainly use a lot of electronics in almost all of experimental physics, although we may not be designing CPUs or radios per se any more). Even math majors will benefit from seeing just how *useful* complex algebra can be in solving certain classes of differential equations, although our treatment will stop short of doing anything with e.g. Fourier Transforms.

13.7.1: Single Circuit Elements – Complex Solution

Before proceeding further we need to demonstrate how to solve the "stuff in the box" above for the currents in the case where the voltage itself is considered a ***complex exponential function of time*** and where the desired solution is – ultimately – the ***real part of this complex solution***. Instead of devoting a subsection to each element as I did before, I'll just do them all at once – as you'll see the solution is actually a bit *simpler* than it was before.

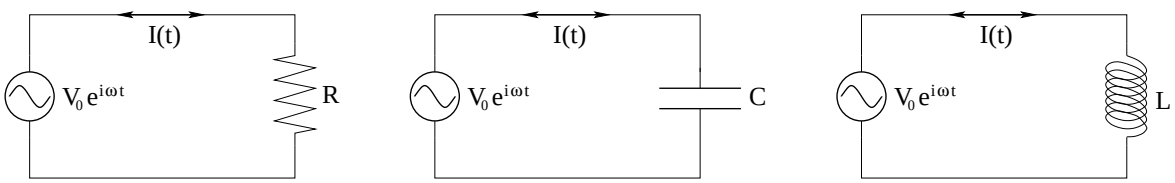


Figure 13.34: ***Complex*** AC voltages across R, L and C

From figure 13.34 we can write Kirchoff's Loop Rule three times, one for each circuit element. Remember, although we are drawing the voltage source symbol for a "power supply", in each case it stands only for the actual voltage drop across the element in parallel with – the actual *source* for the voltage is the entire circuit upstream/downstream from the wire containing the element. That is:

$$V_{\text{element}} = V_0 e^{i\omega t}$$

where element $= R$, C, or L in each sub-figure and V_0 is the *amplitude* of the complex voltage drop across the element.

For each figure, then, let's find the relation between the voltage drop across the element and the current. As before, we'll find:

$$I_C = \frac{d}{dt}CV_C(t) \quad \text{and} \quad I_L = \int \frac{1}{L}V_L(t)\,dt$$

(with no constant of integration). Thus:

$$V_R(t) = V_0 e^{i\omega t} = I_R(t)R \qquad \Leftrightarrow \qquad I_R(t) = \frac{V_R(t)}{\chi_R} = \frac{V_0}{R}e^{i\omega t} \tag{13.141}$$

$$V_C(t) = V_0 e^{i\omega t} = \frac{Q}{C} \qquad \Leftrightarrow \qquad I_C(t) = \frac{V_C(t)}{\chi_C} = i\omega C V_0 e^{i\omega t} \tag{13.142}$$

$$V_L(t) = V_0 e^{i\omega t} = L\frac{dI}{dt} \qquad \Leftrightarrow \qquad I_L(t) = \frac{V_L(t)}{\chi_L} = \frac{V_0}{i\omega L}e^{i\omega t} \tag{13.143}$$

Now, instead of having to convert sine functions back into cosine functions with an additional $\pm\pi/2$ phase as we are forced to do for the "real" phasor+trig solutions above, we simply identify the ***complex reactances***:

$$\chi_R = R \qquad \chi_C = \frac{1}{i\omega C} \qquad \chi_L = i\omega L \tag{13.144}$$

where clearly we don't need the symbol χ_R, I just included it to make the analogy clear. In any event, I will henceforth omit it in favor of the real value of R alone.

The use of a complex exponential voltage, as you can see, makes it *trivial* to express the voltages themselves in terms of the currents – the factor of i in the reactances automatically keeps track of phase shifts because of the two identities:

$$i = e^{i\pi/2} \qquad \frac{1}{i} = -i = e^{-i\pi/2}$$

This lets us easily obtain the "real" solutions we put in the box above by e.g.

$$I_C(t) = i\omega C V_0 e^{i\omega t} = \omega C V_0 e^{i\omega t} e^{i\pi/2} = \omega C V_0 e^{i(\omega t + \pi/2)}$$

The real part of this is clearly identical to the result we used above for the capacitance:

$$I_C(t) = \frac{V_0}{\chi_{C,\text{real}}}\cos(\omega t + \pi/2)$$

with $\chi_{C,\text{real}} = 1/\omega C$ or vice versa to express the voltages in terms of the currents. **Note well:** We never do the taking of the real part until the ***very end of the algebra*** because we don't *want* to mess with harmonic trig functions until we absolutely cannot avoid it.

We can can now ***completely ignore the explicit time-dependence*** $e^{i\omega t}$. It is there by assumption in every term. Indeed, the mathematically sophisticated way to obtain all of these relations is to take the fourier transform of the inhomogeneous differential equation of motion, whereupon the single-frequency results are precisely the equations and relations above! We can just write:

$$V_R = I_R R \qquad V_C = I_C \chi_C \qquad V_L = I_L \chi_L \tag{13.145}$$

These are *algebraically identical* to what we'd write down for the voltage across a simple resistance in a DC circuit, with the sole "complication" being "complexation"[216]. All of our

[216] You see what I did, there? Joke, joke.

phase information *relative* to the now ignored time dependence is tied up in the *algebraic form* of the complex reactances! All we have to do solve problems with algebra, no trig or trig identities needed, and the put the results in a more or less standard form, take the real part, and we're done! This is what makes this approach so (ultimately) simple and powerful!

Best of all, we are now no longer limited to single loop or simple parallel circuits – we can, in principle, do *any combination of elements in multiple circuit loops* the exact same way we would do them if they were all resistances in an ordinary DC circuit, except, well, for the complex bit. Time to practice up on complex algebra!

We'll start by putting all of this in a box as before and then applying "stuff in the box" to the same important circuits as before:

$$\boxed{\begin{array}{ccc} I_R = \dfrac{V_R}{R} & \Leftrightarrow & V_R = I_R R \\ I_C = \dfrac{V_C}{\chi_C} & \Leftrightarrow & V_C = I_C \chi_C \\ I_L = \dfrac{V_L}{\chi_L} & \Leftrightarrow & V_L = I_L \chi_L \end{array}}$$

Kind of boring, right? That's the *point!* We *like* boring!

Note that all of these *are* to be multiplied – at the end – by the *common factor* $e^{\imath\omega t}$ as these are equations for the *complex amplitudes* of the voltages across and current through each element. In the meantime, no differential equations to solve. No explicit time dependence. Just ***Ohm's Law*** – kind of – where we use (possibly) complex ***reactances*** instead of *resistances*.

This is why electrical engineers are forced to master the complex approach. Sure, you pay a – rather small, really – penalty at the beginning while you (re)master complex algebra, but after that, it's *all algebra*, no trig or trig identities needed! Indeed, as I've tried to demonstrate along the way, the easiest way to *derive* trig identities in many cases is using complex algebra and the Euler relation.

13.7.2: Series LRC Circuit – Complex Solution

We'll start by writing Kirchoff's Loop Rule for a series LRC circuit:

$$V_L(t) + V_R(t) + V_C(t) - V(t) = 0 \tag{13.146}$$

In this equation, $V(t)$ is *now* the actual voltage applied across the entire series collection of elements L, R, and C. Using the stuff in the *new* box in the previous section, we can write down:

$$V_L(t) = I_0 \chi_L e^{i\omega t} \qquad V_R(t) = I_0 R e^{i\omega t} \qquad V_C(t) = I_0 \chi_C e^{i\omega t} \tag{13.147}$$

and:

$$V(t) = V_0 e^{i\omega t} \tag{13.148}$$

In this, only V_0 and R are guaranteed to be real numbers; all of the other parameters (including I_0) are generally *complex*.

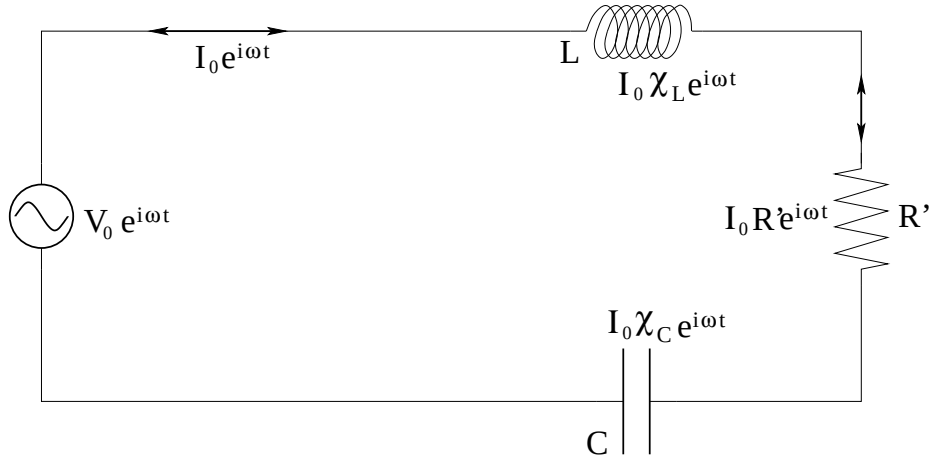


Figure 13.35: An LRC circuit, where R' is real but where the voltage, the current, χ_L and χ_C are all *complex*.

This leads us to draw figure 13.35 as a good representation of the system, including the current and the voltage drops across each element. To keep the picture and algebra as simple as possible initially while we are still (re)learning complex algebra, I've bundled up $R' = r + R$ into a single effective series resistance – we can always split it up *a la* voltage divider later. Note one last time that *everything* in the circuit has the complex time dependence of $e^{i\omega t}$ of the applied voltage; all other phase information is encoded in the complex amplitudes and reactances *relative* the phase of the applied voltage.

Life is now easy. All the hard work is already done, except for some complex algebra, which I'll do very completely and carefully so you can see how it works even if you never really learned how to do it before. We start by ***cancelling the common $e^{i\omega t}$ out of Kirchoff's Loop Rule***[217] to transform it into a single purely algebraic version of Kirchoff's Loop Rule:

$$I_0\chi_L + I_0 R + I_0\chi_C = (\chi_L + R' + \chi_C)I_0 = I_0 Z = V_0 \tag{13.149}$$

In this, I've introduced the ***complex impedance*** Z such that:

$$I_0 = \frac{V_0}{Z} \quad \Leftrightarrow \quad V_0 = I_0 Z \tag{13.150}$$

which is a fourth line that in *this* context we could add to the box above. Note that this implies that the impedance must have the ***opposite complex phase of I_0*** so that their product is, as required, real and equal to V_0. This is *already* a valuable insight, as we shall see!

Now we just *solve algebraically* equation 13.149 for (complex) Z' *directly* and then put it into the *Euler form* for a complex number. No need for calculus any more – it went *into the complex reactances of the circuit elements themselves* and we're done with it! We get:

$$Z = (R' + \chi_L + \chi_C) = R' + i\left(\omega L - \frac{1}{\omega C}\right) = |Z|e^{i\delta} \tag{13.151}$$

where (recall) we are using the total resistance $R' = r + R$ throughout even though I didn't explicitly draw r in on the figure above.

[217] Or, *project out* this term using a Fourier transform. However, if you don't know what that is yet, don't worry about it. Dividing it out is entirely permitted in this context as it is *never zero*.

We now have to find $|Z|$ and δ from this algebraic form. I'm going to be doing this – finding a complex number given as a sum of real and imaginary parts as a magnitude times a unimodular complex exponential at some phase angle – a *lot*, so I will do a lightning review of the steps here. Welcome to ***Complex Algebra 101!***

Complex Algebra 101

Given $z = x+iy$, we want $z = |z|e^{i\theta}$ instead. This is very similar to going between cartesian and polar coordinates in ordinary geometry! Recall that the ***complex conjugate*** of z is given by changing the sign of i only:

$$z^* = x - iy$$

so that:

$$z = x + iy = |z|e^{i\theta} \quad \Rightarrow \quad zz^* - |z|^2 e^{i(\theta-\theta)} = |z|^2 = x^2 - (iy)^2 = x^2 + y^2$$

or

$$|z| = +\sqrt{x^2 + y^2}$$

Then, from the Euler relation:

$$e^{i\theta} = \frac{z}{|z|} = \frac{x}{|z|} + i\frac{y}{|z|} = \cos\theta + i\sin\theta$$

with:

$$\theta = \tan^{-1}\frac{y}{x}$$

End of Complex Algebra 101

Thanks, Euler!

Now we just recapitulate these steps for Z. This time I'm going to go ahead and note *first* that:

$$\frac{\omega L}{R'} = \mathcal{Q}\frac{\omega}{\omega_0} = Q\beta \qquad \frac{1}{\omega R' C} = \mathcal{Q}\frac{\omega_0}{\omega} = Q\beta^{-1}$$

(as explicitly shown earlier) so we can write:

$$\begin{aligned} Z^2 = ZZ^* &= \left\{R' + i\left(\omega L - \frac{1}{\omega C}\right)\right\} \times \left\{R' - i\left(\omega L - \frac{1}{\omega C}\right)\right\} \\ &= R'^2 + \mathcal{Q}^2\left(\beta - \frac{1}{\beta}\right)^2 \end{aligned} \tag{13.152}$$

or:

$$|Z| = R'\sqrt{1 + \mathcal{Q}^2\left(\beta - \frac{1}{\beta}\right)^2} \tag{13.153}$$

The phase angle now follows from Euler:

$$\tan\delta = \frac{\Im m(Z)}{\Re e(Z)} = \frac{\left(\omega L - \frac{1}{\omega C}\right)}{R'} = \mathcal{Q}\left(\beta - \frac{1}{\beta}\right) \quad \Rightarrow \quad \delta = \tan^{-1}\left\{\mathcal{Q}\left(\beta - \frac{1}{\beta}\right)\right\} \tag{13.154}$$

precisely as shown in the original phasor approach. The difference is that *now* we can write the full complex impedance Z ***once and for all*** for this problem with this δ as:

$$Z = R'\sqrt{1 + \mathcal{Q}^2\left(\beta - \frac{1}{\beta}\right)^2}\, e^{i\delta} \tag{13.155}$$

Now finding the complex current as the peak current V_0/R' times a dimensionless, universally scalable form is literally a one liner:

$$I_0 = \frac{V_0}{Z} = \frac{V_0}{R'} \times \frac{1}{\sqrt{1 + \mathcal{Q}^2\left(\beta - \frac{1}{\beta}\right)^2}} e^{-i\delta} = \frac{I_{\max}}{\sqrt{1 + \mathcal{Q}^2\left(\beta - \frac{1}{\beta}\right)^2}} e^{-i\delta} \tag{13.156}$$

(where $I_{\max} = V_0/R'$ as expected from our work before and common sense). How simple is that! No need for phasor diagrams, no figuring out what is ahead of and what is behind the phase of what trig function. No guesses as to the form of the current.

Oh, sure, this is the *complex* form of the current and the *complex* form of the impedance, but that's a simple matter to rectify. We just restore the time dependence to form:

$$I_{\text{complex}}(t) = \frac{V_0}{Z} e^{i\omega t} = \frac{V_0}{|Z|} e^{i(\omega t - \delta)} \tag{13.157}$$

and ***take its real part:***

$$\boxed{I(t) = \Re e\left\{I_{\text{complex}}(t)\right\} == \frac{1}{\sqrt{1 + \left(\beta - \frac{1}{\beta}\right)^2}} \frac{V_0}{R} \cos(\omega t - \delta)} \tag{13.158}$$

Of course we can replace the cosine with sine by simply starting our clock a quarter cycle ahead or behind, or add an arbitrary phase to describe an arbitrary clock (and phase) in the original power supply.

But what about *power*? It turns out that the complex formulation of our answers has a great deal to say about power, in some sense a lot *more* to say about it than we were able to find in the original real+trig phasor approach. To explain this, however, I have to introduce the ***complex power***, defined to be:

$$S = V(t)I(t)^* = V_0 e^{i\omega t} \times \frac{V_0}{|Z|} e^{-i\omega t} e^{+i\delta} = \frac{V_0^2}{|Z|} e^{+i\delta} = \frac{V_0^2}{|Z|}(\cos\delta + i\sin\delta) = P + iQ \tag{13.159}$$

The first term of this is the *magnitude* of the *real* power in the circuit. This is the part that is *in phase with the applied voltage* and is put *into* the circuit by the power supply and taken *out of* the circuit by the load resistance. Its form is:

$$P = \frac{V_0^2}{|Z|} \cos\delta$$

where $\cos\delta$ is the power factor (discussed above).

The second term is new. It is called the (magnitude of the) ***reactive power***:

$$Q = \frac{V_0^2}{|Z|} \sin\delta$$

This has one important use. It is representative of the power going *into and out of* the total reactance of the system *without* being turned into heat by the load resistor. This concept is directly connected to the relative efficiency of a circuit – the power supply has to deliver a larger current than expected in terms of "just" the average load power, and in real world engineering,

where the power supply itself has internal resistance (and may well have internal reactances as well) this has its costs!

To fully explore this is (finally) beyond the scope of this course, even for EEs and physics/math majors – we'd have to enter the sublime realm of impedance matching and much more, but there are entire *textbooks* devoted to this subject, too much for even a useful summary here[218].

From this we can conclude several things. First, the phasor approach *works* for this problem – it gives us the rigorously correct answer from little more than an inspired (or informed) guess and the geometry of triangles. Second, the complex solution is potentially ***much more general*** – it can be used to analyze what happens in the LRC or $rLRC$ circuit when the input is a *more general* waveform voltage $V(t)$ that is *not* a single-frequency harmonic function, once one masters Fourier analysis[219]. It can also be used (as we will see below) to handle at least *simple* multiloop circuits of interest as they for various "frequency filters" that are commonly used in electronics.

For the benefit of those future EE's and other interested readers, I'm going to state without formal proof a very important theorem in filter design that again we have no room to develop or prove here – the ***Maximum Power Theorem***[220] or ***Jacobi's Law*** that states that in general when a power source has a complex internal impedance Z_S and the load has a complex impedance Z_L, maximum power is transferred when

$$Z_L = Z_S^* \tag{13.160}$$

or, the impedance of the load has ***the same amplitude but the opposite phase of the impedance of the source***. This theorem works for purely resistive loads – in fact in its simplest application it simply describes the energy distribution between two resistors R_S and R_L in series!

When designing simple radios, however, an antenna can only deliver a *tiny bit of power* to the input of an electrical circuit designed to receive and decode just one signal out of many incident on the antenna. It then helps a lot to (when possible) match the impedance of the antenna one hopes to use with it in order to maximize the *possible* delivery of power from that particular antenna to the load, and similarly, it helps if the load placed on e.g. a transformer (which often has an inductive component to its internal impedance) is matched to the transformer or vice versa.

It turns out that a parallel $rLRC$ circuit is often more useful for this purpose than a series one, hence we will treat it next. In this section, you'll see the real power of the complex approach laid bare.

13.7.3: Parallel $rLRC$ Circuit – Complex Treatment

Next, let's re-examine the case of a parallel LRC circuit, only *this time* we'll include an explicit source resistance. In this "multiloop" example, the advantages of the complex treatment are

[218] See, for example, https://www.amazon.com/Electric-Circuits-11th-James-Nilsson/dp/0134746961, chapter 10.

[219] This doesn't mean one can always *write down a solution* for these general problems, but if $V(t)$ can be evaluated *numerically*, any computational numerical library worth its salt will have "fast fourier transform" (FFT) routines that will do it all for you.

[220] Wikipedia: http://www.wikipedia.org/wiki/Maximum Power Theorem.

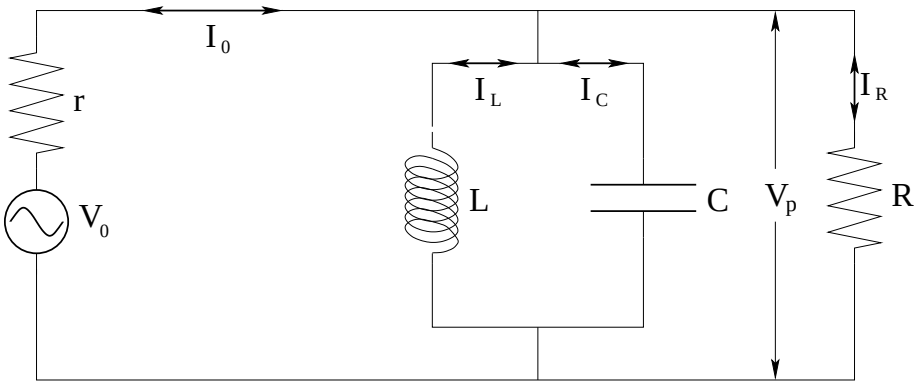


Figure 13.36: Parallel AC circuit *with* internal resistance r

made clear.

One of the beauties of using complex reactances is that complex units keep track of all phases for you, allowing you to – within limits – analyze the circuit almost exactly the same way you would do if the power source were a simple DC battery. As noted above, the voltage drops across the various circuit elements can be written to "look like" a (complex) resistance times a current:

$$\chi_R = R \qquad \Leftrightarrow \qquad V_p = V_R = I_R\chi_R = I_R R \tag{13.161}$$

$$\chi_L = i\omega L \qquad \Leftrightarrow \qquad V_p = V_L = I_L\chi_L \tag{13.162}$$

$$\chi_C = \frac{1}{i\omega C} \qquad \Leftrightarrow \qquad V_p = V_C = I_C\chi_C \tag{13.163}$$

The time dependence of the currents and voltage drops must all be $e^{i\omega t}$ within the phases implicit in the reactances. As before, while V_0, r and R are all real, V_p and all of the I's and χ's will, in general, end up being complex numbers with this time dependence cancelled[221] here and in all subsequent algebra.

The advantage of introducing V_p as the voltage in parallel across all three main circuit elements is that we can express the currents through each one in terms of it much as we did for parallel resistors in the chapter on resistance and DC circuits. This is a shorthand way of writing the other loop equations in the parallel portion of the circuit. This yields:

$$I_R = \frac{V_p}{R} \qquad I_L = \frac{V_p}{\chi_L} \qquad I_C = \frac{V_p}{\chi_C} \tag{13.164}$$

We substitute these three results into Kirchoff's Junction Rule:

$$I_0 = I_R + I_L + I_C = \left(\frac{1}{R} + \frac{1}{\chi_L} + \frac{1}{\chi_C}\right) V_p \tag{13.165}$$

and apply the result to Kirchoff's Loop equation for any of the loop equations (they all have V_p

[221] Or being projected out via a Fourier Transform.

across them):

$$V_0 - I_0 r = V_p$$

$$V_0 = I_0 r + V_p = r \times (I_R + I_L + I_C) + V_p = \left(\frac{r}{R} + \frac{r}{\chi_L} + \frac{r}{\chi_C}\right) V_p + V_p$$

$$V_0 = \left(\frac{r}{R} + \frac{r}{\chi_L} + \frac{r}{\chi_C} + 1\right) V_p \tag{13.166}$$

We can now solve for V_p in terms of V_0 and known parameters:

$$V_p = \frac{V_0}{\left(\frac{r}{R} + \frac{r}{\chi_L} + \frac{r}{\chi_C} + 1\right)} = \frac{V_0}{r}\left(\left\{\frac{1}{R} + \frac{1}{r}\right\} + \frac{1}{\chi_L} + \frac{1}{\chi_C}\right)^{-1} = \frac{Z}{r} V_0 \tag{13.167}$$

I've written it this way so we can define Z in terms of:

$$R \parallel r = R' = \frac{Rr}{R + r} \tag{13.168}$$

as:

$$\frac{1}{Z} = \left(\frac{1}{R'} + \frac{1}{\chi_L} + \frac{1}{\chi_C}\right) \tag{13.169}$$

We finally end up with:

$$V_p = \frac{Z}{r} V_0 \quad \Rightarrow \quad I_R = \frac{V_p}{R} = \frac{V_0}{R}\frac{Z}{r} \tag{13.170}$$

which somehow looks reasonable, at least once one recognizes (from equation 13.167 above) that:

$$\lim_{r \to 0} \frac{Z}{r} = 1 \qquad \Leftrightarrow \qquad \lim_{r \to 0} V_p = V_0$$

In words, if the power supply has *no* internal resistance, the (peak) voltage across R is just V_0, so the (peak) current is obviously $I_0 = V_0/R$!

To find Z is now a straightforward exercise in complex algebra. Let's let:

$$X = \left(\frac{1}{R'} + i\left\{\omega C - \frac{1}{\omega L}\right\}\right) \tag{13.171}$$

so that:

$$X = 1/Z = |X| e^{i\delta} \quad \Leftrightarrow \quad Z = \frac{1}{|X|} e^{-i\delta}$$

Now we know (using **Complex Algebra 101** above):

$$|X| = \sqrt{\frac{1}{R'^2} + \left\{\omega C - \frac{1}{\omega L}\right\}^2} \quad \text{and} \quad \delta = \tan^{-1}\left\{\frac{\omega C - \frac{1}{\omega L}}{\frac{1}{R'}}\right\}$$

or:

$$Z = \frac{1}{\sqrt{\left(\frac{1}{R'}^2 + \left\{\omega C - \frac{1}{\omega L}\right\}^2\right)}} e^{-i\delta} \tag{13.172}$$

with

$$\delta = \tan^{-1}\left(\omega R' C - \frac{R'}{\omega L}\right) \tag{13.173}$$

This is the result we carried back to the discussion of parallel $rLRC$ circuits and the power delivered to the load resistance R in a previous section. As was worked out there in detail from this starting point, if we use:

$$\mathcal{Q} = \frac{L\omega_0}{R'} = \frac{1}{\omega_0 R' C} \quad \Rightarrow \quad \beta = \mathcal{Q}\frac{\omega}{\omega_0} \quad \text{with} \quad \omega_0 = \sqrt{\frac{1}{LC}}$$

we can transform the equations we just derived into the collection of results:

$$Z = \frac{R'}{\sqrt{1 + \mathcal{Q}^2\left(\frac{1}{\beta} - \beta\right)^2}} \qquad \delta = \tan^{-1}\mathcal{Q}\left(\frac{1}{\beta} - \beta\right) \tag{13.174}$$

Note well that δ for the *parallel* LRC circuit is ***minus the δ obtained for the series LRC circuit!*** Some textbooks acknowledge this by writing:

$$I(t) = I_0 \cos(\omega t + \delta)$$

but since one really has to remember *something* for the two kinds of circuit – either different signs in $I(t)$ (but the same δ) or the same sign in $I(t)$ (but opposite sign of δ) – I opted for the latter.

This means (to summarize the results above) that we write:

$$I_R(t) = \frac{V_0}{R}\frac{Z}{r}\cos(\omega t - \delta) \tag{13.175}$$

and therefore (time averaging and substituting) the average power delivered to the load is:

$$\boxed{P_{\text{avg}} = \frac{1}{2}\frac{V_0^2 R}{(r+R)^2} \times \frac{1}{1 + \mathcal{Q}^2\left(\frac{1}{\beta} - \beta\right)^2} = \frac{1}{1 + \mathcal{Q}^2\left(\frac{1}{\beta} - \beta\right)^2} \times P_{\text{avg,max}}} \tag{13.176}$$

which is ***exactly the same*** as equation 13.107 above for the ***series*** $rLRC$ circuit!

Well, not *exactly*... The order of β and $1/\beta$ has reversed! But since it appears *squared*, the sign does not matter *here*. The thing that ***matters a great deal*** (as previously discussed) is that a parallel LRC *inverts* the effective impedance of the circuit so that the power supply has to deliver a *maximum* current at very high or very low frequences – but diverts almost all of that current through either the capacitor or inductor when they *short out* the load resistor R, causing all of the power the power supply is capable of delivering to be dissipated as heat ***inside the power supply itself!***

At resonance, however, the effective "resistance" of the capacitor/inductor in parallel is *infinite* so all of the current flows through the load resistor R in steady state! This is the *minimum* current delivered to the system by the power supply but the *maximum* current that goes through R!

This makes it sound like this circuit is useless – who wants to burn out a power supply, wasting energy that does no useful work as one does so? – but that is far from the case. This is one of the *preferred* circuits used in e.g. crystal radios, and may also be of use when the power supply itself has some internal e.g. inductance as well as resistance (often the case if it is e.g. a transformer or the upstream part of a complicated circuit). This chapter isn't really a proper introduction to Electrical Engineering, but hopefully it *is* enough to get those future EEs amongst you off on the right foot!

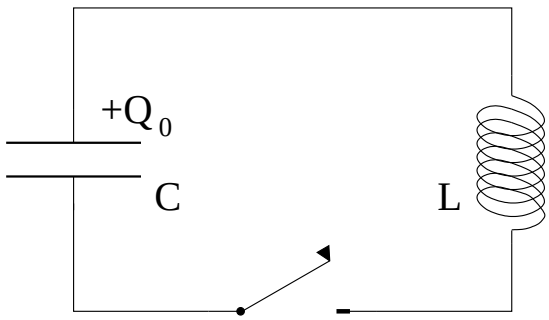
Homework for Week 13

Problem 1.

Physics Concepts

Make this week's physics concepts summary as you work all of the problems in this week's assignment. Be sure to cross-reference each concept in the summary to the problem(s) they were key to. Do the work carefully enough that you can (after it has been handed in and graded) punch it and add it to a three ring binder for review and study come finals!

Problem 2.



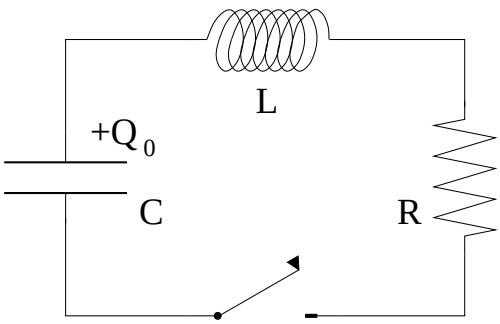
At time $t = 0$ the capacitor in the LC circuit above has a charge Q_0 and the current in the wire is $I_0 = 0$ (there is no current in the wire). ***Derive***:

$$Q(t) = Q_0 \cos(\omega_0 t + \phi)$$

$$\text{with} \quad \omega_0 = \sqrt{\frac{1}{LC}}$$

Then draw a graph of $Q(t)$ with ten periods and $\phi = 0$.

Problem 3.



At time $t = 0$ the capacitor in the LRC circuit above has a charge Q_0 and the current in the wire is $I_0 = 0$ (there is no current in the wire). ***Derive***:

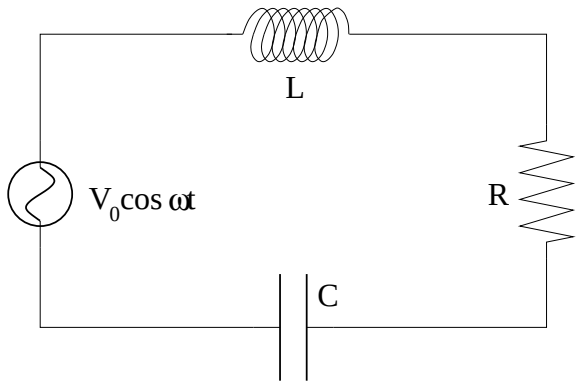
$$Q(t) = Q_0 e^{-t/\tau} \cos(\omega' t + \phi)$$

$$\text{with} \quad \tau = \frac{2m}{b} \quad \text{and} \quad \omega' = \omega_0 \sqrt{1 - \frac{b^2}{4km}}$$

Then draw a graph of $Q(t)$ with ten periods in the case that its exponential damping time $\tau = 2m/b = 5$ periods and $\phi = 0$.

Problem 4.

A *series* LRC circuit connected across a variable AC voltage source $V = V_0 \cos(\omega t)$ is drawn below.



a) Assume that the ***steady state*** current $I(t)$ has the form

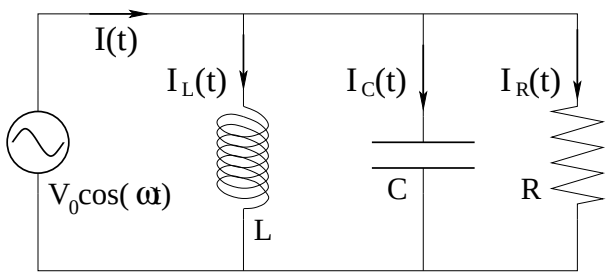
$$I(t) = I_0 \cos(\omega t - \delta)$$

and find I_0 and δ in terms of V_0, ω, L, R, and C.

b) Find the *time average* power P_R delivered to the "load" resistance R.

Problem 5.

A *parallel* LRC circuit connected across a variable AC voltage source $V = V_0 \cos(\omega t)$ is drawn below.



a) Assume that the ***steady state*** current $I(t)$ has the form

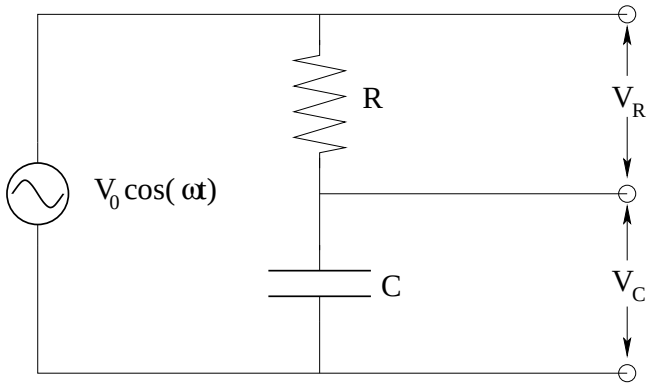
$$I(t) = I_0 \cos(\omega t - \delta)$$

and find I_0 and δ in terms of V_0, ω, L, R, and C.

b) Find the *time average* power P_R delivered to the "load" resistance R.

Problem 6.

In the circuit below, the AC voltage is $V_0 \cos \omega t$.

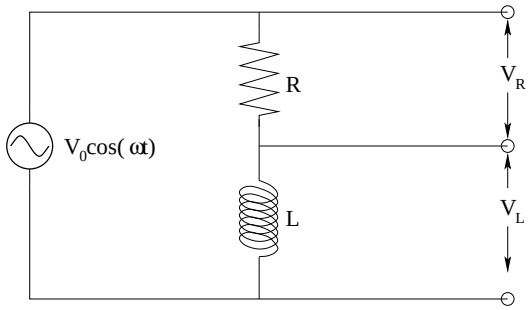


a) Find the current $I(t)$ through the resistor and capacitor, assuming no current is diverted into the branches on the right. Clearly identify the relative phase shift δ between the applied voltage and the current.

b) Find the voltage $V_R(t)$ across the resistor. Factor your answer out so that it is expressed in terms of the dimensionless quantity ωRC. Graph the *amplitude* of $V_R(t)$ as a function of ωRC.

Problem 7.

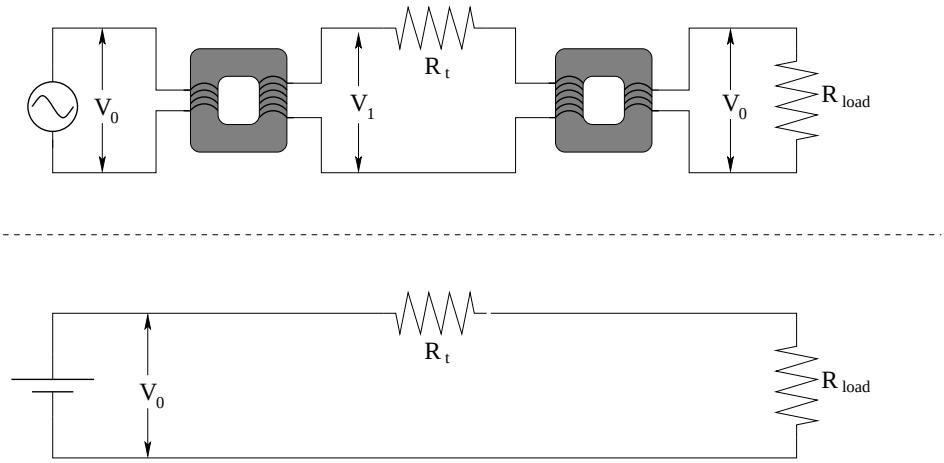
In the circuit shown below, the AC voltage is $V_0 \cos(\omega t)$.



a) Find the current $I(t)$ through the resistor and inductor assuming no current is diverted into the branches on the right. Clearly identify the relative phase shift δ between the applied voltage and the current.

b) Find the voltage $V_R(t)$ across the resistor. Factor your answer out so that it is expressed in terms of the dimensionless quantity $\omega L/R$. Graph the *amplitude* of $V_R(t)$ as a function of $\omega L/R$.

Problem 8.



Let's analyze the problem of ***economical power transmission*** in the early 1880s[222]. Above you can see two alternative schemes for transmitting power long distances being considered at the time.

The first circuit generates AC power at a relatively low voltage V_0 (which is easy). It then steps V_0 up to a very high voltage $V_1 \gg V_0$ and transmits it at high voltage across a long transmission wire of fixed resistance R_t. Finally, it steps it back down to voltage V_0, where the final ***load*** resistance $R_{\rm load}$ is placed across it.

The second circuit simply generates the same low DC voltage V_0 and transmits current down identical transmission lines to where it passes through an identical load $R_{\rm load}$.

Compute the way the power produced by the generator in either case is divided up between $P_{\rm load}$ and P_t, the power wasted heating up the transmission lines. The better solution has $P_t \ll P_{\rm load}$ for reasonable estimates of R_t vs $R_{\rm load}$.

Assume that $V_0 = 100$ volts, $V_1 = 10,000$ volts and $R_t = 100$ ohms. We will imagine the load resistance to be that of (say) one hundred 100 ohm resistances such as old fashioned light bulbs ***in parallel*** in a small neighborhood for a total of $R_{\rm load} = 1$ ohm. Evaluate the ratio of the power lost in the transmission line to the successfully delivered load power:

$$\text{Fractional Loss} = \frac{P_t}{P_{\rm load}}$$

for the two cases and prove that the AC ***high voltage*** solution "wins" in the sense that it minimizes transmission loss compared to delivered power. In the process, get a sense of by how *much* it wins and how that winning *scales* with the transmission voltage compared to the utilization voltage and the load resistance compared to the transmission line resistance.

[222]This problem recapitulates an important historical analysis performed by ***William Stanley, Jr***. Stanley built the first prototype high voltage alternating current energy delivery system in Great Barrington, Massachusetts in 1885, and also, incidentally, invented the "Stanley Thermos" still available today!

Advanced Problem 9.

We wish to evaluate the *Q-factor* for this resonant circuit, as this is an important design parameter for band-pass filters such as those used in radios.

If you correctly solved the driven series LRC circuit problem, you found that:

$$\langle P_R \rangle (\omega) = \langle I^2 \rangle R = \frac{1}{2} \frac{V_0^2 R \omega^2}{\omega^2 R^2 + L^2(\omega^2 - \omega_0^2)^2}$$

is the average power delivered *by* the AC voltage *to* the circuit load R (the inductor and capacitor do not dissipate energy and there is no *net* work done per cycle upon them). In this expression $\omega_0 = \sqrt{1/LC}$ as you should fully understand at this point on the basis of the text, lecture and a previous homework problem as well.

Show that for a sharply peaked resonance (one with "large" Q, say $Q \gtrsim 5$):

$$\Delta\omega \approx \frac{R}{L}$$

so that

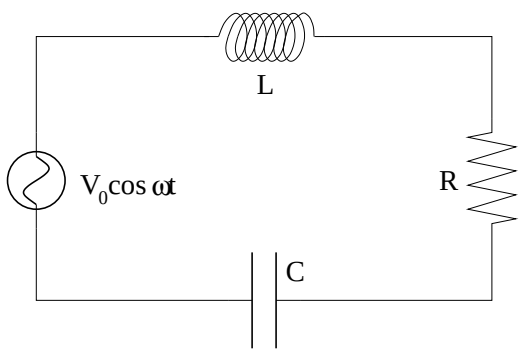
$$Q = \frac{\omega_0}{\Delta\omega} \approx \frac{\omega_0 L}{R}$$

where $\Delta\omega$ is the full width at half maximum of the power curve above.

Hint: To do this, set the expression above equal to the *computed* half-maximum power, and solve for two quadratic roots for ω, assuming that both of them are *very close* to (but not equal to) ω_0 (so that $\omega \approx \omega_0$ when it appears by itself – this is the "sharply peaked" part). You may find the following factorization useful:

$$\omega^2 - \omega_0^2 = (\omega - \omega_0)(\omega + \omega_0) \approx (\omega - \omega_0) \times 2\omega_0$$

Advanced Problem 10.



Derive the full solution to the driven LRC circuit problem using complex exponentials. Start with Kirchhoff's rule for the loop and assume a complex $V(t) = V_0 e^{i\omega t}$ (where by convention V_0 is real). Write $I(t) = I_0 e^{-i\delta} e^{i\omega t}$ – that is, as a general complex amplitude times the *exact same* complex exponential time dependence that the voltage has (that is, so the equation describes a single fourier component of an even more general solution). We'll take the real part of the complex solution at the *end* as usual.

Find an algebraic complex expression that expresses the sum of the voltages. Solve this expression directly using algebra, no pictures of "phasors" really required. Factor out the solution to obtain I_0, δ, Z (the *complex* impedance), and the voltages across each element as a function of time.